"It's not a tumor": A framework for capitalizing on individual diversity to boost target detection

Journal:	*Psychological Science*
Manuscript ID	PSCI-17-1243.R2
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Corbett, Jennifer; Brunel University London, College of Life and Health Sciences, Division of Psychology; Brunel University London, Center for Cognitive Neuroscience Munneke, Jaap; Brunel University London, College of Life and Health Sciences, Division of Psychology; Brunel University London, Center for Cognitive Neuroscience
Keywords:	Visual Search, Signal detection, Wisdom of crowds

Averaging diversity boosts detection

"It's not a tumor": A framework for capitalizing on individual diversity to boost target detection

Jennifer E. Corbett[1,2] * & Jaap Munneke[1,2]
[1] *Brunel University London; College of Life and Health Sciences, Division of Psychology*
[2] *Brunel University London; Center for Cognitive Neuroscience*

***Address correspondence to:**
Jennifer E. Corbett
Brunel University London
Division of Psychology, MJ-122
Kingston Lane, UB8 3PH
Uxbridge, London
Phone: +44 756 309 3475
jennifer.e.corbett@gmail.com

Abstract

Even experts routinely miss infrequent targets, such as weapons in baggage scans or tumors in mammograms because the visual system is not equipped to notice the unusual. To date, limited progress has been made towards improving human factors that mediate such critical diagnostic tasks. Here we present a novel framework for pairing individuals' estimates to increase target detection. Using a "wisdom of crowds" approach capitalizing on the visual systems' ability to efficiently combine information, we demonstrate how averaging two, non-interacting individuals' continuous estimates of whether a briefly presented image contained a pre-specified target can significantly boost detection across a range of tasks. Furthermore, we show how pairing individuals' estimates to maximize decorrelated patterns of performance in one task can optimize performance on a separate task. Results make significant advances towards combating severe deficits in target detection, with straightforward applications for maximizing performance within limited pools of observers.

Despite spectacular developments in image processing and automated search algorithms (Liu et al., 2017), humans still perform the majority of critical visual search tasks such as breast cancer and airport baggage screenings. In stark contrast to a growing class of computer vision algorithms which input the entire image in parallel (Itti & Koch, 2001; Serre, Oliva, & Poggio, 2007) the human visual system is only capable of processing a few objects in detail at any given moment (Luck & Vogel, 1997). To circumvent these capacity restrictions, the visual system acts like a "statistician," averaging out redundant and specific information, "filling in" details based on spatiotemporal context, and leaving us with a representation of abstract, statistical regularities to efficiently represent the maximum amount of information (Attneave, 1954; Baddeley, 1997; Barlow, 1961; Kersten, 1987; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001). However, relying heavily on current and prior context means that we are designed not to notice targets that are out of the ordinary (Simons & Chabris, 1999). For example, even experts are prone to missing infrequent targets, such as weapons in baggage scans or tumors in routine mammograms (Wolfe, Horowitz, & Kenner, 2005).

For over a decade since Wolfe and colleagues' initial discovery that we miss approximately 30% of rare targets (Wolfe et al., 2005), researchers have struggled to develop new methods to improve the detection of unusual and important targets (Van Wert, Horowitz, & Wolfe, 2009; Wolfe et al., 2007; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013; Wolfe et al., 2005). Such attempts have mainly been aimed at increasing individual observers' "vigilance," for example, by interspersing bursts of frequent targets into otherwise rare target searches (Wolfe et al., 2007, 2013). Instead of trying to augment the performance of a system equipped to function optimally within its biological and computational constraints, our approach is grounded in observations abstracted from the visual system's inherent structure and function. Although it

may seem that a system with restricted capacity for detail is destined for failure in complex search tasks, these apparent weaknesses can be translated into strengths if we approach this problem from the understanding that the visual system capitalizes on regularities by sampling the most unique, decorrelated aspects of visual input. For example, individual neurons coding for the most discrepant information in an image (Barlow, 1961) mimic the "wisdom of crowds" approach of asking multiple individuals for different ground truth guestimates (Galton, 1907; Hogarth, 1978). Both cases take advantage of regression towards the mean, such that the more independent the samples, the more information that is accrued from averaging, and therefore the greater the reduction in error. Along these lines, studies of the "crowd within" have focused on similar methods for averaging guesses from a single individual to improve performance on a range of cognitive and perceptual tasks (Corbett, Fischer, & Whitney, 2011; Herzog & Hertwig, 2009; Vul & Pashler, 2008). All of these findings point to the intriguing and potentially life-saving possibility of even greater benefits of independence from combining two estimates from *different* observers. To these ends, we developed and tested the novel framework illustrated in Figure 1 for pairing individuals' continuous estimates of target presence to increase detection performance across a range of tasks.

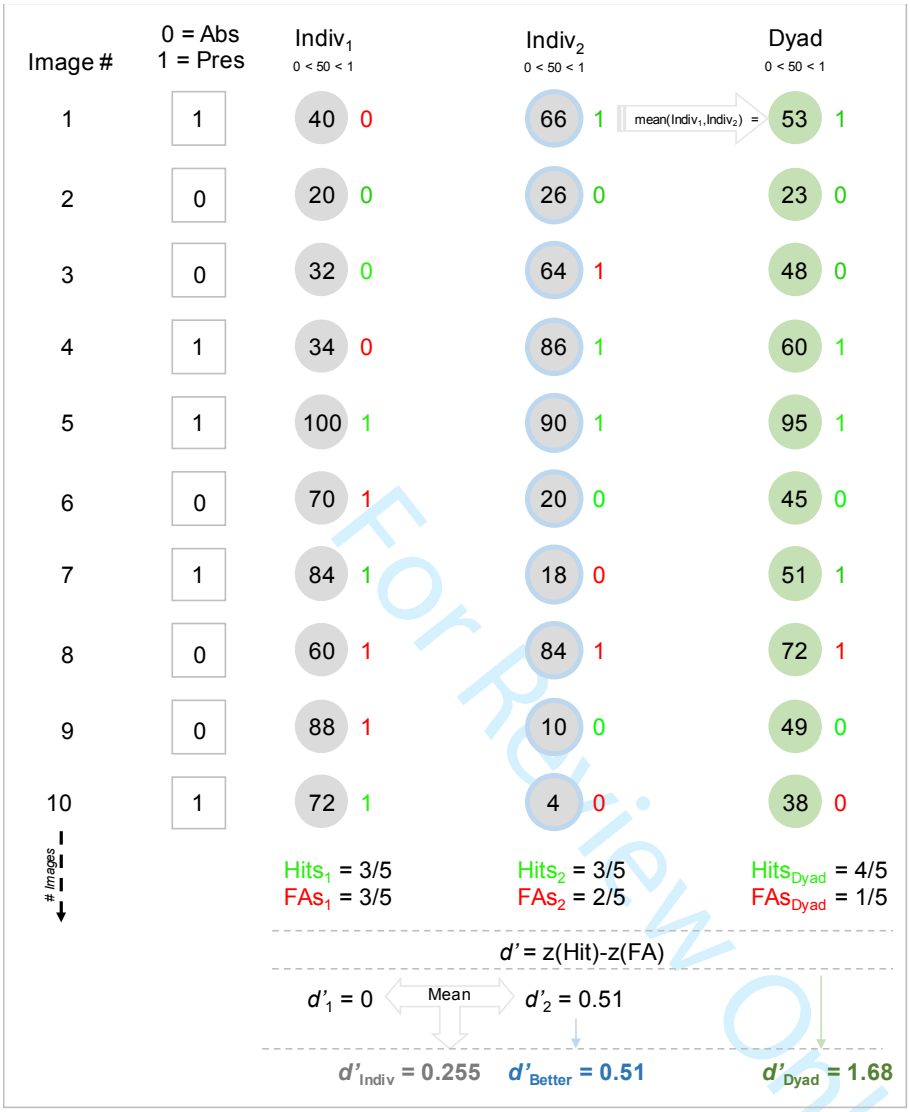

Figure 1. Framework for improving performance by averaging individuals' estimates. For each image, individuals provide a continuous estimate (from 0=No, to 100=Yes) about whether a pre-specified target is present. Individuals perform the task independently. Each row represents hypothetical responses to a single image. The first column corresponds to the specific image number (Image #), and the number inside the square in the second column indicates whether a target is present (1) or absent (0) in that image. Numbers inside gray circles in the third and fourth columns represent the corresponding continuous responses of the two observers in the hypothetical dyad, and red and green 0's and 1's to the right of these estimates represent how that continuous response translates into a binary response. Responses less than 50 are scored as 0=Absent, and responses greater than 50 are scored as 1=Present. Green numbers indicate correct responses and red numbers indicate incorrect responses, such that green 1's correspond to Hits and red 1's correspond to False alarms (FAs). Numbers inside green circles in the final column represent the dyad's average estimate (the average of $Indiv_1$ and $Indiv_2$'s continuous responses), and the red and green 0's and 1's to the right of these estimates represent the dyad's corresponding binary response. Each individual's *d'* (d'_1 and d'_2) is calculated by subtracting their normalized FA rates (FAs_1 and FAs_2) from their respective normalized Hit rates ($Hits_1$ and $Hits_2$), and the dyad's *d'* (d'_{Dyad}; green text) is calculated by subtracting the dyad's normalized FA rate (FAs_{Dyad}) from the dyad's normalized Hit rate ($Hits_{Dyad}$). The average *d'* for the dyad (d'_{Indiv}; gray text) is calculated as the average of the two individuals' *d'*s (the mean of d'_1 and d'_2). $Indiv_2$ (gray circles outlined in blue) had a higher *d'* than $Indiv_1$ ($d'_2 > d'_1$), and is therefore used as the *d'* of the better observer in the dyad (d'_{Better}; blue text). Overall, averaging individuals' estimates results in lower error rates (Hits and FAs) and therefore a higher rate of target detection (*d'*) than the average of the individuals' detection rates or the detection rate of the better individual in the dyad.

Using this framework, we conducted two experiments to test the predictions that: 1) pairing individuals' estimates will result in better performance compared to the performance of either individual (Experiments 1 & 2), and 2) pairing estimates from individuals with the most decorrelated performance on one task will result in the greatest improvements on a separate task (Experiment 2).

In a first experiment, we investigated how averaging independent estimates from two different individuals might improve target detection. We presented naïve observers with unilateral mammography x-rays verified as having either an abnormality (a tumor) or consisting of normal, healthy tissue. Observers' task was to rate the extent to which they thought a tumor was present on a pseudo-continuous scale ranging from 0 (No) to 100 (Yes), excluding 50. Participants completed two "Mammos" conditions in which either 5% or 50% of the total number of trials contained tumor present images (see Methods and Figure 2a for more details). Given the noted benefits of combining independent information (Barlow, 1961; Corbett et al., 2011; Galton, 1907; Herzog & Hertwig, 2009; Hogarth, 1978; Vul & Pashler, 2008), we predicted that the average of two observers' estimates would result in higher *d'* values (a signal detection measure that accounts for both types of possible miss and false alarm errors) than either of their individual estimates.

Methods

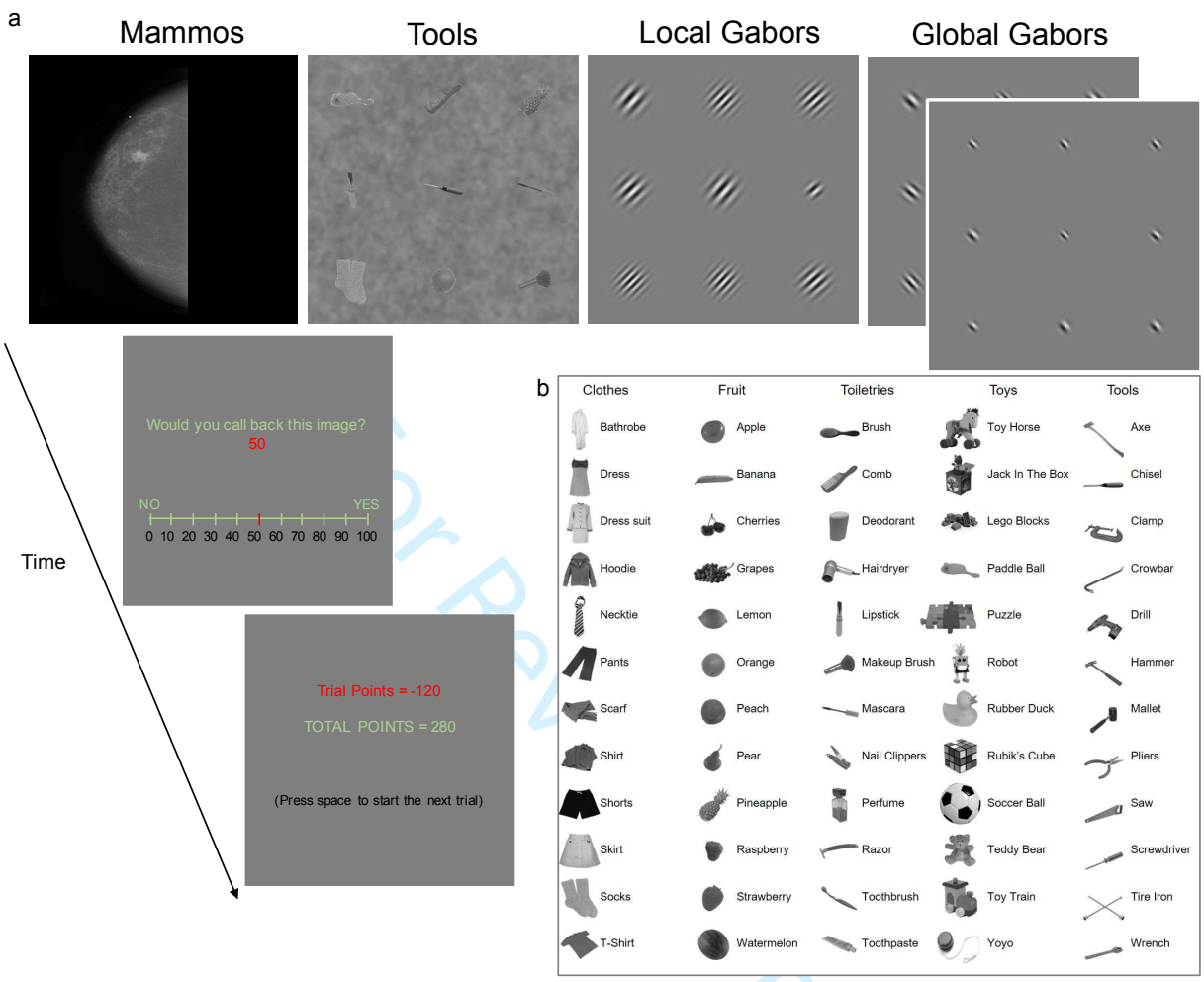

Fig. 2. a. Trial Sequence. On each trial, observers were presented with a search display for 500 ms, followed by a response slider to indicate the extent to which they thought the image should be called back for containing a pre-specified target. Feedback about the points earned on the trial and the total points accumulated over the entire duration of the experimental block was displayed after a response was made. *Mammos (Exp. 1):* Observers were presented with a unilateral mammogram and their task was to decide whether a tumor was present. They completed two conditions in which either 5% or 50% of the trials contained tumor present images (5% Mammos and 50% Mammos) in a blocked design. *Tools (Exp 2):* In the Tools task, observers were presented with a display of nine objects and their task was to decide if a tool target was present. They completed two conditions (blocked) with a tool target present on either 5% or 50% of trials (5% Tools and 50% Tools). *Local Gabors (Exp 2):* On each trial in the Local Gabors condition, observers were presented with an image of nine Gabors, and their task was to decide whether or not to call the image back for containing a target Gabor that was larger or smaller than all the other Gabors. *Global Gabors (Exp 2):* On each trial in the Global Gabors condition, observers were presented with two successive displays of nine Gabors, temporally separated by a 1000 ms blank screen, and their task was to decide whether to call back the sequence for having a "target" difference in global mean size between the first and second images. In both Gabors conditions, 50% of trials contained a target. **b. Objects used in the Tools tasks.** As in Wolfe and colleagues (Wolfe et al., 2005), gray-scaled objects (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Konkle, Brady, Alvarez, & Oliva, 2010) from five different stimulus categories (including tools) were used to create the stimulus displays in the Tools task. Each category consisted of 12 exemplars.

Participants: There were 25 participants in Experiment 1, and 23 participants in Experiment 2. All participants were students at Bilkent University, participated in only one of the experiments, had normal or corrected-to-normal vision, and volunteered to take part in the experiment either for course credit or money. In each experiment, the participant with the highest points won an additional 50 Turkish Lira (see *Points/Feedback*). Seven participants were dismissed from each experiment for failing to pass a pre-established performance threshold during practice blocks (see *Points/Feedback)*. Only the data from the remaining 18 observers in Experiment 1 (11 female, mean age=20.56 years, age range=19-25), and 16 observers in Experiment 2 (14 female, mean age=21.25 years, age range=18-25) were included in subsequent analyses. We had to restrict our sample size to 16 participants in Experiment 2, as this was the maximum number of individuals for which we could generate all possible paired combinations given computational limitations (see *Decorrelated Dyads*). All procedures and protocols were in accordance with the guidelines of Bilkent University's ethical review board.

Task & Trial sequence: Participants were instructed to act like radiologists (Experiment 1) or airport baggage screeners (Experiment 2), and estimate whether they detected the presence of a target item from a pre-specified category (e.g., a tumor, a tool) in a briefly presented image. On each trial in Experiment 1, participants were presented with a unilateral mammogram for 500 ms. On each trial in Experiment 2, participants were presented with an image of nine objects for 500 ms. In both experiments, the target screen was followed by a response slider ranging from 0-100 (adapted from Evans, Georgian-Smith, Tambouret, Birdwell, & Wolfe, 2013; Evans, Haygood, Cooper, Culpan, & Wolfe, 2016). Participants used the mouse to adjust the slider to indicate whether to call back the image based on whether they detected a target (Figure 2a). To

implement a modified version of a two-alternative forced-choice task (2AFC), we restricted the slider responses so that participants could choose any value from 0-100 except 50. After they finished adjusting the slider, participants were presented with feedback on their points for the given trial and their total points for the block until they pressed the space bar to continue to the next trial.

Stimuli & Procedure: All participants performed the tasks individually and independently, in a dimly lit room. Participants had no knowledge that their performance would be paired with another individual offline after the experiment, nor did they receive any feedback that had any relationship to the performance of any other participant. An HP PC was used to present stimuli at a viewing distance of 57 cm, on a 21.3" NEC LCD monitor with a 60-Hz refresh rate and a resolution of 1600 x 1200 pixels. MATLAB (Version 2016a) and the Psychophysics Toolbox (PTB-3; Brainard, 1997; Pelli, 1997) controlled all presentation, timing, response functions, and data collection. See the Supplementary Material available online for details of the training and practice blocks associated with each experiment.

Experiment 1: Each participant completed two experimental conditions (5% and 50% target prevalence) in a single, one-hour experimental session. The two conditions were run in separate blocks, and the order of blocks was counterbalanced over participants. In each condition, observers were presented with a unilateral mammogram centered in the middle of a black screen. Mammograms were acquired from the University of South Florida Digital Database for Screening Mammography (DDSM; Heath et al., 1998; Heath, Bowyer, Kopans, Moore, & Kegelmeyer, 2001), and presented in native pixel resolution (subtending

approximately 10^{o} x 18^{o} of visual angle). There were 400 unique images in each condition. In the 5% condition, a tumor was present in 20 of the 400 images. In the 50% condition, a tumor was present in 200 of the 400 images. All participants saw the same set of 400 images in the 5% condition and another set of 400 images in the 50% condition. The order of the images in each set was randomized, and each participant saw each image only once throughout the experiment. See the Supplementary Material available online for a detailed description of the mammogram stimuli and the corresponding supplementary image database.

Participants completed two stages of training before beginning the experimental blocks. First, they were presented with 50 of the target present mammograms and 50 of the target absent (healthy) mammograms, in random order. Participants were not required to make a response, but only to study the images so that they became familiarized with typical target present and absent images. Each image remained on the screen until the participant pressed the space bar. If the image contained a target, the original mammogram was re-presented with an overlay showing the region identified as abnormal circled in red. The 100 trials in the second training stage were identical to trials in the main 50% experimental condition, except: 1) the training images were presented simultaneously with the response slider and remained on the screen until the participant responded, and 2) if the image contained a target, the image was re-presented with the corresponding overlay of the abnormality. None of the images used in the training and practice blocks were used in the experimental blocks.

Experiment 2: Each participant completed four experimental conditions (Local Gabors, Global Gabors, 5% Tools, and 50% Tools; Figure 2) in a single experimental session lasting

approximately 2 hours. The two Gabors conditions and the two Tools conditions were run in immediate succession, but the order of these two types of conditions and the order of the individual conditions within each type were counterbalanced over observers. In each condition, observers were presented with nine stimuli arranged equidistantly in a 3 x 3 grid, subtending approximately 16^o x 16^o of visual angle and centered in the middle of a gray screen.

- *Local Gabors*: Participants were presented with a total of 288 images, with a target defined as a Gabor that was smaller or larger than the other eight Gabors present in 50% of the images. For each image, the size, spatial frequency, contrast, and orientation of the nine individual Gabors were pseudo-randomly assigned from two possible sets of nine values per dimension (chosen based on the results of previous pilot studies). The small size set ranged in diameter from 1.9^o to 2.1^o in 0.02^o steps, and the large size set ranged from 3.4^o to 3.6^o, also in 0.02^o steps. The low spatial frequency set ranged from 1.2 cycles per degree (cpd) to 2 cpd in 0.1 cpd steps, and the high spatial frequency set ranged from 6 cpd to 6.8 cpd in 0.1 cpd steps. The low contrast set ranged from 73% to 81% contrast relative to the gray background, and the high contrast set ranged from 83% to 91% contrast. The left-tilted orientation set ranged from -49^o to -41^o of tilt from vertical, and the right-tilted set ranged from 41^o to 49^o of tilt from vertical.

  144 target absent images were generated by factorially combining the 2 Sizes x 2 Spatial frequencies x 2 Contrasts x 9 Locations, with two repetitions of each of the combinations. The set of leftward or rightward orientations was randomly selected for the nine Gabors in each image. For each target absent image, a corresponding target present image was constructed from an identical set of parameters except that the

Gabor representing the mean size of the set in the target absent condition was replaced by a Gabor representing the mean size of the opposite set in the target present condition. With the exception of the target size and location in target present images, the individual values in each dimension were randomly assigned to the nine possible locations in each image.

- *Global Gabors*: Participants were presented with a total of 256 pairs of successive images of nine Gabors each. The target was present in 50% of the pairs, and was defined as a difference between the mean size of the items in the first image of nine Gabors and the mean size of the nine Gabors in the second image. There were nine possible sets of nine individual Gabor sizes, constructed by multiplying a base set ranging from 1.3^{o} to 1.7^{o} in 0.05^{o} steps by a constant value from 0 to 8. In target present pairs, the sizes of the Gabors in the first or second image (determined randomly for the pair) were taken from sets 1, 2, 3, 4, 6, 7, 8, or 9. The sizes of the Gabors in the other image were taken from set 5, for a total of eight possible mean size differences between the two images in each pair. In 50% of target absent pairs, the sizes of both images were taken from the same set ("repeated pairs"; 1, 2, 3, 4, 6, 7, 8, or 9). In the other half of target absent pairs, the sizes of Gabors in both images were taken from set 5 ("set 5 pairs"). This manipulation guarded against the tendency to call back the pair of images on a given trial as having a target difference present if either of the images in the pair contained the set 5 sizes. The same spatial frequencies, contrasts, and orientations were used as in the Local Gabors condition. The entire set of 256 image pairs was constructed from two repetitions of the full factorization of

the 8 possible Mean size differences x 2 Present/Absent conditions x 2 Spatial frequencies x 2 Contrasts x 2 Absent pair types (repeated pairs/set 5 pairs). Each target present image had a corresponding target absent image with all of the same parameters except that both target absent images contained the same nine sizes (either repeated pairs or set 5 pairs). As in the Local Gabors condition, the set of leftward or rightward orientations for the nine Gabors was randomly selected for each pair of the 128 target present images, and the same orientation set was used for the corresponding target absent pair of images. In all images in all sequential pairs, the individual values in each dimension were randomly assigned to the nine possible locations.

- *5% Tools*: There were a total of 360 images, each comprised of nine objects. A single tool target was present in 5% of the images, and the other eight items were drawn from one of four categories (Clothing, Fruit, Toiletries, and Toys; Figure 2b). There were 12 objects per category (including Tools), for a total of 60 possible items. The images for individual items were taken from open-source databases published in previous work (Brodeur et al., 2010; Konkle et al., 2010). Each item subtended 2.5^{o} x 2.5^{o} (the average size of the individual Gabors). Individual items were presented at 60% transparency, each item in each image was tilted a random ±15^{o} from vertical in 1^{o} steps, and a pink noise mask was added to the entire image in attempt to better equate the task to a real-world baggage screening task (Wolfe et al., 2005). There were 18 target present images (two per location) and 342 target absent images. For each target present image, there were 19 corresponding target absent images, each

containing the same nine non-tool objects drawn randomly from one of the four non-tool categories. Although the same nine items were used in the target absent images, the individual items were placed randomly in the nine locations in each image. In the corresponding target present image, a tool was randomly selected to replace the non-tool item in the target location, with the locations of the eight other non-tool items randomly determined.

- *50% Tools*: A tool target was present in half of the 360 total images. All parameters were identical to those used in the 5% Tools condition, except that there were 180 target present images (20 per location) and 180 target absent images, with target absent images containing nine randomly selected items from the four non-tool categories and the corresponding target present image containing a tool randomly selected to replace one of the non-tool items and presented in the target location.

Points/Feedback: On each trial, participants earned points for a correct response or lost points for an incorrect response (adapted from Wolfe et al., 2005). In all 50% conditions, participants could earn a maximum of 40 points when there was a target present and they correctly called back the image (i.e., adjusted the slider to a value above 50%; Hit), and 2.5 points when there was no target present and they correctly did not call back the image (Correct rejection). They could lose a maximum of 40 points if there was a target and they did not call the image back (Miss), and 75 points if there was no target and they incorrectly called back the image (False alarm). In the 5% conditions, participants could earn a maximum of 400 points for a Hit, and a maximum of 2.5 points for a Correct rejection. They could lose a maximum of 400

points for a Miss and a maximum of 75 points for a False alarm. We calculated the specific number of points earned or lost on each trial as a function of the extremeness of the given response using the following formula:

$$Points = (x * MaxPoints) * (1-((2*Response)/100))$$

where x = -1 for signal present trials, x = 1 for signal absent trials, MaxPoints was the maximum number of points that could be earned or lost, and Response was the participant's estimate between 0 and 100 made via the slider adjustment on the trial.

In replicating the rare target deficit originally reported by Wolfe and colleagues (2007, 2005), it was necessary to implement a points system that prevented participants from simply responding 0 on each trial in the 5% prevalence condition (which would result in 95% accuracy), or simply hovering in the middle of the response range without actually performing the detection task. Along these lines, the points system allowed us to better equate participants' response criteria (tendency to say yes or no more frequently) across the two types of conditions (Healy & Kubovy, 1981; Wolfe & Van Wert, 2010), such that there was a greater penalty for a bias to respond "YES" in the 50% conditions when targets were frequent, but a greater penalty for a bias to respond "NO" in the 5% conditions when targets were rare. Participants were informed of the specific points for each type of response at the start of each block, and required to have a positive score (points > 0) at the end of every block.

Results

Dyads versus individuals: To gauge whether pairing individuals' estimates yielded improved detection performance, we first paired the 18 (Experiment 1) and 16 (Experiment 2) participants in n-1 unique arrangements of pairings. Each arrangement included all 18 or 16 participants. Only the pairings of the participants into nine (Experiment 1) and eight (Experiment 2) dyads was different over arrangements, such that each possible dyad was present only once in the resultant 17 (Experiment 1) and 15 (Experiment 2) arrangements. For example, the dyad formed by participants 1 and 2 only appeared in one of the 17 possible arrangements of nine dyads in Experiment 1. In other words, all participants were included in each arrangement of pairings, but each dyad was unique across all pairings such that the combination of the two individuals forming a dyad was never repeated (see Table S1 for an illustration of this procedure). As we planned the statistical comparisons illustrated in Figure 3 and Table S2 between three groups, we conducted our analysis using only these unique arrangements of pairings based on multiple combinations of the individual participants.

As illustrated in Figure 3a, averaging individual estimates from two observers improved signal detection (*d'*) over estimates obtained from the corresponding individuals in both the 5% Mammos condition (12.89% average improvement) and the 50% Mammos condition (10.81% average improvement). In both conditions, pairing estimates also improved performance relative to the performance of the better individual in each dyad. We conducted a follow-up analysis to explicate how pairing estimates affected both types of miss and false alarm errors. Although pairing always decreased both misses and false alarms compared to estimates made by the corresponding individuals, there was a noteworthy trade-off between misses and false alarms in the 5% Mammos condition, suggesting that estimates made by the better of the two observers in each dyad were less prone to misses whereas dyad estimates were less likely to result in false

alarms (Figure 3b&c). In all reported analyses, paired comparisons refer to two-tailed t-tests that were Bonferroni-corrected for multiple comparisons. Table S2 in the Supplemental Material available online presents the full results of all planned comparisons.

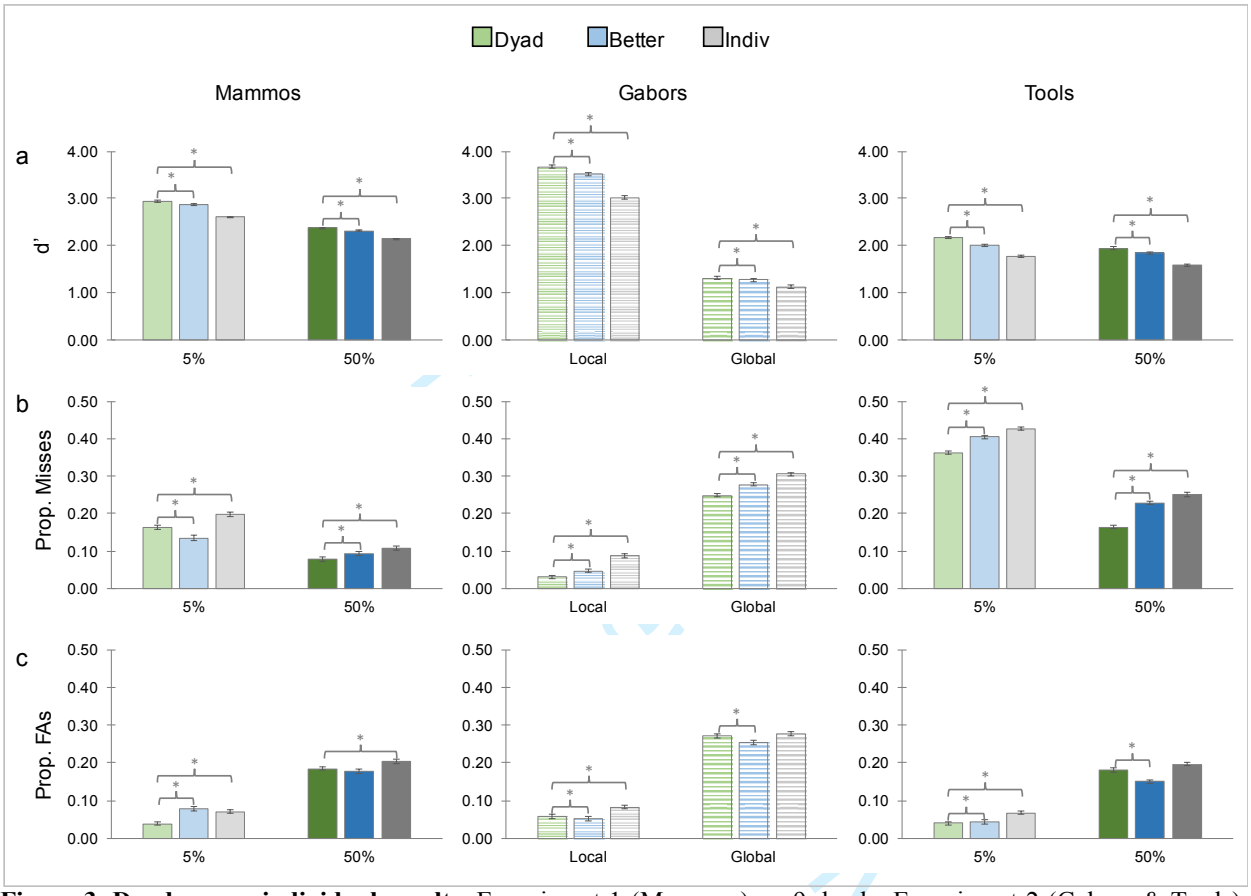

Figure 3. Dyad versus individual results. Experiment 1 (Mammos): n=9 dyads; Experiment 2 (Gabors & Tools): n=8 dyads. **a.** *Detection performance (d'):* For each task, dyads' (green) average *d'* was significantly higher compared to the average *d'* of the corresponding individuals (gray) as well as the *d'* of the better of the two individuals (blue). **b.** *Misses:* Also for each task, the miss rate was significantly lower for dyads compared to individuals, and for dyads compared to the better of the two individuals, with the exception of the 5% Mammos condition. **c.** *False alarms:* False alarms were significantly lower for dyads versus individuals in all tasks except the Global Gabors and 50% Tools conditions, false alarms were significantly lower for dyads versus the better of the two individuals in the 5% Mammos condition, and false alarms were significantly greater for dyads versus the better of the two individuals in the Global Gabors and 50% Tools conditions. Asterisks represent significant differences, with $\alpha = 0.0125$ using the Bonferroni correction for four multiple comparisons. Error bars represent the 95% within-subjects confidence intervals for the two-way interaction depicted in each graph (Loftus & Masson, 1994).

Decorrelated dyads: After confirming the significant benefits of pairing individuals' independent estimates of the likelihood of tumor presence in mammography images, we sought to generalize this benefit using other types of visual search tasks in Experiment 2. In particular,

we first replicated the results of Experiment 1 using a tool detection task in Experiment 2 that is typically used with naïve observers as a proxy for weapons detection in baggage scans (Fleck & Mitroff, 2007; Wolfe et al., 2005; Figure 2). Given that previous findings suggest greater benefits of averaging as estimates become more independent (Corbett et al., 2011; Herzog & Hertwig, 2009; Hogarth, 1978; Vul & Pashler, 2008), we further extended our investigation to examine whether pairing estimates from observers with the most independent, decorrelated patterns of performance in separate Gabors tasks could maximize performance in the main Tools search tasks. Given previous findings that suggest detection might be based in part on a global "gist" signal (Drew, Evans, Võ, Jacobson, & Wolfe, 2013; Evans et al., 2013, 2016; Rensink, 2004), we used local and global versions of the Gabors task (Figure 2). This resulted in 4 main conditions: Local 5%, Global 5%, Local 50%, and Global 50%. We expected the improvement in *d'* from averaging dyads' performance in the Tools tasks to increase with their decreasingly correlated performance in the Gabors tasks.

Similar to real-world situations like airport baggage screenings where a pair of security officers typically work at one of several stations, we sought to optimize overall *group* performance in the Tools tasks by pairing individuals into eight dyads, such that performance on the Gabors task was maximally decorrelated over the entire group of 16 individuals. Importantly, and unlike the analysis used in both Experiments 1 and 2 to compare the performance of dyads and individuals (Figure 3 and Table S2), we only compared dyad performance as a function of correlation on the Gabors tasks and did not restrict this analysis to only one unique occurrence of a particular dyad. Instead, we began by constructing all possible pairings of the 16 individuals using a custom-written recursive MATLAB function, such that one set of the eight dyads was

distinct relative to any other combination (the only consideration was the uniqueness of the dyads within a combination relative to the dyads in all previous combinations. Dyads could repeat over different combinations of pairings, but not within the same combination.). For n=8 possible dyads of the 16 observers, this resulted in

$$\frac{(2n)!}{n!} * \frac{1}{2^n} = 2{,}027{,}025 \text{ unique combinations of 16 observers into 8 dyads}$$

Using this approach, the number of unique combinations increased exponentially with each additional dyad, such that standard computational power limited our analysis to a maximum of 16 individuals grouped into eight dyads.

Next, separately for the Local and Global Gabors tasks, we calculated the average correlation between dyads' responses for the individual images (the 288 images in the Local Gabors condition and the 256 images in the Global Gabors condition) for each set of 2,027,025 combinations. More specifically, the average correlation in responses for each set of eight dyads was found by summing the eight correlations between each pair of individuals ($Indiv_1$ and $Indiv_2$), then dividing by the total number of (8) dyads. After finding the average correlation for each set of combinations, the combinations were sorted from most decorrelated to most correlated. We then used these combinations on each of the Local and Global Gabors tasks to compute dyad performance for each of the 2,027,025 combinations in the 5% and 50% Tools tasks, resulting in four main conditions (Local 5%, Global 5%, Local 50%, and Global 50%). Finally, we averaged the data from the 2,027,025 combinations over 9,009 bins of 225 pairings each to facilitate graphical and statistical comparisons.

Although participants' responses were not speeded, their average response times for both rare (5% Tools: *M*=2020 ms, *SD*=1332 ms) and frequent (50% Tools: *M*=1902 ms, *SD*=1076 ms) target detection were similar to those previously reported by Wolfe and colleagues (2005).

After first replicating the improved rates of target detection for dyad versus individual estimates observed in the mammogram task in Experiment 1 for both Gabors and Tools conditions in Experiment 2 (E2: 16.06% – 22.14% average improvement from averaging; Figure 3 and Table S2), we tested the prediction that pairing estimates from observers such that their overall performance was maximally decorrelated in the Gabors tasks would result in the greatest improvements from averaging in the Tools tasks. In support of this hypothesis, we observed a strong negative correlation (r=-0.972), such that *d'* values on the 5% Tools task increased as the average correlation over a given set of pairings in the Global Gabors task decreased (Global 5%). On the contrary, there was no evidence of a relationship between *d'* on the 5% Tools task and the average correlation over a set of pairings in the Local Gabors task (Local 5%; r=0.196). Improved performance in the 50% Tools task was also well-predicted by observers' decorrelated performance in the Global (Global 50%; r=-0.937) and Local (Local 50%; r=-0.972) Gabors tasks. Taken together, these patterns of results in Figure 4a suggest that pairing estimates from individuals with decorrelated performance on a simple global detection task can maximize performance in a separate, more complex rare target detection task, whereas pairing estimates from individuals with the most decorrelated patterns of performance regardless of the local or global scale of the simple task may optimize the detection of frequently occurring targets in a more complex task.

These conclusions are further supported by the different patterns of improvement in miss and false alarm rates in the two Tools tasks that resulted from pairing the most decorrelated observers in the two Gabors tasks. In the 5% Tools condition where dyads showed the lowest miss and false alarm rates (Figure 3b&c), the more decorrelated the dyads' performance in the Global Gabors condition, the lower their miss and false alarm rates (Figure 4b&c). In the 50%

Tools condition, dyads showed the lowest miss rates (Figure 3b) but higher false alarm rates than the better of the two individuals in each dyad (Figure 3c). Also in the 50% condition, the more decorrelated the dyads performed in the Global and especially the Local Gabors conditions, the greater the improvement in false alarm rates (Figure 4c) with little change in the miss rates (Figure 4b), such that the greatest improvements from averaging decorrelated dyads were observed where improvements were most needed. Further evidence of the benefits of capitalizing on the more fine-grained information available in participants' continuous estimates are illustrated by the inferior results from the same analyses using participants' dichotomized estimates in the Gabors tasks instead of their continuous responses (Figure S1).

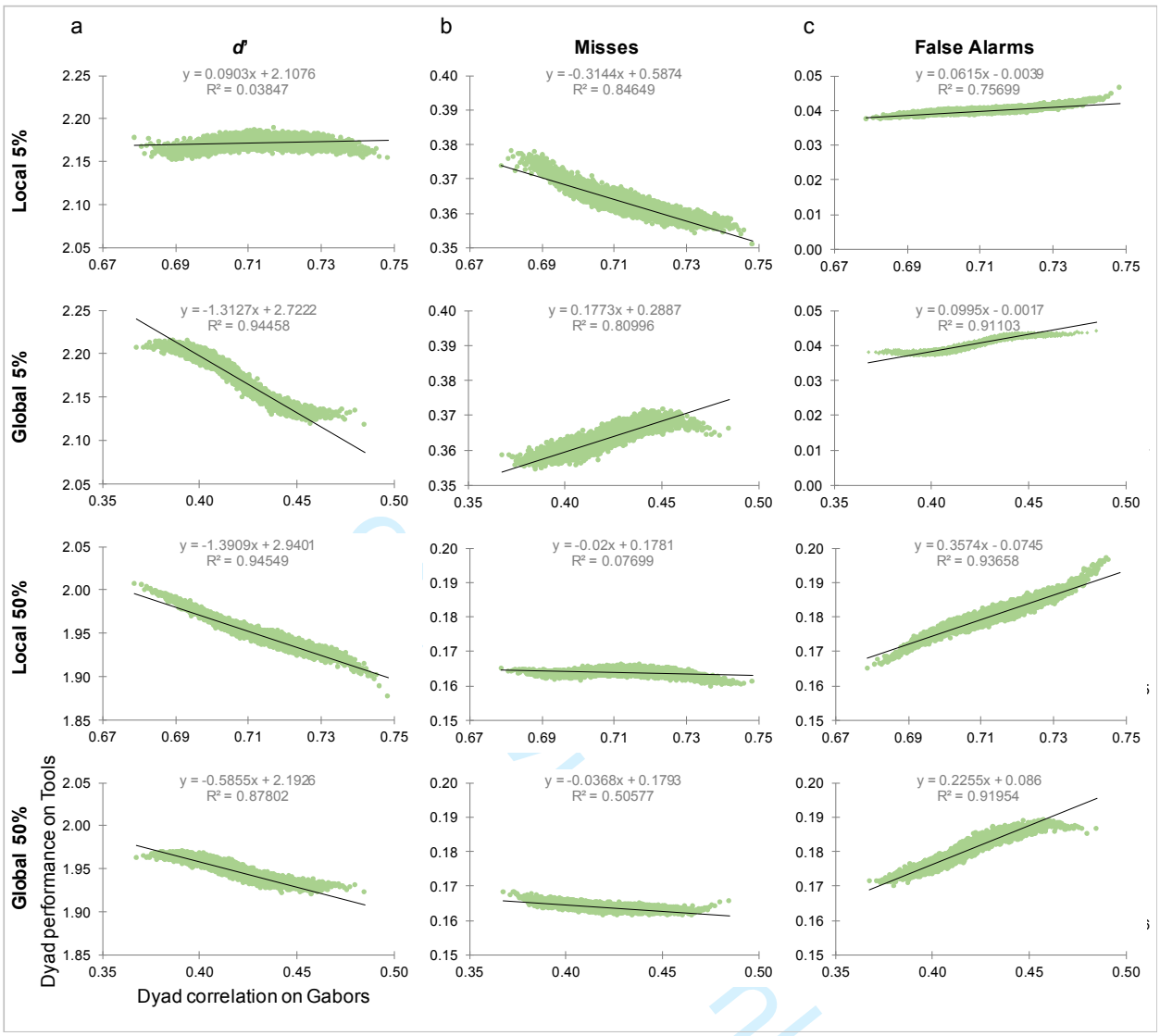

Figure 4. Decorrelated dyads results. a. *Detection performance (d'):* For all conditions except Local 5%, there was a strong negative correlation between dyads' correlated performance on the Gabors task and their corresponding *d'* values on the Tools task. **b.** *Misses:* Whereas dyads' miss rate in the 5% Tools task decreased with increasingly correlated performance in the Local Gabors task, it decreased with decreasingly correlated performance in the Global Gabors task. There was no strong relationship between dyads' correlated performance in either of the Gabors tasks and dyads' miss rates in the 50% Tools task. **c.** *False alarms*: In the 5% Tools condition, false alarms slightly increased as dyads' performance became increasingly correlated in both the Local and Global Gabors conditions. In the 50% Tools condition, there was a strong positive correlation between dyads' false alarm rates and correlated performance in both Local and Global Gabors tasks. All plots depict the results of simple regression analyses with all Fs(1,9008)>360 and all ps<0.001.

Discussion

The current framework represents a significant advancement towards improving human factors that mediate target detection. Building on parallels between the visual system's ability to

capitalize on the variance inherent in the surrounding environment (e.g., Barlow, 1961) and findings that the "wisdom of crowds" (e.g., Galton, 1907) can be approximated by aggregating the most conflicting estimates from the same individual (e.g., Corbett et al., 2011; Herzog & Hertwig, 2009; Vul & Pashler, 2008) the present study demonstrated that averaging estimates from individuals with the most decorrelated patterns of performance can improve detection across a range of visual tasks.

Several critical differences set the current framework apart from methods used in previous work (Van Wert et al., 2009; Wolfe et al., 2007, 2013, 2005). First, given evidence that observers can rapidly detect but not localize changes (Rensink, 2004), and that experts can detect but not localize the presence of abnormalities in mammography images within a fraction of a second (Drew et al., 2013; Evans et al., 2013, 2016), we used brief 500 ms stimulus presentations in attempt to engage the nonselective visual pathway thought to be responsible for the rapid extraction of global, statistical information (Drew et al., 2013; Evans et al., 2013, 2016; Rensink, 2004; Wolfe, Võ, Evans, & Greene, 2011). Second, instead of having observers provide a binary present/absent judgment or a binary judgement followed by a confidence rating, we required them to make a continuous response by adjusting a slider to indicate the extent to which they thought an image should be called back for further inspection because it contained a target. This allowed us to average individuals' estimates for each of the unique images, instead of limiting us to multiplying their average error rates to obtain a measure of paired performance (Wolfe et al., 2007) or calculating a separate measure of metacognitive confidence from binary responses and confidence ratings (e.g., Maniscalco & Lau, 2012). Third, the current framework does not require people to interact to reach a joint decision. Human interaction decreases the independence of individuals' judgements and results in joint performance that is sub-optimal

(e.g., Bang et al., 2017), as it can be no greater than improvements obtained from relying on the answers of the more confident (Bahrami et al., 2010) and/or accurate (e.g., Erev, Gopher, Itkin, & Greenshpan, 1995) member of each dyad. Instead, observers performed all tasks individually without interacting or any knowledge that their responses would be combined with those of other participants. Finally, we used performance on a basic visual task (Gabors) to pair individuals' estimates on a separate, more complex detection task (Tools). Averaging estimates from individuals with the most decorrelated patterns of performance in either the Local or Global Gabors task lead to the greatest improvements in detecting frequent Tools targets. In contrast, only averaging estimates from individuals with the most decorrelated performance on the Global Gabors task lead to improvements in detecting rare Tools targets. Whereas previous attempts to improve rare target detection have interspersed bursts of frequent targets into searches for infrequent targets (Wolfe et al., 2007, 2013), the present results suggest that the optimal solution is to pair estimates from observers using different cognitive strategies.

The most straightforward applications of our results are to begin employing continuous bounded response scales in safety-critical signal detection tasks and averaging individuals' estimates. For example, in-line with a recent report by Mayo clinic researchers underscoring the importance of second opinions in increasing the accuracy of medical diagnoses (Van Such, Lohr, Beckman, & Naessens, 2017), different radiologists can quickly scan through mammograms and continuous estimates from the individuals with the most decorrelated patterns can be combined offline. Importantly, our framework can be applied without requiring any communication between individuals and without requiring the delay between estimates that is necessary to obtain similar benefits from aggregating repeated estimates from the same individual (Corbett et al., 2011; Herzog & Hertwig, 2009; Vul & Pashler, 2008). Therefore, estimates made by an expert at

one location (e.g., a transportation security officer at an airport checkpoint) could be paired in real-time with estimates made by a second person viewing the same images at an offsite location. In the future, it may even be possible to simulate a decorrelated pattern of performance that can be averaged online with an existing individual's performance, which could be particularly cost-effective and useful when personnel or data is sparse. Although we note several possible applications in human and computer vision, the potential application domain extends broadly across a diverse range of fields from navigation to engineering to economic forecasting.

In conclusion, we have demonstrated a simple yet effective method for capitalizing on the diversity afforded by averaging continuous estimates of target presence from different individuals to boost target detection. Not only does this method yield benefits from combining independent samples in a range of different visual tasks, but our results strongly suggest that detection in complex tasks can be optimized by pairing estimates from individuals with the most diverse patterns of performance on a secondary basic detection task.

Acknowledgements: We thank Bahar Aykut, Miray Kapan, Elif Beyza Koş, Sümeyra Özçalık, and the members of the Visual Perception and Attention Lab (VPAL) for their help with data collection. We also thank Aaron Clarke for supplying the code for pre-processing the mammography images and for his help rating the initial set of mammogram stimuli. Many thanks to Chris Oriet for helpful comments on earlier versions of this work.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. https://doi.org/10.1037/h0054663

Baddeley, R. (1997). The Correlational Structure of Natural Images and the Calibration of Spatial Representations. *Cognitive Science*, *21*(3), 351–372. https://doi.org/10.1207/s15516709cog2103_4

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., … Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117. https://doi.org/10.1038/s41562-017-0117

Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). MIT Press.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLOS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Corbett, J. E., Fischer, J., & Whitney, D. (2011). Facilitating stable representations: Serial dependence in vision. *PloS One*, *6*(1), e16701.

Drew, T., Evans, K., Võ, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in Radiology: What Can You See in a Single Glance and How Might This Guide Visual

Search in Medical Images? *Radiographics*, *33*(1), 263–274. https://doi.org/10.1148/rg.331125023

Erev, I., Gopher, D., Itkin, R., & Greenshpan, Y. (1995). Toward a Generalization of Signal Detection Theory to N-Person Games: The Example of Two-Person Safety Problem. *Journal of Mathematical Psychology*, *39*(4), 360–375. https://doi.org/10.1006/jmps.1995.1034

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*(6), 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

Fleck, M. S., & Mitroff, S. R. (2007). Rare Targets Are Rarely Missed in Correctable Search. *Psychological Science*, *18*(11), 943–947. https://doi.org/10.1111/j.1467-9280.2007.02006.x

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature*, *75*(7), 450–451.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354. https://doi.org/10.1037/0278-7393.7.5.344

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., & Munishkumaran, S. (1998). Current Status of the Digital Database for Screening Mammography. In *Digital Mammography* (pp. 457–460). Springer, Dordrecht. https://doi.org/10.1007/978-94-011-5318-8_75

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, P. (2001). The Digital Database for Screening Mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography* (pp. 212–218). Medical Physics Publishing.

Herzog, S. M., & Hertwig, R. (2009). The Wisdom of Many in One Mind: Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science*, *20*(2), 231–237. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40–46. https://doi.org/10.1016/0030-5073(78)90037-5

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Kersten, D. (1987). Predictability and redundancy of natural images. *JOSA A*, *4*(12), 2395–2400. https://doi.org/10.1364/JOSAA.4.002395

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., … Stumpe, M. C. (2017). Detecting Cancer Metastases on Gigapixel Pathology Images. *ArXiv:1703.02442 [Cs]*. Retrieved from http://arxiv.org/abs/1703.02442

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. https://doi.org/10.1038/381607a0

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Rensink, R. A. (2004). Visual Sensing Without Seeing. *Psychological Science*, *15*(1), 27–32. https://doi.org/10.1111/j.0963-7214.2004.01501005.x

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429. https://doi.org/10.1073/pnas.0700622104

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Simons, D. J., & Chabris, C. F. (1999). Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception*, *28*(9), 1059–1074. https://doi.org/10.1068/p281059

Van Such, M., Lohr, R., Beckman, T., & Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. *Journal of Evaluation in Clinical Practice*.

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception, & Psychophysics*, *71*(3), 541–553. https://doi.org/10.3758/APP.71.3.541

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

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3), 33–33. https://doi.org/10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. https://doi.org/10.1038/435439a

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology. General*, *136*(4), 623–638. https://doi.org/10.1037/0096-3445.136.4.623

Wolfe, J. M., & Van Wert, M. J. (2010). Varying Target Prevalence Reveals Two Dissociable Decision Criteria in Visual Search. *Current Biology*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001