Memory & Cognition (2021) 49:1067–1081
https://doi.org/10.3758/s13421-021-01160-7

Relationships between expertise and distinctiveness: Abnormal medical images lead to enhanced memory performance only in experts

Hayden M. Schill[1] · Jeremy M. Wolfe[2,3] · Timothy F. Brady[1]

Accepted: 21 February 2021 / Published online: 14 April 2021

Abstract

Memories are encoded in a manner that depends on our knowledge and expectations ("schemas"). Consistent with this, expertise tends to improve memory: Experts have elaborated schemas in their domains of expertise, allowing them to efficiently represent information in this domain (e.g., chess experts have enhanced memory for realistic chess layouts). On the other hand, in most situations, people tend to remember abnormal or surprising items best—those that are also rare or out-of-the-ordinary occurrences (e.g., surprising—but not random—chess board configurations). This occurs, in part, because such images are distinctive relative to other images. In the current work, we ask how these factors interact in a particularly interesting case—the domain of radiology, where experts actively search for abnormalities. Abnormality in mammograms is typically focal but can be perceived in the global "gist" of the image. We ask whether, relative to novices, expert radiologists show improved memory for mammograms. We also test for any additional advantage for abnormal mammograms that can be thought of as unexpected or rare stimuli in screening. We find that experts have enhanced memory for focally abnormal images relative to normal images. However, radiologists showed no memory benefit for images of the breast that were not focally abnormal, but were only abnormal in their gist. Our results speak to the role of schemas and abnormality in expertise; the necessity for spatially localized abnormalities versus abnormalities in the gist in enhancing memory; and the nature of memory and decision-making in radiologists.

Keywords Expertise · Radiology · Recognition memory · Long-term memory

Our ability to remember information is deeply dependent on our existing knowledge structures, or schemas (Bartlett, 1932; Hintzman, 1986). Even superficially identical information is better remembered if it is integrated into a set of knowledge rather than simply seen as arbitrary. For example, people are better at remembering that someone *is* a baker than that someone's *name* is Baker, because the profession baker activates a rich set of meaningful associations that the name Baker does not (McWeeny et al., 1987); and people remember visual images better if they recognize them as faces than if identical images are not recognized, but seen as meaningless texture (e.g., Brady et al., 2019).

Different people have different knowledge and schemas, in part based on their expertise, and this has consequences for memory: Imagine after playing a round of chess, you are asked to recreate the board from some critical moment in the game. For most people, this task would prove very difficult. However, if you were a world-class chess player, this might be quite easy. Becoming an expert in a domain such as chess changes our memory for items in that domain of expertise (Chase & Simon, 1973; de Groot, 1946), allowing us to store more information as long as this information is consistent with the expectations we have formed as a result of our expertise (Gobet & Simon, 1996).

A large literature is devoted to quantifying memory benefits in experts compared with novices (e.g., Ericsson & Kintsch, 1995; Engle & Bukstel, 1978; Gobet & Simon, 1996; Vincente & Wang, 1998). For example, car experts can remember more car images in visual working memory (Curby et al., 2009); baseball experts can remember more baseball-related information in long-term memory (Voss

✉ Hayden M. Schill
hschill@ucsd.edu

[1] Department of Psychology, University of California, San Diego, San Diego, CA, USA

[2] Department of Surgery, Brigham & Women's Hospital, Boston, MA, USA

[3] Departments of Ophthalmology and Radiology, Harvard Medical School, Boston, MA, USA

et al., 1980); and expert radiologists have better long-term memory for mammograms—but not natural scenes or real-world objects—compared with controls (Evans et al., 2011).

Why do experts show this increase in memory performance for their domain of expertise? In the literature on expertise, many authors posit that memory improvement occurs because existing knowledge allows experts to know what variation to expect for information in an expert's domain (e.g., Vincente & Wang, 1998). That is, existing schemas make the relevant part of the information predictable and thus easier to encode and remember (Graesser & Nakamura, 1982). Thus, in many ways, memory benefits in experts may be considered a manifestation of a broader phenomenon where information that is understood as meaningful—and thus integrated into a schema—is easier to correctly recognize or recall (Bartlett, 1932). For experts, there may simply be a wider variety of meaningful concepts and schemas, resulting in a richer ability to understand and remember stimuli in their domain of expertise (e.g., Ericsson & Kintsch, 1995). This is sometimes known as an organizational processing account of expertise: that experts can have improved memory because they are better able to chunk this information and otherwise create effective knowledge structures (Ericsson & Kintsch, 1995; Rawson & Van Overschelde, 2008).

Is better organization the sole reason for better memory in experts? Beyond schemas and knowledge organization, experts in some domains—particularly those where the expertise is more perceptual, like radiologists looking at mammograms or car experts focusing on the details of cars—may have developed specialized processing mechanisms for their domain of expertise which take advantage of the way stimuli vary in that domain. For example, experts in some domains employ more holistic processing strategies for objects of their expertise (Bilalić et al., 2011; Gauthier et al., 2000; Gauthier et al., 1999; Richler et al., 2011; Watson & Robbins, 2014). Enhanced perceptual expertise may allow experts to process more information about an item even in the same amount of time, and lead to richer memory traces (Ericsson & Kintsch, 1995).

In addition to building richer knowledge structures and better perceptual encoding, there is a third factor that could explain experts' improved memory performance in domains of expertise, which has often been overlooked in studies of memory: increased distinctiveness of items when they are items of expertise (Rawson & Van Overschelde, 2008). In contrast to views that claim memorability is an intrinsic aspect of a stimulus (e.g., Bainbridge et al., 2013), a significant body of literature argues instead that the critical driver of how memorable an item is in a given context is its distinctiveness from other items currently being stored in memory. Imagine, for example, you are given a list to remember that has 30 animal names and also the word "bread" on it. People tend to remember this distinctive word ("bread") most accurately—and this is true even if it appears first on the list, so its unique status is not yet known and it is not differentially attended or processed (Calkins, 1894; Hunt, 2006). Memory models naturally predict this effect because most memory models propose that memory is strongly limited by interference at retrieval, and having more unique features allows easier retrieval (e.g., Shiffrin & Steyvers, 1997).

This is broadly consistent with the idea that abnormal or schema-inconsistent items tend to be *better* remembered than expected, schema-consistent items (Friedman, 1979; Hollingworth & Henderson, 2003; Light et al., 1979; McDaniel & Einstein, 1986; Pedzek et al., 1989). For example, people tend to better remember unexpected aspects of images (Friedman, 1979).

How does such distinctiveness interact with expertise? For experts, many items may be unique from other items in a set in a way that would not be noticed by nonexperts, thus enhancing memory for those items as they would then be more unique in the set for experts than nonexperts (Rawson & Van Overschelde, 2008).

In summary, experts are often better at accurately recognizing or recalling information in their domain of expertise. This can arise from at least three factors, each of which has been independently studied: experts may have changed perceptual processing strategies; may benefit from general usage of schemas to organize memory; or may benefit from increased distinctiveness of items in memory. However, the way these effects interact has rarely been studied, and many have been studied primarily in domains with limited or no perceptual expertise available (e.g., in word lists).

The current experiments: Memory for mammograms in novices and expert radiologists

To understand how expertise effects memory, and how each of these three factors may play a role, the current experiments ask how expertise affects memory for mammograms (comparing novices and expert radiologists), and test whether expert radiologists have better memory for abnormal images (i.e., cancerous mammograms), when compared with normal images (i.e., noncancerous mammograms). While for normal mammograms, perceptual encoding benefits, schemas, and distinctiveness all likely play a role in expert's memory, abnormal mammograms provide a unique case study. Abnormal mammograms do not violate a radiologists' schema (as they are trained to look for abnormalities), but abnormal cases do provide distinctive retrieval cues (e.g., this mammogram has calcifications in this location) which would not be available to nonexperts who have no idea that those little white spots are significant. Nor would these cues be available in normal mammograms. Abnormal mammograms therefore present an

interesting case; they are schema-consistent, while also potentially providing a unique window into the role of distinctiveness in expert's memory.

To measure memory performance, we will use receiver operating characteristic (ROC) analysis to take into account the possibility of differential false alarms and differential response criterion, which is critical to understand whether any effects we observe are truly changes in memory strength. We predict that experts will have improved performance compared with nonexperts for both normal and abnormal mammograms because of their perceptual expertise and because they have developed schemas over time to represent these complex images. We also predict that abnormal items might show even more benefit for radiologists compared with nonexperts because for radiologists and radiologists alone, these images have unique and distinctive retrieval cues available.

We focus on radiologists' memory for mammograms for two reasons: First, search for signs of breast cancer involves a usefully specific perceptual expertise. For instance, only 2-3 kinds of local abnormalities are typically present in abnormal mammograms, and radiologists have significant perceptual expertise whether looking at normal or abnormal medical images.

Second, there are two senses in which a mammogram might be considered "abnormal": (1) It could contain a focal abnormality. In our study, these are masses or architectural distortions that are subsequently proven to be malignant. (2) Given a mass (for example) in one breast, the other breast could be considered abnormal in the sense that the image comes from a patient with cancer. We assess the impact of each of these two kinds of abnormality on memory. Note that a mammogram might be considered "abnormal" if it showed a benign mass. We did not use such stimuli in this study.

Radiologists are explicitly trained to recognize an image as abnormal if they detect the presence of a visible, localized abnormality, like a mass or calcification. In addition, recent research has shown that, if asked in an experimental setting, radiologists have an ability to detect a "gist" of abnormality in the breast contralateral to the lesion. They perform at above chance levels when asked to categorize images as coming from normal or abnormal patients (Evans et al., 2016). In other words, this study suggests that radiologists do not always need to see a localized physical lesion to know that an image is abnormal. This global signal of abnormality is relatively subtle. More importantly, for present purposes, work on this gist signal is new enough that most radiologists are unfamiliar with the concept. Thus, any impact on memorability could be considered to be the result of an implicit effect of abnormality.

Published studies of the gist of abnormality have involved giving radiologists only a brief (250–500 ms) glance at an image. While this seems sufficient for expert radiologists to gain some evidence of abnormality, it remains unknown whether this ability impacts radiologists' memory for normal versus abnormal images.

To summarize, the questions guiding this experiment are the following: Do radiologists show improved memory performance for abnormal images compared with normal images? If so, does global gist produce enhanced expert memory for images of the breast contralateral to the breast that contains focal signs of cancer? Alternatively, does any abnormality advantage in memory depend upon having a focal abnormality that can draw spatial attention?

Experiment 1 is a baseline study with novice observers, whose performance can be compared with radiologist performance in Experiment 2. In addition, Experiment 1 allows us to determine whether our stimulus set contains images that are memorable regardless of expertise. Experiment 2 assesses memory performance in expert radiologists. To anticipate our results, Experiment 1 reveals patterns in our image set that we take into account in Experiment 2. In Experiment 2, we find a large memory benefit for radiologists relative to novices as well as an abnormality advantage in radiologists for focal abnormalities. We find no evidence that experts make use of a nonfocal gist of abnormality either in judgment or memory.

Experiment 1: Novices

Experiment 1 was conducted using novice (nonradiologist) observers. The design, number of observers, exclusion, and analysis plan for this experiment were preregistered (URL for this experiment: http://aspredicted.org/blind.php?x=xr3843).

In this experiment, novice observers viewed a series of mammograms and judged whether each case was normal or abnormal and whether they remembered seeing the image earlier in the experiment. We would expect both, judging whether an image is normal or abnormal as well as remembering the images to be difficult, as this task is designed for expert radiologists. However, novice performance provides a useful baseline for comparing radiologist performance and provides a baseline of memory in novice observers. In particular, the results of this experiment can indicate if particular images are particularly distinctive in the absence of any mammographic expertise.

Method

Participants

Sixty participants (23 female participants, mean age 38 years) were recruited for this experiment through Amazon's Mechanical Turk, which offers monetary compensation for participation in online tasks. Mechanical Turk workers are reasonably representative of the American adult population (Berinsky et al., 2012; Buhrmester et al., 2011; Difallah

et al., 2018), and provide data that are comparable to data obtained when participants are tested in experimental psychological laboratories (e.g., see Brady & Alvarez, 2011, for a comparison in a visual memory context). All participants gave informed consent, were compensated at a rate of approximately $10/hour, were located in the United States, and had a hit approval rate greater than 95%. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Stimuli and procedure

Participants viewed single breast mammograms in this study. The stimulus set consisted of 80 abnormal (cancerous) cases and 40 normal (noncancerous) cases. All images were deidentified. All images were preclassified by a group of trained radiologists who did not participate in the study. Normal images were noncancerous and did not contain benign lesions. Abnormal images consisted either of histologically verified malignant masses or architectural distortions (see Evans et al., 2016, for a more detailed description of this stimuli set). Half of the abnormal images contained a visible abnormality (i.e., a lesion was present) and half were images of the breast contralateral to the breast with the lesion (i.e., still an abnormal case, but with no focal indication of that abnormality). Thus, the entire set consisted of 40 normal images, 40 focal-abnormality images (herein referred to as abnormal), and 40 non-focal abnormality images (images contralateral to the breast with the focal abnormality), herein and henceforth referred to as contralateral-abnormal. Each image subtended approximately 16 × 20 degrees of visual angle at an estimated viewing distance of approximately 60 cm from the screen.

On each trial, one image was present for 3 seconds, followed by a new screen containing response questions. The mammogram was randomly chosen to be either normal, abnormal, or contralateral-abnormal. Critically, each image was also either a new image (presented for the first time in the experiment) or a repeated image from 3 trials back or 30 trials back (3-back and 30-back, respectively). Of the images that were later repeated, 50% were repeated at 3-back, and 50% were repeated at 30-back. The experiment was balanced such that ~20% of trials in the first and second half of the study were 3-back and 30-backs, respectively. In fact, due to sampling different streams of images for each participant, in our exact pool of radiologists, 18% of trials were 3-backs in the first half of the trials, versus 23% in the second half of the trials, and 22% were 30-backs in the first half of the trials, and 20% in the second half. In total, with repetitions, there were 210 trials: 120 new images (40 per condition), plus 90 repeat images (30 per condition, split evenly between 3-back and 30-back).

After being displayed for 3 seconds, each image was immediately followed by two response questions: (1) Was the image abnormal or normal? (2) Have you seen this image before? Using a standard computer mouse, participants were told to indicate their level of confidence on a 6-point rating scale ranging from confident yes/abnormal to confident no/normal (see Fig. 1). We collected confidence ratings instead of simple yes/no answers to allow for ROC analysis. There was no time constraint imposed on responding. The initiation of the next trial was contingent on answering both questions of the current trial.

Before the experiment began, participants were presented with instructions and several demographic questions (Gender; Age; "Are you a radiologist?"; "Do you have a job where you read medical images; i.e., tech, medical physicist?"). Instructions were written for a novice population with no medical training. For novice participants, abnormal cases were broadly defined as "images that might contain lesions, or cancer, or otherwise might be something worthy of follow-up if you were a radiologist."

Exclusion criterion and analysis plan

Our exclusion criteria and analyses were decided in advance (see preregistration, above). Individual trials were excluded if participants took less than 1,500 ms or more than 15,000 ms to respond (based on pilot data). Participants were excluded if they took less than 15 minutes (zero excluded) or more than 1 hour to complete the study (3 excluded). Radiologists were excluded (1 excluded) as were those with other prior experience reading medical images (zero excluded). Participants were also excluded if they had more than 80% identical responses (e.g., picked the exact same answer on nearly every trial; one excluded) or had more than 20% of trials excluded on the basis of the reaction time criteria (one excluded). After applying these a priori exclusion criteria, seven participants were excluded from analysis, leaving a final sample of 53 participants.

Following our preregistered analyses (above), we did not conduct an overall analysis of variance (ANOVA) initially but rather followed our specific targeted tests. We first analyzed the confidence ratings of classifying an image as abnormal or normal. We subsequently analyzed the confidence ratings representing memory for images. In order to do this, we conducted ROC analysis for 3-back and 30-back as a function of image type (normal/abnormal/contralateral-abnormal). We also generated ROCs for the normal/abnormal judgments. ROCs were summarized by area under the curve (AUC) and compared using *t* tests. As noted, we are interested in whether, within the group of novice participants, there is a benefit in memory performance for any type of image (e.g., as judged by normal vs. abnormal AUC). Since the novices lack medical experience, any such effect would give us insight into the nature of the image set (i.e., memorability or distinctiveness).

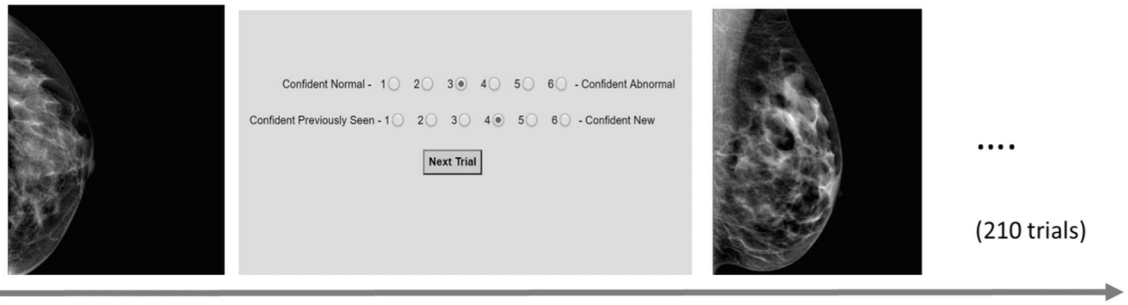

Fig. 1 Method. $N = 60$ nonexpert novice participants rated a sequence of 210 images on normal/abnormal and old/new. Images could repeat either after three or 30 subsequent images and be either normal, abnormal, or contralateral-abnormal

Finally, we conducted image similarity analyses to quantify how image differences might be influencing these results.

Image similarity comparison

Because normal, focally abnormal and contralateral-abnormal images are necessarily different image sets, it is useful to compare how distinctive each set of images is from all the other images in order to look at the effect this has on memory. One way to accomplish this is to have individuals give similarity ratings between images. However, this would require 120 × 120 = 14,440 similarity ratings. Instead, to streamline the process, we relied on previously established computer vision techniques designed to give similarity measurements for natural scenes. Specifically, we conducted a Gabor wavelet pyramid (GWP) analysis, which computes features of the images and compares them (Greene et al., 2016; Kay et al., 2008). To assess the level of similarity in the different image types, the GWP represents each image as the output of a bank of multiscale Gabor filters. Prior work has shown that these features can successfully model object representation in early visual areas (Kay et al., 2008). Following the exact procedure and parameters provided by Greene et al. (2016), each image was converted to grayscale, down sampled to 128 × 128 pixels, and represented with a bank of Gabor filters at three spatial scales (3, 6, and 11 cycles per image with a luminance-only wavelet that covers the entire image), four orientations (0, 45, 90, and 135 degrees) and two phases (0 and 90 degrees). This gave a set of features for each image, which we then compared with all 120 images to compute a distance/dissimilarity score by computing the dot product of each images features to each other images after subtracting the mean across images and normalizing the feature vectors to unit length.

Results (Experiment 1: Novices)

Performance on the classification task

First, we looked at how confident novices were at classifying an image as either normal or abnormal (see Fig. 2). We found a significant difference between normal and abnormal images, $t(52) = 4.78$, $p < .001$, but not between normal and contralateral-abnormal images, $t(52) = 1.94$, $p > .05$.

While participants did not have training to distinguish normal from abnormal medical images, a small number of images in the set are extremely saliently abnormal (i.e., a single bright white spot would look questionable even to novice viewers). Looking at ratings by image (see Fig. 3) reveals that these images are largely responsible for the significant difference between normal and abnormal ratings. In short, for at least for a small subset of images, even novice participants can notice the abnormality, leading to above-chance classification performance broadly. But for most images, novices seem to have little information about normality versus abnormality.

Note that the *y*-axis in Fig. 2 represents the confidence ratings for novices. It is clear that the novices are generally not confident in distinguishing any image type, with average responses tightly clustered near the middle of the rating scale for all conditions. Another way of visualizing this data is on an

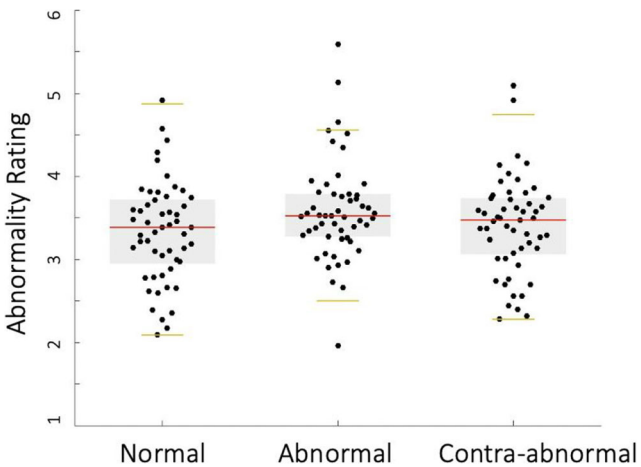

Fig. 2 Classification task: Overall performance of novices on labeling an image as normal or abnormal. The confidence rating scale is now plotted on the *y*-axis. Each point in the plot represents the rating for a particular image. We found a significant difference in confidence in classifying normal versus abnormal images, which seems to be driven by a few salient abnormal images. Novices are not confident in distinguishing between any image type (most responses tend to be in the middle of the confidence scale, no matter the image type). Error bars represent standard error of the mean

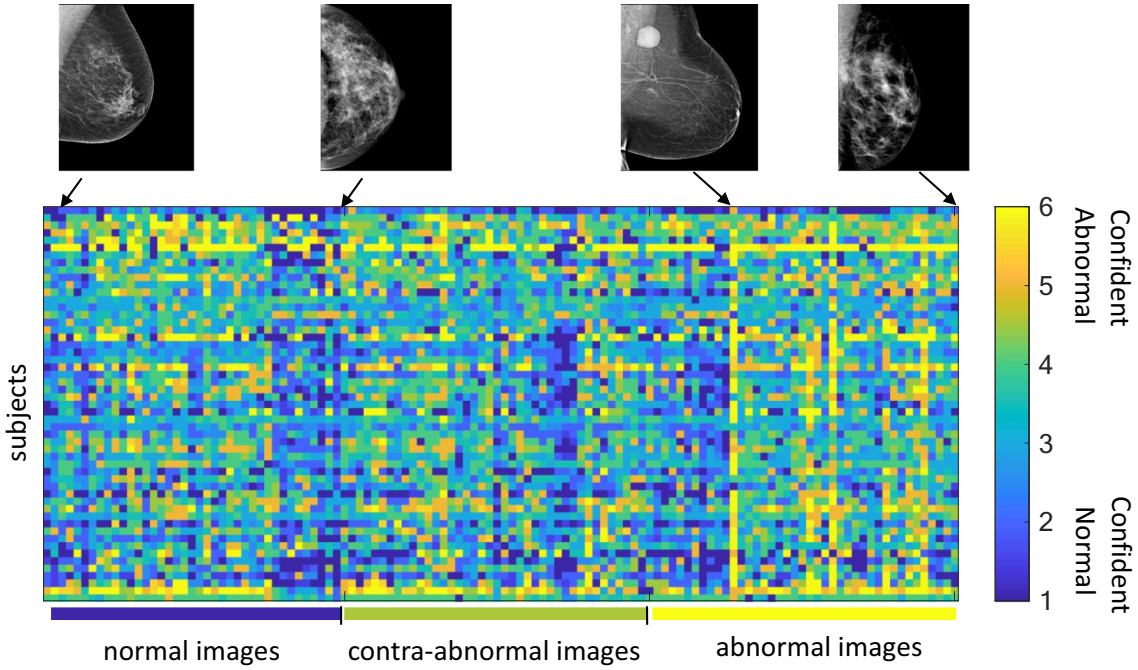

Fig. 3 Image ratings in the classification task. Example images and their confidence ratings for each participant in the classification task. As can be seen with the third pictured image, most participants rated this as abnormal with high confidence. Altogether, the two or three brightly striped vertical lines in the abnormal image set indicate that those and only those images were reliably rated as abnormal by a large majority of participants

ROC curve (see Fig. 4), where novices fall almost on top of the dotted diagonal line indicative of chance performance, with an AUC of only 0.54 (where 0.50 is chance and 1.0 is perfect). Although, as stated above, this difference from chance is highly reliable across participants, $t(52) = 4.21$, $p < .001$, largely because of the few images that participants could all reliably classify.

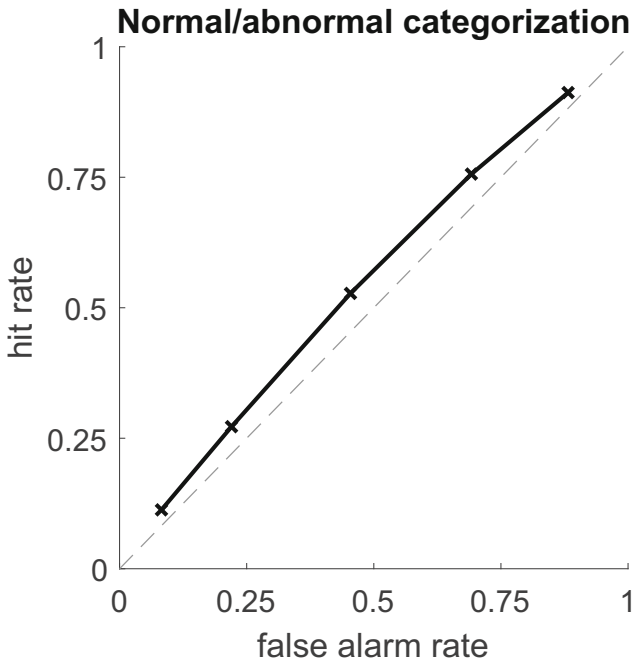

Fig. 4 ROC for normal/abnormal categorization. Novices are very close to the diagonal line representative of chance performance, indicating that they do not perceive a strong difference between normal and abnormal images. The significant effect is driven by a select few salient images (see Fig. 3)

Memory for abnormal images

Figure 5 shows the ROCs for the 3-back and 30-back memory tasks. Since novices were not, for the most part, able to perceive contralateral-abnormal images as different from normal images in the classification task, we focused exclusively on memory differences between normal and abnormal images. Overall, independent of image type, and as expected, novices have better 3-back memory (averaged AUC of 0.70 for detecting 3-backs) than 30-back memory (averaged AUC of 0.64 for detecting 30-backs), $t(52) = 6.59$, $p < .001$. Interestingly, breaking down performance across image conditions reveals that novices show a small normality benefit: they remember normal images better than abnormal images in both the 3-back condition and the 30-back condition, with only the 3-back yielding a significant difference. We found an AUC benefit of 0.069 for normal images at 3-back, $t(52) = 5.48$, $p < .001$, compared with abnormal, and an AUC difference of 0.026 for normal images at 30-back, $t(52) = 1.70$, $p = .096$, compared with abnormal.

Given the weak performance at discriminating normal from abnormal images, it is rather surprising that normality had any effect. Therefore, we examined the data for evidence of more basic effects of visual similarity. We found that the lower

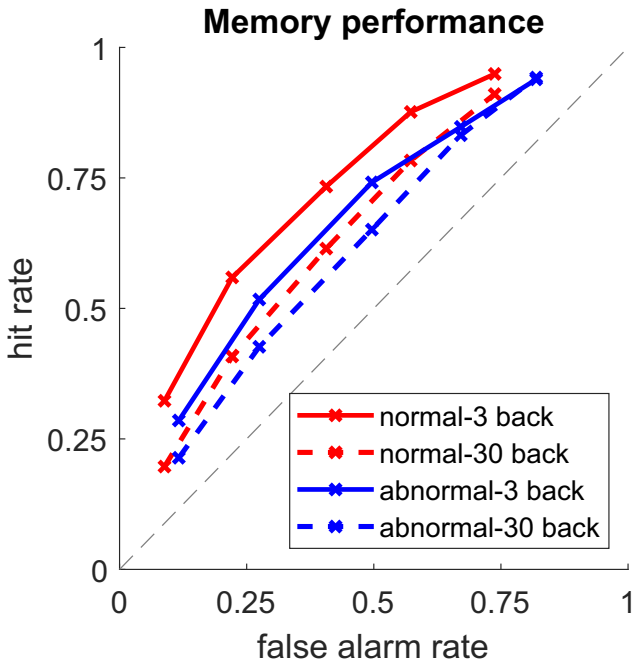

Fig. 5 Novice performance on the memory task. As noted, the gray dashed line indicates chance, and more bowed out curves represent better memory performance. Novices had stronger memory for images in the 3-back condition than in the 30-back condition. Novices also show a small effect of normality, with memory for normal images being better than for abnormal images in both 3-back and 30-back conditions

memory performance in the abnormal conditions was largely driven by an increased false-alarm rate in the contralateral-abnormal and abnormal image sets. Here, we are classifying as "new" all images with a confidence rating >3. This is consistent with an image similarity account in which novices would be more likely to false alarm to new images in the contralateral-abnormal and abnormal conditions simply because these images are more similar to one another than images in the normal set (as predicted by summed similarity accounts of memory; e.g., Nosofsky, 1991). In other words, if the normal images were somewhat more dissimilar to each other compared with the other images, this could explain why novices have somewhat better memory for the normal condition (i.e., it is easier to determine if an image of a dog is new if that dog is presented in a series of different animals than if it is presented in a set of similar dogs. Obviously, the similarity effects in our stimuli are much smaller.). We test this hypothesis next.

Similarity matrix—Gabor wavelet pyramid analysis

We tested this image similarity hypothesis by measuring similarity between our images as described in the Methods (Greene et al., 2016; Kay et al., 2008). We found increased *dissimilarity* among normal images relative to contralateral-abnormal and abnormal images (normal = 0.174; abnormal = 0.139; contralateral-abnormal = 0.133). In other words, normal images were more different, on average, from one another (and thus more discriminable in memory) than either abnormal or contralateral-abnormal images. This is consistent with the hypothesis that the small difference in memory favoring normal images is driven by image similarity differences across sets. Thus, the small normality benefit found in the current study is likely a result of image similarity. Critically, this can provide a useful baseline for considering memory for the same images in expert radiologists in Experiment 2.

Experiment 2: Radiologists

Experiment 2 was the same as Experiment 1, except conducted on radiologist observers.

Method

Participants

Thirty-two expert radiologists (14 female participants, average age = 49 years) were recruited during the 2018 Radiological Society of North America (RSNA) conference in Chicago, Illinois. All radiologists gave informed consent and were not compensated beyond being entered into a lottery for a $500 gift card. Informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Data from participants would have been excluded if they took less than 15 minutes or more than 1 hour to complete the study, had more than 80% identical responses, or had more than 20% of trials excluded. Under these guidelines, no radiologists were excluded from analysis, leaving a final sample of 32 participants.

Stimuli and procedure

The stimuli and experimental design were the same as described in Experiment 1. The main procedural difference was that the experiment was conducted at the RSNA conference where the experimenter explained the instructions in person. Unlike in Experiment 1, in Experiment 2, we gave more general instructions, asking for any abnormality rather than specifically asking participants to look for focal lesions or cancer: "For each image, please judge whether the image is abnormal or normal, and whether you have previously seen it during the course of the experiment."

Results

In this section, we compare the performance of expert radiologists to the performance of novice participants in Experiment 1. In particular, we investigate how nonexperts compare to experts' judgments of image classification (i.e., normal vs.

abnormal), and critically, whether experts show differential memory for abnormal versus normal images. While analyzing expert performance, we take into account idiosyncrasies in our image set that we learned from Experiment 1, such as that our normal images are more dissimilar and therefore inherently slightly more memorable.

Performance on the classification task

Similar to Experiment 1, we first analyzed performance on the classification task by looking at the confidence ratings of classifying each image as either normal or abnormal. How good are radiologists at simply distinguishing abnormal from normal images? Unsurprisingly, radiologists are very good at distinguishing abnormality (see Fig. 6a). Radiologists were significantly more confident that an abnormal image was abnormal instead of normal, $t(31) = 13.17$, $p < .001$. Figure 6b shows the ROC curve for distinguishing focal-abnormal images from normal images in radiologists. ROCs were summarized by area under the curve (radiologist AUC = 0.72; recall that novice AUC = 0.54). As noted in Experiment 1, controls are close to the diagonal line indicative of chance, whereas radiologists elicit a typical curvilinear ROC indicative of a perceived (and significant) difference between normal and abnormal images with an AUC well above chance, $t(31) = 19.8$, $p < .001$.

Next, we looked at if radiologists could detect abnormality in the contralateral-abnormal images. There was not a significant difference between the normal and contralateral-abnormal image conditions, $t(31) = 0.43$, $p = .67$. In the original study of Evans et al. (2016), they found an effect of abnormality in the gist information (i.e., in a very short presentation time of ~250 ms). Our instructions and stimulus set may have biased participants against reporting contralateral images as abnormal. In a set of images that include visible lesions (the abnormal cases) and in the absence of an instruction to look for asymptomatic images from symptomatic patients (the contralateral cases), it is, perhaps, not surprising that radiologists reserved their abnormal ratings for the abnormal cases with lesions. Furthermore, it is possible that our instructions could have primed radiologists to look for both benign and malignant lesions, although no benign lesions were present in the current study. Future studies could investigate the effects of instruction on this task. Recall, however, that our interest in the present experiment is in radiologists' memory for these images. Contralateral-abnormal images, for instance, might still be remembered better if their vaguely suspicious appearance caused radiologists to devote more attention to them.

Memory for abnormal images

Figure 7 shows radiologist performance on the memory task. Radiologists have better memory for abnormal images in both memory conditions, but the advantage for abnormal images is only significant in the 30-back condition, $t(31) = 2.86$, $p = .008$, AUC difference = .051. We found an AUC advantage of 0.02 for abnormal images at 3-back. Although this was not significant, $t(31) = 1.62$, $p = .12$, it follows the same trend as the 30-back condition.

Radiologists showed no memory benefit for the contralateral-abnormal images, even at long delays ($p = .24$). Since radiologists were not able to distinguish between contralateral-abnormal images and normal images in the classification task, this result might be expected; though, recall that we were looking for evidence that an implicitly recognized abnormal gist might enhance memory. That is not what we found. Overall, independent of image type, radiologists have better memory at 3-back (averaged AUC of .852 for detecting 3-backs) than 30-back (averaged AUC of .752 for detecting 30-backs) for medical images. Why are radiologists better at 3-back than at 30-back? While it seems clear that this difference largely reflects typical effects of forgetting and interference (e.g., Wixted, 2005), it is also possible that observers would be more likely to "catch on" to the presence of 3-back rather than the 30-back repetitions. If so, they might adopt a strategy that prioritized the 3-back task. However, given that the 3-back and 30-back tests were equally likely and equally distributed throughout the task, and that observers consistently said they remembered seeing mammograms from 30 images back (and therefore were distinctly aware that 3-back wasn't the only *n*-back test present), it seems unlikely that observers would transition to a strategy that only prioritized 3-back memory task. Taken together, these results suggest that experts have better memory overall at 3-back than at 30-back, but that a memory benefit for abnormal images compared with normal images is significant only at 30-back.

In recognition memory studies, it is almost always found that ROCs are not consistent with an equal variance signal detection model (e.g., Egan, 1958; Wixted, 2007). One way to look at this is to convert the hit and false-alarm rates to *z* scores and to plot zROC functions. On a zROC graph, equal variance produces data with a zROC slope of 1.0. Instead, as is typical in recognition memory tasks, the slopes of our zROCs were reliably below 1.0 in 3 of the 4 memory conditions. We fit a linear mixed-effect model with slope and intercept as random per-subject factors (mean slope[*M*] = 0.68 for 3-back for normal images, difference from 1.0: $p < .001$; $M = 1.05$ for 30-back for normal images, not different from 1, $p = .60$; $M = 0.52$ for 3-back for abnormal images, $p < .001$; 0.82 for 30-back for abnormal images, $p = .005$). Collapsing across all conditions, thus allowing the slope to be more reliably estimated, the mean zROC slope was 0.68, significantly different from 1.0 ($p < .00001$). Taken together, then, the ROCs we observed in memory were inconsistent with an equal variance signal detection model and consistent with an unequal variance model, potentially due to variation in memory strength between different items. This is typical of recognition

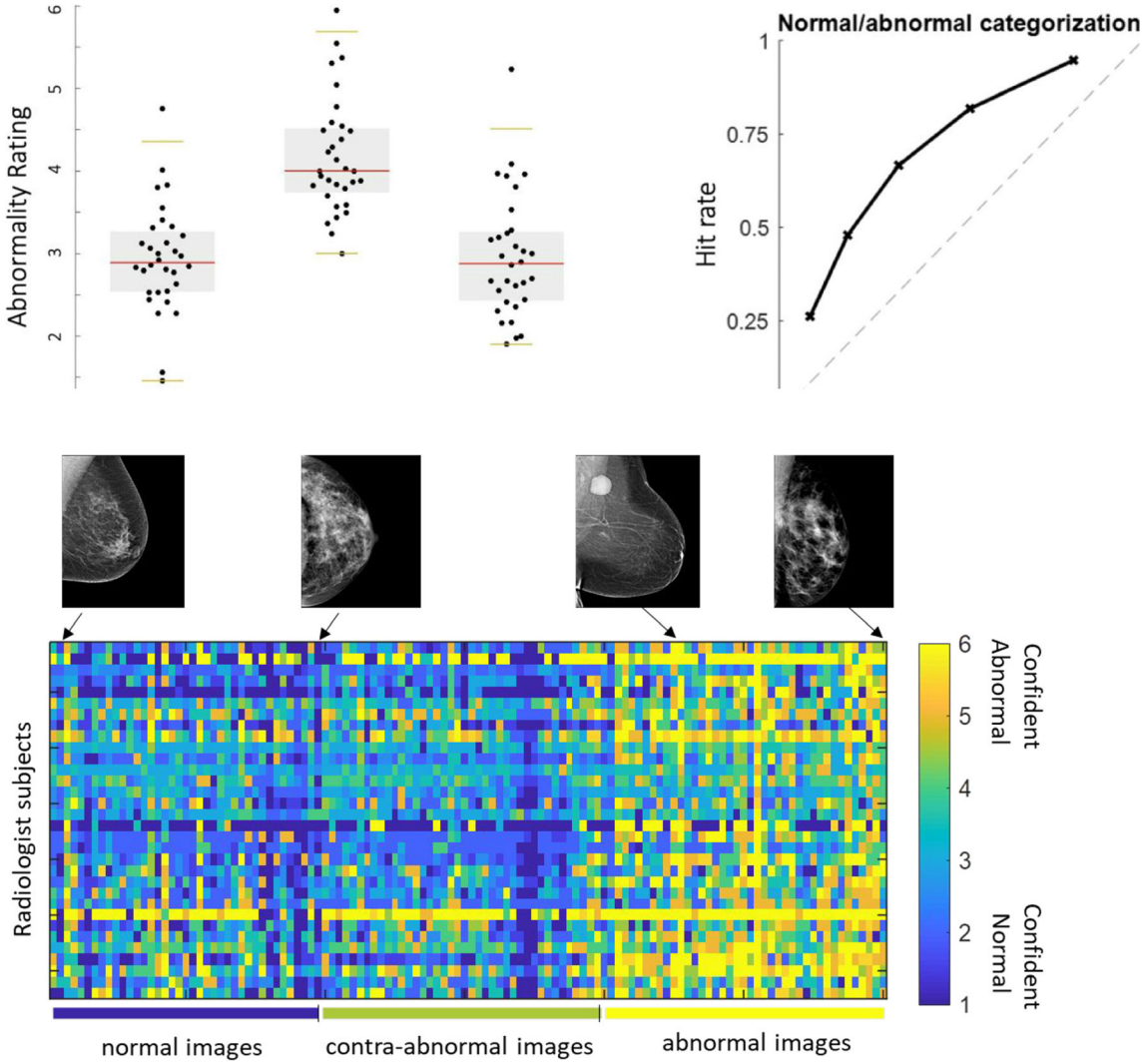

Fig. 6 a (top left): Classification task: Overall performance of radiologists on labeling an image as normal or abnormal. Once again, each point in the plot represents the average rating for a particular image. Radiologists clearly distinguished abnormal from normal images, but they did not distinguish between contralateral-abnormal and normal images. **b (top right):** ROC depiction of performance for labeling an abnormal image as abnormal instead of normal (ignoring contralateral-abnormal images). **c (bottom):** Classification by image. Unlike novices, experts reliably classify most of the abnormal images as abnormal and most of the normal images as normal, with performance not largely driven by any particular subset of images

memory and the reason that collecting confidence judgments and performing ROC analysis is necessary in order to assess memory strength. Simple *d'*, in this context, does not properly account for response criteria differences (e.g., Dougal & Rotello 2007).

Recall from the similarity analysis in Experiment 1 that the normal images in our data set are less similar to each other than the abnormal images, and thus memory for normal images should be better than abnormal (as it was in novices). In fact, it is memory for the abnormal images that is better in radiologist observers. This suggests that the effect of expertise more than compensates for differences between the stimulus categories in image similarity. To see what the effect of abnormality is, independent of baseline image similarity differences, we can compare radiologists' memory performance to novices' performance with the same images. To do this, we compare the benefit—in terms of AUC of the ROC—for radiologists relative to controls in each condition. Doing so reveals a significant abnormality benefit at both 3-back, $t(31) = 6.67$, $p < .001$, and 30-back in expert radiologists, $t(31) = 4.33$, $p < .001$, where, taking their performance after baselining relative to the performance of novice participants, radiologists were specifically better at remembering abnormal images (see Fig. 8).

Extracting additional information with a second presentation

Due to the structure of this experiment, designed to probe memory, each item in the memory set has two classification ratings (for normal/abnormal). Thus, while we set out to probe

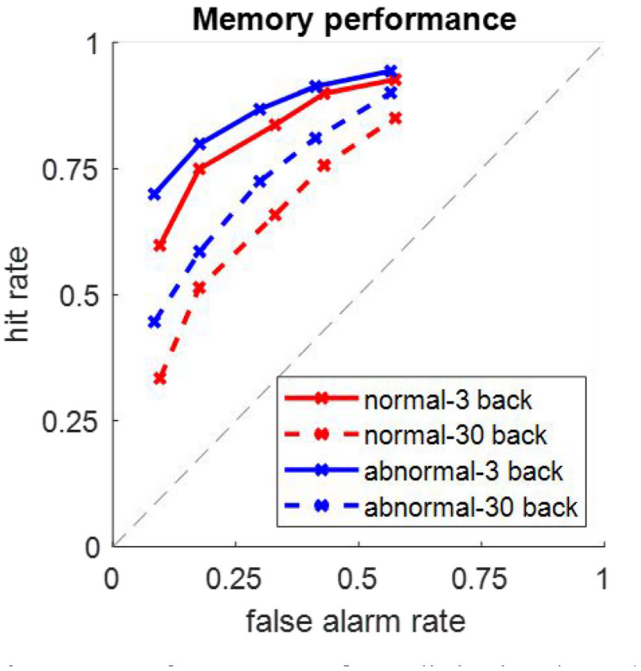

Fig. 7 Radiologist performance on the memory task. Radiologists have better memory for abnormal images in both of the memory conditions. However, only memory at long delays (30-back) was significant

memory, the experiment also makes it possible for us to combine both ratings in order to examine whether there is a "crowd-within" effect in this situation (Vul & Pashler, 2008). The authors proposed the crowd-within as a variant for the "wisdom of the crowd." They found that averaging a single individual's responses to repetitions of same question led to better performance than single responses alone. This is what one would expect if a single judgment did not incorporate all of the information people could possibly have about a question. If this is true for assessments of mammograms by expert radiologists, we would expect that averaging a radiologist's ratings of abnormality from two exposures to the same mammogram should result in better accuracy than looking at either rating alone. Note that in this situation, however, unlike Vul and Pashler (2008), participants actually have additional information the second time—they get to see the image again before the second judgment, they are not just asked again. Thus, in this case, the crowd-within effect here could arise from actual new information being incorporated (e.g., the observer might scrutinize different parts of the image), rather than internal sampling.

We find a modest but significant advantage to incorporating both judgments: Averaging radiologists' responses from the first and second time that they saw an image resulted in slightly higher performance in the 30-back condition (AUC = 0.745) compared with single item performance (AUC = 0.716), $t(31) = 3.46$, $p = .002$ (see Fig. 9, left). The effect was not significant in the 3-back condition (joint AUC = .712, single AUC = .705), $t(31) = 1.15$, $p = .259$. Unsurprisingly, this effect was not present in novices, since their performance was very poor on both responses (see Fig. 9, right; all $ps > .10$).

Thus, expert performance can be improved (albeit, rather modestly) by averaging more than one response. It remains to be seen whether this benefit would occur if radiologists were offered unlimited time to process each image, rather than the 3 seconds in the current study. The limited viewing time here may have particularly enhanced radiologists' ability to extract new information in the second viewing of the mammogram.

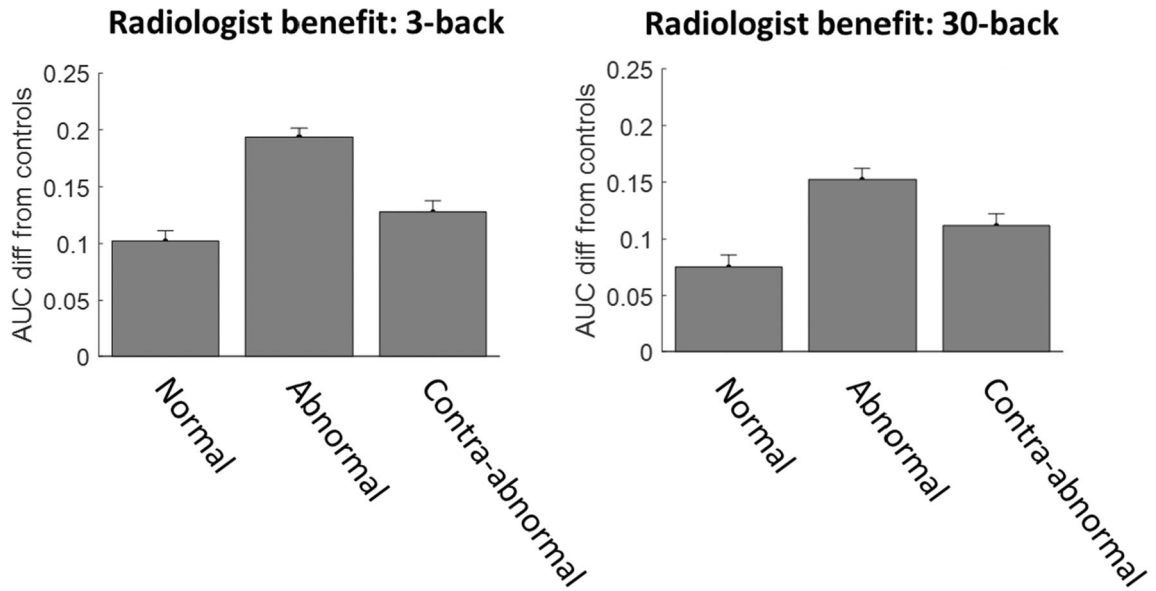

Fig. 8 Using novices as a baseline to account for image similarity, there were robust abnormality memory benefits for radiologists at both 3-back and 30-back

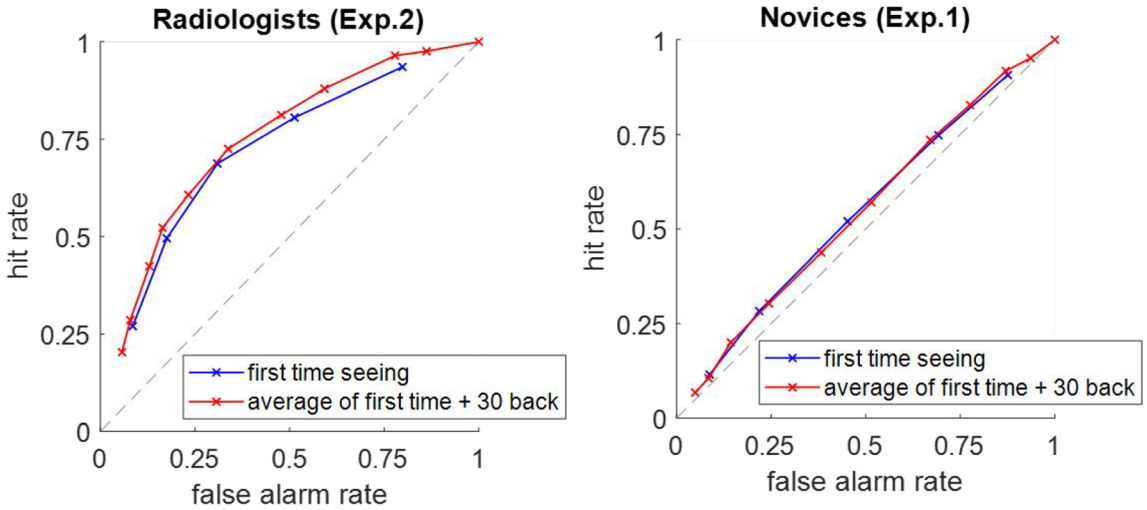

Fig. 9 Crowd-within analysis. Left: Radiologists (Exp. 2). The blue line is the ROC for distinguishing focally abnormal mammograms versus normal mammograms when the radiologists first see the image. The red line is the average of the first time seeing it and their responses seeing it at 30-back. Right: Novices (Exp. 1). Once again, the blue line is the ROC for distinguishing focally abnormal mammograms versus normal mammograms when novices first see the image. The red line is the average of the first time seeing it and their responses seeing it at 30-back

General discussion

In the current study, we examined memory performance by nonexpert novices and expert radiologists for normal versus abnormal mammography images as a case study in understanding the role of schemas, distinctiveness, and expertise in memory. To do so, we relied on ROC analysis, designed to properly measure memory independent of differences in response criteria and to take into account both enhanced memory for seen items as well as the possibility of false alarms.

First, we looked at how confident and how competent novice and expert observers were at classifying medical images as either normal or abnormal. Unsurprisingly, radiologists were much better than novices at this task. Novices did show some ability to distinguish abnormality, although this appeared to be largely the result of a few salient images.

Second, we examined our main question of interest: memory for the images. In Experiment 1, we examined memory for mammograms in novices, who have none of the expertise or schemas needed to process these images. We found poor performance overall, as well as a small normality benefit in novice participants' memory, which could be explained by the greater image dissimilarity of normal images. Thus, Experiment 1 (on novices) gave us not only a baseline for memory performance, but also an understanding of the intricacies of our image set, showing that some abnormal images were quite salient, and that our normal images were more dissimilar from each other.

Even though the normal images in our set were more visually distinctive, in Experiment 2, we found that radiologists had better memory for abnormal images, and had far superior memory performance to novices. This gives insight into how expertise changes memory: not only enhancing the encoding of normal items but also enhancing the distinctiveness of abnormal items. Thus, while experts might have access to perceptual encoding benefits, distinctiveness and/or schemas/chunking to enable them to outperform novices, our finding of an extra benefit of expertise for abnormal images is most consistent with a special role of distinctiveness. For experts, the abnormal images have unique features that make them distinct from other items in memory; whereas for novices, these features are not appreciated and so these images are just like any other image. For example, one possibility is that rather than encoding the entire image, in the case of abnormal images, radiologists specifically encode the abnormality and not the rest of the image into memory. This might reduce the load on memory for that image and might make the memory trace for that image more distinctive.

Broadly speaking, then, we find strong evidence for a role of schemas and distinctiveness in memory, even after taking into account false memory and the possibility of response criterion shifts: We find experts significantly outperform novices, and that memory for abnormal cases with a visible, focal lesion is better than memory other images. There was no evidence for a memory benefit for "abnormal" contralateral cases.

Measuring memory: False alarms and ROC analysis

In the current studies, we used ROC analysis to examine memory. This is because, in previous work, it has often been unclear if benefits for schema-consistent information like those reported in experts are, in fact, improvements in memory, as opposed to changes in response criteria. To determine whether memory has actually improved, it is not adequate to simply find a reliable increase in the rate with which observers

correctly report having been exposed to some piece of information (the true positive, or "hit" rate). The observer could simply be saying "yes, I have seen it" more often. This would produce an increase in false-positive (or false-alarm) errors. In the context of memory research, these false-positive errors can be seen as a form of false memory. In theory, signal detection models and measures like d' can distinguish between these two, but in practice, the prerequisites for d' to properly adjust for response bias (equal variance; zROC slopes = 1.0) are almost never present in recognition memory contexts, and were not present here. Thus, ROC analysis is needed to distinguish between the difference in the ability to remember as opposed to criterion shifts, which would reflect an increase tendency of observers to say that they remember (e.g., Wixted & Mickes, 2015).

Is false memory a true concern? In fact, previous work has found that organizing information in memory via schemas can have both positive and negative consequences—and in particular, does often increase false alarms, making it difficult to tell whether memory is genuinely improved. In particular, while greater understanding—as in expertise—may allow encoding of only the relevant details, reducing memory load, it may also cause us to falsely remember information that was not present (e.g., Owens et al., 1979). For example, in recognition tests, people are more likely to false alarm to schema-consistent relative to schema-inconsistent lures. They would be more likely to falsely report seeing books in a graduate student's office than inconsistent objects like a piece of tree bark or a pair of pliers (Brewer & Treyens, 1981; Lampinen et al., 2001). And while participants are more likely to correctly remember schema-consistent information in a briefly presented scene (Biederman et al., 1982; Brewer & Treyens, 1981), they are also more likely to falsely remember such information (e.g., Hollingworth & Henderson, 2003; Pedzek et al. 1989).

Thus, measuring fully ROCs—rather than attempting to infer how response bias would change performance using measures like A', d', or hits minus false alarms—often reveals surprising answers about memory, particularly in situations like expertise and consistent/inconsistent items where it is known that both hit and false-alarm rates are affected. For example, Dougal and Rotello (2007) used ROC analysis to show that the well-known effect of "improved memory" for emotional words compared with neutral words is a response bias effect, not a true difference in memory between the words. Similarly, Mickes et al. (2012) showed in the domain of eyewitness memory that sequential lineups, which reduce both false alarms and hit rates relative to simultaneous lineups, are actually inferior to simultaneous lineups, contrary to a large body of literature suggesting the opposite (e.g., Wells et al., 2011), as the major "benefit" arises simply from a response criterion shift, not a change in memory strength.

Thus, the current experiments provide unique evidence that expertise and distinctiveness that is apparent only to experts do, in fact, enhance memory—and that this is not just a response criterion shift.

What explains radiologists outperforming novices

Consistent with a wide variety of work on expertise, we find that expert radiologists outperform novices in remembering mammograms. One likely possibility is that this occurs because of experts knowledge about these images: they have relevant knowledge that allows them to understand these images in a way novices do not, and likely have perceptual expertise built into their visual system from years of experience (e.g., in the form of greater holistic processing; e.g., Richler et al., 2011). In particular, for an expert, the abnormal images would have an added attribute (that mass, that calcification), learned over years of experience, that would help to distinguish the item in memory.

However, in the current study, we did not attempt to directly match our experts to our novices. Our novice pool was sampled from the internet, which is much more broadly representative of the demographics of the United States than an undergraduate population (e.g., Difallah et al., 2018), but still likely differs in a number of ways from our radiologists (in demographic and socioeconomic factors, as well as motivation to focus on mammogram images). Thus, Experiment 1 should be taken as only an approximate baseline: it revealed important image features in our stimulus set, and points to the possibility of strong expertise effects, but does not directly confirm these are based solely on knowledge rather than other factors.

Memory and abnormality judgments in radiologists

Previous work has found mixed results when investigating memory improvements in radiologists. For example, Hardesty et al. (2005) investigated radiologists' long-term memory for medical images presented months later and found that none of the radiologists remembered cases that they had read previously. Evans et al. (2016) found mixed results when investigating whether abnormality improves memory in expert observers, including radiologists. Our results provide context to these ambiguities, as they suggest that expert radiologists do have stronger memory for abnormal images even in a long-term memory setting and even when response bias is properly taken into account using ROC analysis. However, our long delays were only on the order of minutes, not months, and so it remains unclear how such advantages would last over long durations.

It is worth noting that in the classification task, radiologists performed on average much more poorly than would be expected of radiologists in the clinic with unlimited viewing time (d' = 2.5–3.0, as in D'Orsi et al., 2013). One reason for this might be that each image in our study was only presented for 3

seconds each. For instance, Evans et al. (2013) showed radiologists only a brief glimpse of mammograms and varied timing from 250 ms to 2,000 ms. The respective AUC's for radiologists in their experiment for 500 ms, 1,000 ms, and 2,000 ms viewing times was 0.65, 0.66, and 0.72, respectively. In our experiment with a presentation time of 3,000 ms, we found an AUC of 0.72. Thus, our 3,000-ms presentations resulted in a similar level of performance to the 2,000-ms presentations of Evans et al. (2013), which, while well below what is expected with unlimited viewing time, is consistent with other studies and consistent with viewing time being the main constraint that lead to lower performance.

The "crowd-within" effect in radiologists

Because our study had radiologists answer the same classification question about an image multiple times, we looked at whether averaging radiologists' responses when they judge the same image twice resulted in better performance (a "crowd-within" effect; Vul & Pashler, 2008). We found that radiologist performance improved when averaged across the same image twice compared with either response alone, but only in the 30-back condition and only modestly even then. This indicates that by the time radiologists were presented with the same image 30 images later, they gave a response that is somewhat independent of their first response. This suggests that, under the current experimental conditions, there might be information the radiologists are not using the first time they see an image—and that the opportunity to see the image again allows the radiologist to glean additional useful information. Future studies might determine whether such benefits persist when experts are given unlimited time to process the images as well as whether this effect can be made larger with an even longer delay between the first and second presentation of an image (as found by Vul & Pashler, 2008).

The "gist" of abnormality

Given the Evans et al. (2016) finding that there is a "gist of abnormality" present in the contralateral breast when no localizable abnormality is present, we were interested to know whether these contralateral-abnormal images had any advantage over normal images in expert memory. We found no such evidence. In our experiment, we also found no difference in the classification of abnormality between contralateral-abnormal images compared with normal. While at first this might seem to contradict earlier work, there are a number of methodological differences that make it difficult to compare our results directly with Evans et al. (2016). It is possible that we did not find this result because we presented images for longer encoding time (3,000 ms). Typical stimulus exposure in mammogram "gist" studies has been less than a second; 500 ms is typical. It is possible that presenting images for longer encoding times might actually obscure the gist information—overwriting an initial "gist" impression with more semantic or meaningful information. Recall, also, that our radiologists were not informed about gist and likely reserved their "abnormal" ratings for cases where they could localize a lesion. It is possible that we would observe a contralateral-abnormal effect even at long encoding times if we explicitly directed participants to look for a more general abnormal texture or gist. Given these methodological differences, the current study cannot be readily compared with Evans et al. (2016). However, this seems to be a promising avenue for future work.

Conclusion

Using radiologists as a case study, we find an advantage for memory in experts as well as an advantage for abnormal images—even when properly measuring memory via ROC analysis. This is broadly consistent with the literature on schemas. Our findings have important implications for both applied fields that utilize expert intelligence in making inferential decisions as well as theoretical fields interested in how memory changes with expertise. In particular, understanding the structure of memory in experts is critical in situations where decisions need to be made by people who have significant expertise.

Acknowledgements All persons who contributed to this project are authors on the final paper.

Authors contributions All authors contributed to the original hypothesis, and read and approved the final manuscript.

H.M.S. contributed to data collection, data analysis, and writing of the manuscript. T.F.B. contributed to data collection, data analysis, and editing of the manuscript.

J.M.W. provided general guidance and contributed to editing of the manuscript.

Funding This research was supported by NSF BCS-1829434 to T.F.B.

Data Availability For data and material, please contact the corresponding author.

Declarations

Ethics approval and consent to participate All participants gave informed consent. For all experiments in this study, informed consent procedures were approved by the Institutional Review Board of the University of California, San Diego.

Consent for publication Not applicable.

Competing interests n/a

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge University Press.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Bilalić, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many faces of expertise: Fusiform face area in chess experts and novices. *Journal of Neuroscience*, *31*(28), 10206–10214. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., Alvarez, G., & Störmer, V. (2019). The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli. *Journal of Neuroscience, 39*(6) 1100–1108. https://doi.org/10.1523/JNEUROSCI.1693-18.2018

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*(2), 207–230. https://doi.org/10.1016/0010-0285(81)90008-6

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Calkins, M. W. (1894). Experimental. *Psychological Review, 1*(3), 327–329. https://doi.org/10.1037/h0065852

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 94–107. https://doi.org/10.1037/0096-1523.35.1.94

de Groot, A. D. (1946). *Het denken van den schaker: een experimenteel-psychologische studie* [The thinking of the chess player: An experimental-psychological study]. Noord-Hollandsche Uitgevers Maatschappij.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of Mechanical Turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135–143). ACM. https://doi.org/10.1145/3159652.3159661

D'Orsi, C. J., Getty, D. J., Pickett, R. M., Sechopoulos, I., Newell, M. S., Gundry, K. R., Bates, S. R., Nishikawa, R. M., Sickles, E. A., Karellas, A., & D'Orsi, E. M. (2013). Stereoscopic digital mammography: Improved specificity and reduced rate of recall in a prospective clinical trial. *Radiology, 266*(1), 81–88. https://doi.org/10.1148/radiol.12120382

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

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, 58–51,* ii, 32.

Engle, R. W., & Bukstel, L. (1978). Memory processes among bridge players of differing expertise. *The American Journal of Psychology, 91*(4), 673–689. https://doi.org/10.2307/1421515

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Evans, K. K., Cohen, M. A., Tambouret, R., Horowitz, T., Kreindel, E., & Wolfe, J. M. (2011). Does visual expertise improve visual recognition memory? *Attention, Perception, & Psychophysics*, *73*(1), 30–35. https://doi.org/10.3758/s13414-010-0022-5

Evans, K. K., Georgian-Smith, D., Tambouret, R., Birdwell, R. L., & Wolfe, J. M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin & Review, 20*, 1170–1175. https://doi.org/10.3758/s13423-013-0459-3

Evans, K. K., Haygood, T. M., Cooper, J., Culpan, A.-M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(37), 10292–10297. https://doi.org/10.1073/pnas.1606187113

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355. https://doi.org/10.1037/0096-3445.108.3.316

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–573. https://doi.org/10.1038/9224

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82–94. https://doi.org/10.1037/xge0000129

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted positions: Implications for the theory of expertise. *Memory & Cognition, 24,* 493–503. https://doi.org/10.3758/BF03200937

Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 59–109). Academic.

Hardesty, L. A., Ganott, M. A., Hakim, C. M., Cohen, C. S., Clearfield, R. J., & Guret, D. (2005). "Memory effect" in observer performance studies of mammograms. *Academic Radiology, 12*(3), 286–290. https://doi.org/10.1016/j.acra.2004.11.026

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hollingworth, A., & Henderson, J. M. (2003). Testing a conceptual locus. *Memory & Cognition*, *31*(6), 930–940. https://doi.org/10.3758/BF03196446

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. https://doi.org/10.1038/nature06713

Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1211–1222. https://doi.org/10.1037/0278-7393.27.5.1211

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 212–228. https://doi.org/10.1037/0278-7393.5.3.212

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 54–65. https://doi.org/10.1037/0278-7393.12.1.54

McWeeny, K. H., Young, A. W., Hay, D. C., & Ellis, A. W. (1987). Putting names to faces. *British Journal of Psychology*, *78*(2), 143–149. https://doi.org/10.1111/j.2044-8295.1987.tb02235.x

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376. https://doi.org/10.1037/a0030609

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19,* 131–150. https://doi.org/10.3758/BF03197110

Owens, J., Bower, G. H., & Black, J. B. (1979). The "soap opera" effect in story recall. *Memory & Cognition, 7,* 185–191. https://doi.org/10.3758/BF03197537

Pedzek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectations. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 587–595. https://doi.org/10.1037/0278-7393.15.4.587

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*(3), 646–668. https://doi.org/10.1016/j.jml.2007.08.004

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science*, *20*(2), 129–134. https://doi.org/10.1177/0963721411402472

Shiffrin, R. M., Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. https://doi.org/10.3758/BF03209391

Vincente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(1), 33–57. https://doi.org/10.1037/0033-295X.105.1.33

Voss, J. F., Vesonder, G. T., & Spilich, G. J. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior, 19,* 651–667. https://doi.org/10.1016/S0022-5371(80)90343-6

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Watson, T. L., & Robbins, R. A. (2014). The nature of holistic processing in face and object recognition: Current opinions. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00003

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2011). *A test of the simultaneous vs. sequential lineup methods an initial report of the AJS National Eyewitness Identification Field Studies.* https://mn.gov/law-library-stat/archive/urlarchive/a100499.pdf

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science, 14*(1), 6–9. https://doi.org/10.1111/j.0963-7214.2005.00324.x

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

Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, *4*(4), 329–334. https://doi.org/10.1016/j.jarmac.2015.08.007