Rare targets are often missed in visual search

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*, 439-440.

Our society relies on accurate performance of visual screening tasks (e.g. for knives in luggage or tumors in mammograms). These are visual searches for rare targets. We report that target rarity leads to disturbingly inaccurate performance.

Visual search is the subject of a voluminous laboratory literature [1]. Typically, observers perform several hundred searches and targets are presented on 50% of trials. Target prevalence in baggage screening or cancer screening is much lower (~0.3% in routine mammography [2]). We compared performance on high and low prevalence versions of an artificial baggage-screening task. Observers looked for "tools" among objects drawn from other categories. Semi-transparent objects were presented on noisy backgrounds and could overlap (Fig 1). The number of objects in a display was 3, 6, 12, or 18. Target prevalence was 1%, 10% or 50%. At 1% prevalence, 12 paid volunteer observers had to be tested for 2000 trials each (broken into 250 trial blocks) to obtain a mere 20 target-present trials each, Each observer was tested for 200 trials in the 10% and 50% conditions. Observers were given feedback on their performance, including a point system designed to emphasize the importance of finding the target (see supplementary methods). Low prevalence search has some similarity to vigilance tasks where observers wait for fleeting signals [3,4]. However our search stimuli are continuously visible until observers choose to respond.

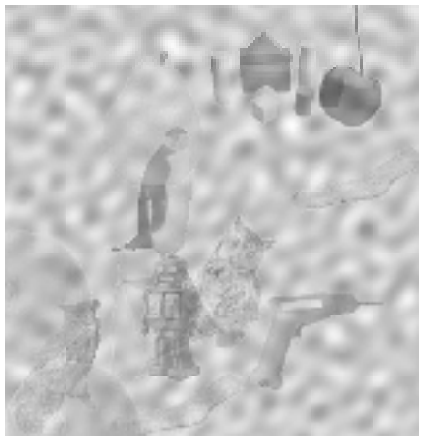

Figure 1: Stimuli: Observers searched for tools in displays with semi-transparent objects placed randomly on a noisy background.

Figure 2a shows error rates as a function of number of objects. 50% prevalence produced 7% miss errors, typical for laboratory search tasks of this sort. However, errors increased dramatically (and reliably) as prevalence decreased. 10% prevalence produced 16% errors, while at 1% prevalence errors soared to 30%. Errors were primarily "misses" (failing to notice a target). "False alarms" (saying "yes" when targets are absent) were vanishingly rare (0.03%), despite incentives to produce the opposite behavior (see supplementary methods). Simply changing prevalence produced a fourfold increase in error rate. If similar effects occur in socially important searches, this could have significant consequences.

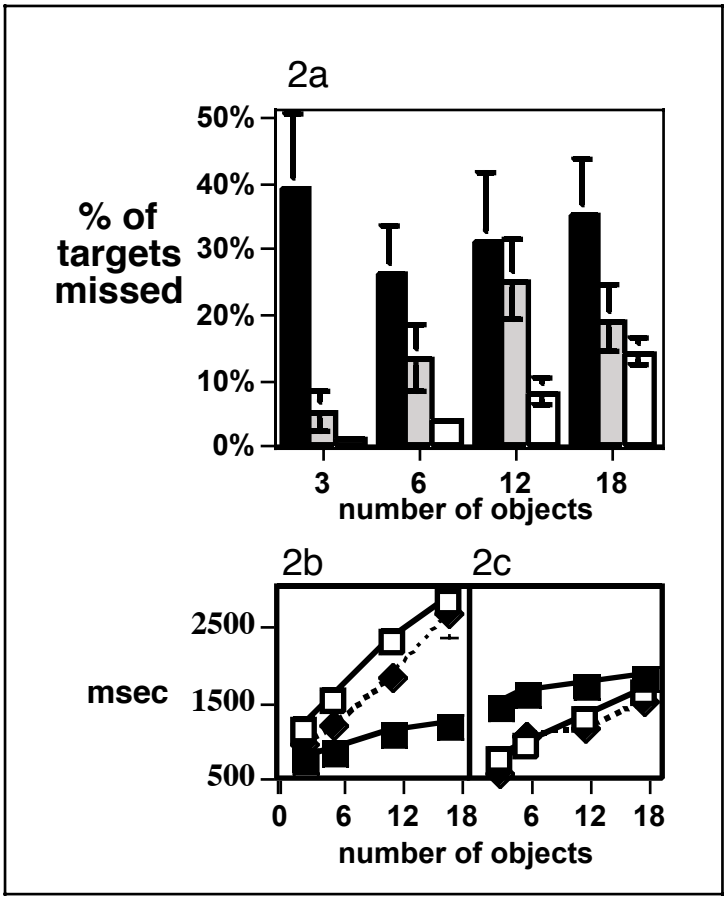

Figure Two: the effects of target prevalence on search performance.

2a: Error data: When targets were rare (1% prevalence - black bars) observers made more than 4X times the errors made when targets were common (50% prevalence - white bars). Data are averages of 12 observers. Error bars show +/- 1 s.e. for those 12 error rates. Gray bars show 10% prevalence results.

2b: Left: Reaction Times for 50% prevalence. Typical reaction times are longer when the target is absent (open symbols) than when targets a present (closed) Miss error RTs are shown by diamonds.

2c: Right: Reaction Times for 1% prevalence.However, when prevalence is low, observers make "absent" responses that are faster than the "present". This leads to elevated error rates. For 1c&d, error bars (s.e.) fall within data points.

Why does this happen? The reaction time (RT) data (Fig 2b&c) provide some clues. Observers require a threshold for quitting when no target has been found. This threshold is constantly adjusted; observers slow down after mistakes and speed up after successes[5]. When targets are frequent, fast "no" responses will often lead to mistakes. As a result, "no" RTs are slower than "yes" RTs in high prevalence search (2b). With infrequent targets, observers can successfully say "no" almost all of the time, driving down the quitting threshold. As seen in Fig 2c, the result is too rapid target-absent search. Observers abandon search in less than the average time required to find a target.

The problem cannot be solved simply by adding pseudotargets to increase prevalence (e.g. search baggage for iPods *and* weapons) In a second experiment, we mixed common (44% prevalence), rarer (10%), and very rare (1%) targets such that *some* target was present on 50% of trials. Here, observers missed just 11% of common targets but 25% of rarer targets and 52% of very rare targets (see supplementary methods & results). Is the prevalence effect just a by-product of naive Os unfamiliarity with the targets? In a separate investigation, we compared the miss error rate for 4000 trials at 1% prevalence (40 targets, 41% miss errors) to the miss error rate for the first 100 trials at 34% prevalence (34 targets, 11% miss errors). It appears to be prevalence, not just number of targets presented that is critical (See Supplementary Methods).

Visual search is a ubiquitous human signal detection task [6]. Heuristics that produce acceptable performance over a wide range of target prevalence may betray us at low prevalence. Because the experiments are burdensome, we do not have a clear idea whether these effects occur in the field [7,8]. A scoring system in the lab cannot duplicate

the motivation to find a gun or a tumor nor the motivation to move the check-in line along. The training of lab volunteers differs from that of professionals. Nevertheless, there are sufficient similarities between lab and field to strongly suggest that we should find out if the massive increases in error shown here occur in socially important search tasks.

Jeremy M Wolfe (1,2), Todd S Horowitz (1,2), & Naomi M Kenner (1)

1. Visual Attention Lab, Brigham and Women's Hospital, Boston, MA

2. Dept. of Ophthalmology, Harvard Medical School, Boston, MA

64 Sidney St.

Cambridge, MA 02139

wolfe@search.bwh.harvard.edu

References

1. Wolfe, J. M. in *Attention* (ed. Pashler, H.) 13-74 (Psychology Press Ltd., Hove, East Sussex, UK, 1998).

2. Gur, D. et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst* **96**, 185-90 (2004).

3. Warm, J. S. in *Workload transition: Implications for indvividual and team performance* (ed. Wicken, B. M. H. C. D.) 139-170 (National Academy Press., Washington, DC:, 1993).

4. Mackworth, J. *Vigilance and Attention* (Penguin Books, Harmondsworth, England, 1970).

5. Chun, M. M. & Wolfe, J. M. Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology* **30**, 39-78 (1996).

6. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res* **40**, 1227-1268 (2000).

7. Kundel, H. L. in *Medical Imaging 2000: Image Perception and Performance,* (ed. Krupinski, E. A.) 135-144 (2000).

8. Gur, D., Rockette, H. E., Warfel, T., Lacomis, J. M. & Fuhrman, C. R. From the laboratory to the clinic: the "prevalence effect". *Acad Radiol* **10**, 1324-6 (2003).

Supplementary information accompanies the communication on www.nature.com/nature.

SUPPLEMENTARY MATERIALS

Experiment One – Methods

Apparatus & stimulus details

The experiment was run on PowerMacIntosh G4 computers programmed in Matlab using the Psychophysics toolbox [1,2]. Stimuli were presented on 21" color CRT monitors. Targets and distractors were 60 photorealistic pictures from the Hemera Photo-Objects Collections. Items were presented on a noise background consisting of the sum of ten sinusoidal gratings of different orientations and spatial frequencies. Ten backgrounds were used at random from trial to trial. Each object was converted to grayscale and assigned an opacity of 40%. Since they were semi-transparent, entire objects could be seen, even when overlapping with another object. Figure 1a shows an example.

On each trial, 3, 6, 12, or 18 objects were presented. At a viewing distance of 57.4 cm, the background subtended 23.3º by 23.3º and each object subtended 4.5º by 4.5º of visual angle. There were five object categories with six instances of each category. Twelve observers searched for a *tool* among items drawn at random from the other categories (Table S1).

Toy	**Fruit**	**Clothing**	**Bird**	**Tool**
Puzzle	Grape	Shirt	Penguin	Hammer
Blocks	Peach	Dress	Duck	Saw
Kite	Apple	Shorts	Owl	Pliers
Robot	Pear	Socks	Eagle	Axe
Yo-yo	Cherry	Pants	Parrot	Drill
Ball	Orange	Vest	Chicken	Wrench

Table S1: Stimuli were drawn from five categories with six instances per category.

Procedural details

The critical variable in this experiment was target frequency. In different blocks, targets could appear on 1%, 10% or 50% of trials. In the 10% and 50% conditions, observers were tested for 50 practice trials followed by 200 experimental trials. Note that with 10% target-present trials, there are on average only 20 target-present trials per observer. In order to have even that meager number of target-present trials when target frequency was 1%, observers completed 100 practice trials and 8 blocks of 250 trials (2000 trials) in that condition. Observers could take breaks between blocks. Blocks lasted about 20 minutes (depending on the observer). Order of conditions was randomized for each observer. All observers were tested in all conditions.

Observers searched for tools that they were informed would appear "frequently", "somewhat rarely", or "very rarely", depending on the condition. They pressed one key if they saw a tool and another if they didn't. In an effort to generate some interest in an otherwise tedious experiment, the study was structured as a game with points earned or lost according to observers' performance. Scores for each block of trials were posted on screen, allowing observers to strive to improve their scores. This scoring system was designed to emphasize the importance of finding the target, as shown in the payoff matrix (Table S2). While we cannot simulate the pressures and payoffs of a real-world task like baggage screening, we designed the payoff with simulation of such tasks in mind. For example, a false alarm at the airport checkpoint is a nuisance while a miss could be a disaster. Accordingly, in the low prevalence condition, the penalties for incorrect "no" responses (misses) and the rewards for correct "yes" responses (hits) are much higher than the penalties for incorrect "yes" responses (false alarms) and the rewards for correct "no" responses (correct rejections).

Condition	Target Frequency	Total trials per observer	Hit points	Miss Points	Correct Rejection Points	False Alarm Points
1	1 %	2000	+2000	-2000	+2	-75
2	10 %	200	+200	-200	+2	-75
3	50 %	200	+40	-40	+2	-75

Table S2: Conditions and payoff matrices for Experiment One.

A tone preceded each trial by 500 ms and a fixation cross (0.5º by 0.5º) appeared in the center of the background image. Once presented, the search display remained on the screen until response. Screen feedback was given after every response. Additionally, a string of four fast beeps marked errors.

Experiment One – Further results & discussion

As shown in Fig 2a, the critical result is the very large increase in miss error rates. These misses seem to occur because the observers are abandoning search too quickly. The problem can be seen in the RT data (Fig 2c). At 1% prevalence, target-absent responses are faster on average than target-present responses. A clearer picture of the roots of the problem can be seen if we examine RTs for correct target-absent responses as a function of their position in the sequence of trials relative to a target-present trial. This is shown in Figure S1. RTs relative to a hit trial are shown in green. RTs relative to miss errors are shown in red.

When targets are present on 50% of trials, the RT after a successful trial is a little faster on average than the previous trial. RTs after errors are slower by a more significant amount [3]. Behavior at 1% prevalence has a number of unusual features (denoted with letters on the figure). RTs before a miss are about 500 msec slower than RTs before a

hit (Point A). Misses occur when the average time to say "no" has gotten dangerously low. This illustrates how "yes" responses come to be slower than "no" responses at low prevalence. After a low prevalence miss, observers are much slower on the subsequent trial (Point B). This jump of about 750 msec is far greater than what is seen in high prevalence conditions and reflects the high "value" of rare targets. As expected, no jump is seen after a hit. However, what would be an adaptive correction is counteracted by another force. RTs speed up markedly after a target-present trial (Point C). Even after a miss, RTs are faster within 4 trials. This appears to be a form of the "gambler's fallacy". The probability of a target is 1% on all trials, but observers behave as though a target on one trial means a reduced chance of a target on subsequent trials. Similar results have been seen in other vigilance tasks [4]. This deserves more study since it would be very unfortunate if the radiologist who found a tumor on one film behaved as though this reduced the probability of a tumor on the next. To summarize, there are rules that allow observers to end searches in a reasonable manner when targets are fairly frequent. These rules lead observers into maladaptive behavior when targets are rare.

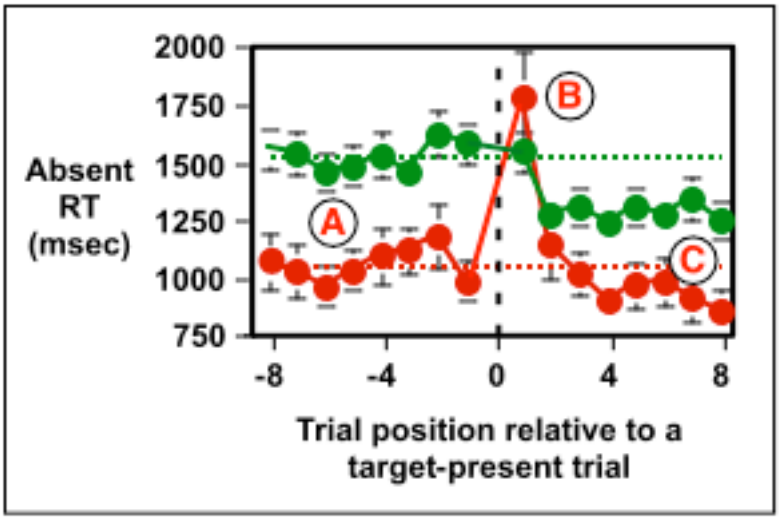

Figure S1: Target-absent RTs as a function of their occurrence relative to a target-present trial. RTs relative to "Hit" trials are shown in green. RTs relative to "Miss" trials are shown in red. Letters indicate interesting aspects of the data discussed in the text. Error bars are +/- 1 s.e.

Order effects: Because the order of conditions was counterbalanced, six observers performed the 50% condition before the 1% conditions (two of these had the 10%

condition intervening). These observers would have had extensive feedback after finding targets and more familiarity with target appearance. Four observers performed the 1% condition first and would have seen very few targets. Two observers had performed only the 10% condition before the 1% condition. Observers who performed the 50% condition before the 1% condition averaged 26% miss errors in the 1% condition, while observers who performed the 1% condition first averaged 46% errors. Observers who had performed in only the 10% condition before the 1% condition produced an intermediate 31% error rate in the 1% condition. There is a striking difference in error rates of observers who had experience with 50% prevalence before doing the 1% task and observers who did not have that experience. However, it is merely suggestive since variability between observers and the small numbers in each group make the difference statistically insignificant. The topic is worth further investigation because it points to the possibility of training regimens that might counteract the effects of low prevalence.

Experiment Two – Methods

Experiment Two was designed to determine if low prevalence error rates could be reduced if observers knew that *some target* was present half of the time. Eleven observers searched for three types of target: Guns, knives, and clocks (as a stand-in for bombs). One type was common (44% prevalence), one type was rare (10%) and the final type was very rare (1%). Mapping of target category to target prevalence was randomized for each observer and did not make a significant difference in this experiment. All target probabilities were independent, meaning that there could be 0, 1, 2, or 3 targets per display. Target prevalences were chosen to yield at least one target on 50% of trials, if targets are independent of each other. Observers pressed one key if they found a gun, another for knife, and a third for clock. They pressed a fourth key to

terminate the trial. Observers were tested for 2000 trials in blocks of 250 trials. Observers received feedback after each trial informing them of the identity of any missed targets ("You missed a gun").

Experiment Two – Further discussion of results

As noted in the main text, error rates were very high for rare targets. Figure S2 shows average error rates. There is a main effect of prevalence as assessed by the non-parametric Kruskal-Wallis test (KR statistic = 19.12, $p < 0.0001$).

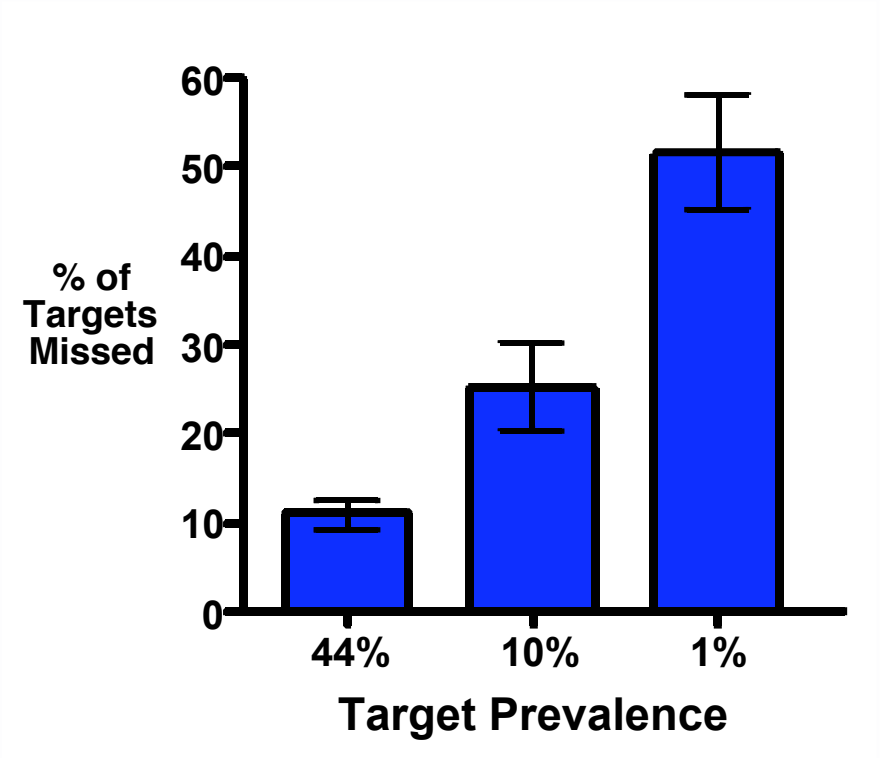

Figure S2: Miss errors as a function of target prevalence in Exp. 2 where any target could appear on any trial. Error bars represent standard errors of error rates of individual observers.

The importance of this result is that it indicates that the effects of prevalence are not just generalized effects of boredom or failures of vigilance. Here observers are finding something on almost half of the trials. Nevertheless, they are missing the rare targets at a very high rate. This suggests that observers are maintaining something akin to separate quitting thresholds for each target type. While they search successfully for the common target, their threshold for giving up on the search for the rare targets is set to a value that

causes them to quit too soon. Again, this is at best an approximation to the situation in socially important search tasks, but "satisfaction of search", the tendency to quit when one has found something rather than everything, is recognized as an issue in real-world search [5].

Experiment Three: Methods

The penultimate paragraph of the Brief Communication refers to a third experiment. As in Experiment One, observers searched for tools among other items. In this case, in the low prevalence condition, one specific tool (a wrench, for example) was present on 1% of trials. No other tools appeared. In the high prevalence condition, tools appeared on 50% of trials. Four different tools appeared with different probabilities (1%, 5%, 10% or 34%). For example, a saw might appear on 1% of trials, a hammer on 5%, a drill on 10%, and an axe on 34%. The actual mapping of tools to probabilities was randomized across observers. The remaining 50% of trials were target absent trials. Observers responded with one key to the presence of *any* tool and with another key to the absence of a tool. Nine observers were tested for 4000 trials in the 1%, low prevalence condition and 4000 trials in the 50%, high prevalence condition.

Experiment Three: Further discussion of results

In the low prevalence condition, observers missed 41% of the target items, replicating the basic prevalence effect. In the high prevalence condition, observers missed 23% of the 1% prevalence targets, 13% of the 5% prevalence targets, 9% of the 10% prevalence targets, and 8% of the 34% prevalence targets. For purposes of the discussion in the Brief Communication, we were interested to see if observers missed a large number of the common, 34% prevalence targets at the beginning of a session when they had only experienced a few searches for such targets. Looking at just the first 100 trials,

observers missed an average of 11% of those 34 targets. This compares to missing 41% of the first 40 targets in the low prevalence condition.

Jeremy M Wolfe (1,2), Todd S Horowitz (1,2), & Naomi M Kenner (1)

1. Visual Attention Lab, Brigham and Women's Hospital, Boston, MA

2. Dept. of Ophthalmology, Harvard Medical School, Boston, MA

64 Sidney St.

Cambridge, MA 02139

wolfe@search.bwh.harvard.edu

1. Pelli, D. G. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision 10, 437-442 (1997).
2. Brainard, D. H. The Psychophysics Toolbox. Spatial Vision 10, 443-446 (1997).
3. Chun, M. M. & Wolfe, J. M. Just say no: How are visual searches terminated when there is no target present? Cognitive Psychology 30, 39-78 (1996).
4. Holland, J. G. Technique for behavioral analysis of human observing. Science 125, 348-50 (1957).
5. Samuel, S., Kundel, H. L., Nodine, C. F. & Toto, L. C. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. Radiology 194, 895-902. (1995).

Acknowledgements: We thank the Transportation Security Administration for financial support.