

## Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks

Jeremy M. Wolfe and Todd S. Horowitz  
Brigham and Women's Hospital and Harvard Medical School

Michael J. Van Wert, Naomi M. Kenner,  
Skyler S. Place, and Nour Kibbi  
Brigham and Women's Hospital

In visual search tasks, observers look for targets in displays containing distractors. Likelihood that targets will be missed varies with *target prevalence*, the frequency with which targets are presented across trials. Miss error rates are much higher at low target prevalence (1%–2%) than at high prevalence (50%). Unfortunately, low prevalence is characteristic of important search tasks such as airport security and medical screening where miss errors are dangerous. A series of experiments show this *prevalence effect* is very robust. In signal detection terms, the prevalence effect can be explained as a criterion shift and not a change in sensitivity. Several efforts to induce observers to adopt a better criterion fail. However, a regime of brief retraining periods with high prevalence and full feedback allows observers to hold a good criterion during periods of low prevalence with no feedback.

*Keywords:* attention, visual search, airport security, low prevalence, signal detection

In a typical laboratory visual search task, observers look for a designated target in a field containing some distracting items. Although these experiments are intended to mimic the ubiquitous search tasks of everyday living, they lack a characteristic feature of an important class of tasks. In the laboratory, a target is usually present on 50% of trials—100% for identification tasks. In stark contrast, targets are very rare in a variety of socially important searches such as airport security (Rubenstein, 2001) or medical screening tasks (Gur et al., 2004; Pisano et al., 2005; P. A. Smith & Turnbull, 1997). We previously demonstrated that miss error rates were far higher at 1% target prevalence than at 50% prevalence (Wolfe, Horowitz, & Kenner, 2005). The average miss rate of about 0.30 in our low-prevalence task would clearly be a matter of concern at either the checkpoint or the clinic.

Low-prevalence visual search is characterized by a marked elevation of miss errors accompanied by a corresponding decrease in reaction time (RT) for target-absent trials (Wolfe et al., 2005). The natural explanation would be that observers simply became fast and careless at low prevalence. However, in this article, we demonstrate that the prevalence effect is not a simple speed–accuracy trade-off. In Experiment 1, two observers viewed the same sequence of stimuli. If prevalence errors were careless lapses, they should not be correlated between observers. However,

errors proved to be strongly correlated. Rather than being the product of carelessness, errors appear to have been caused by shifts of decision criteria that led both observers to miss most of the same targets. Observers faced with very infrequent targets said no quickly and repeatedly, shifting their criteria to a strongly conservative position. This pattern of results is well known in other settings. In decision theory, an optimal criterion is classically defined as the ratio of the probability of Event 1 (here, target absence) to the probability of Event 2 (target presence) multiplied by the payoff matrix associated with correct and incorrect responses to these events (Green & Swets, 1967; Maddox, 2002).<sup>1</sup> This ratio model is known to overpredict criterion shift and the resulting miss errors when prevalence is low; various efforts have been made to modify the model to predict human behavior (e.g., Treisman, 1984).

Perhaps the largest body of work related to the prevalence effect in visual search comes from the extensive work on vigilance in the 1960s. Jane Mackworth (1970) stated that “one of the most important findings in vigilance research has been the discovery that the probability that a signal will be detected is considerably reduced when the background event rate is increased” (p. 60; see also Broadbent & Gregory, 1965; Colquhoun, 1961). The background rate in this context is the number of nontarget events over a period of time. The more nontarget events there are, the lower the target prevalence. The introductory example in Green and Swets’s (1967) classic text on signal detection illustrates that the effect of reducing target prevalence is to make the decision criterion more

---

Jeremy M. Wolfe and Todd S. Horowitz, Visual Attention Lab, Brigham and Women's Hospital, Cambridge, Massachusetts, and Department of Ophthalmology, Harvard Medical School; Michael J. Van Wert, Naomi M. Kenner, Skyler S. Place, and Nour Kibbi, Visual Attention Lab, Brigham and Women's Hospital, Cambridge, Massachusetts.

This work was funded by a grant from the Department of Homeland Security's Transportation Security Laboratory Human Factors Program.

Correspondence concerning this article should be addressed to Jeremy M. Wolfe, Visual Attention Lab, Brigham and Women's Hospital, 64 Sidney Street, Suite 170, Cambridge, MA 02139-4170. E-mail: wolfe@search.bwh.harvard.edu

---

<sup>1</sup> Note that in this case, *criterion* refers to the signal detection bias parameter  $\beta$ . In our analyses later in this article, *criterion* refers to the signal detection theory criterion parameter  $c$ . An unbiased  $\beta$  is 1.00, whereas an unbiased  $c$  is 0.00. Briefly,  $c$  is the location of the criterion, and  $\beta$  is the signal/noise likelihood ratio at the criterion location. The two parameters are related to each other via  $d'$ :  $\ln(\beta) = cd'$ .

conservative. One is less likely to call something a target if the a priori probability of a target is lower.

Vigilance tasks differ in important ways from search tasks. Most vigilance tasks require the observer to detect “a faint and infrequent signal at an uncertain time” (Broadbent, 1964, p. 17). In contrast, search stimuli in the tasks described here remain present until the observer classifies the display as containing a target or not. This difference is critical because it means that a transient lapse in attention will not generate an error. If a baggage screener is distracted from the task, the line slows down until he or she returns to examine the bag. Turning to other differences, vigilance tasks typically present stimuli one at a time. Some tasks may require the observer to distinguish between signal and nonsignal stimuli (e.g., the continuous performance task; Beck, Bransome, Mirsky, Rosvold, & Sarason, 1956; Hsieh et al., 2005), but these targets and distractors occur successively. In other vigilance tasks, only signals are presented (e.g., the psychomotor vigilance task, or PVT; Kribbs & Dinges, 1994). Search tasks, by definition, involve search for a target among simultaneous distractors. Vigilance decrements appear as time-on-task increases. Importantly, many screening search tasks make efforts to minimize these fatigue effects. For example, airport screeners are rotated to different tasks every 20 min, precisely to avoid misses due to vigilance decrements. In our work, we also enforce breaks approximately every 20 min. Thus, although there are interesting theoretical similarities between vigilance decrements and prevalence effects in search, they are not the same phenomenon. The similarities may arise from fundamental aspects of human decision processes that play out in different ways in different settings.

In clinical radiology, there has been some work on prevalence effects in search. Here, the central question is whether prevalence affects the detectability of abnormalities. New display modalities are typically studied at relatively high prevalence. If prevalence changes sensitivity, then the results of high-prevalence testing of a new method in the lab might not be applicable to performance at low prevalence in the field. The widest range of prevalence is in a retrospective examination of lung cancer studies by Kundel (2000), with prevalence ranging from 0.1% to about 50.0%. He reported rather large effects on  $d'$  (1.4–3.9) with higher prevalence yielding lower  $d'$  values. That might seem to indicate a paradoxical increase in sensitivity as targets become rare, but Kundel offered a different account. Receiver operating characteristics (ROCs) form straight lines if plotted as  $z$  scores, and the standard symmetrical ROCs underlying the  $d'$  computation have slopes of 1.0 in  $z$  space. However, Kundel's data suggest a slope of 0.6. If low prevalence shifts criterion to a more conservative point on an asymmetrical ROC of this sort, false alarms will become rare at low prevalence, and estimates of  $d'$  (which assumes a unit-slope ROC) will rise. Kundel did not explain why the ROCs might have a nonunit slope, but we have found almost exactly the same slope of 0.6 in several different studies reported in this article. We return to this point later.

The Kundel (2000) study was retrospective. Ethell and Manning (2001) did a smaller prospective study in the lab, using wrist fracture prevalences of 83%, 50%, and 22%. They reported a modest decrease in  $A_z$  (area under the ROC, a measure of sensitivity that does not assume a unit slope) as prevalence decreased (0.80 at 83%, 0.71 at 50%, 0.68 at 22%). Eggin and Feinstein (1996) reported a similar effect with prevalences of 60% and 20%.

In a larger, well-designed study, Gur et al. (2003) found no significant effect on  $A_z$  and concluded that “with laboratory conditions, if a prevalence effect exists, it is quite small in magnitude; hence, it will not likely alter conclusions derived from such studies” (p. 10). From the point of view of the Gur et al. study, there was no prevalence effect because there was no effect of prevalence on sensitivity as measured by the area under the ROC. They did not report error rates or criterion measures.

The prevalence effect is not mere human frailty. Bond and Kamil (2002) trained blue jays to hunt for artificial moths on a computer screen and found that moths with a rare pattern survived better than those with more common markings. Hughes and her colleagues reported a similar finding in the survival rate of rare and common guppies in the field (Olendorf et al., 2006). These reports from the animal behavior literature may reflect the fact that the prevalence effect is a very sensible behavior unless the goal is to minimize misses, as it is in medical and airport screening tasks. If one's goal is to reduce miss errors and if one is willing to tolerate an increase in false alarms, shifting criterion to a more liberal position would be a solution to the prevalence problem.

There are two obvious solutions to the problem. One option is to change the payoff matrix, the pattern of rewards and punishments for correct and incorrect responses. In pilot experiments, changing payoff matrices did not have a noticeable effect on miss error rates. This finding was anticipated in the literature. Healy and Kubovy (1981), Maddox (2002), and others have noted that payoffs are less successful than target prevalence in moving criterion. Because we were operating at quite extreme prevalence (1%–2%, compared with 25% in Maddox, 2002, e.g.), we suspect that the payoffs required to overcome these levels of prevalence might exceed what is practical in the lab. However, we are currently testing monetary payoffs, though it is hard to imagine how one would simulate the real payoffs of an airport checkpoint or a radiology suite.

The second obvious remedy would be to increase prevalence to 50% by adding artificial target-present trials. Note that simply adding target-present trials would double the number of trials and therefore double the length of the task—hardly an ideal solution at a crowded airport or an understaffed radiology suite. We tested some less burdensome approaches in Experiments 2–7.

Experiments 2–6 continued to show robust prevalence effects. However, in Experiment 7, intermittent short periods of high prevalence with feedback were combined with long periods of low prevalence with no feedback. This appeared to cure the prevalence effect, allowing observers to maintain a criterion that reduces miss errors. In the discussion, we present a model to account for these results.

### Experiment 1a: Paired Observers

As noted, low-prevalence visual search tasks have some similarity to vigilance tasks (Parasuraman, 1986; Warm, 1993). If miss errors were generated by independent attentional lapses, then the joint errors of two observers would be the product of the individual error rates. Thus two observers, each missing 0.30 of rare targets, might be expected to jointly miss only 0.09 of them. Errors in difficult search tasks are not entirely independent of the stimulus. Still, if the additional errors due to low prevalence are random idiosyncratic lapses, then two observers should find a substantially greater proportion of targets than either observer alone.

## Method

### Making Luggage

We created a laboratory version of the task of screening carry-on luggage. The Department of Homeland Security's Transportation Security Laboratory supplied jpeg X-ray images of empty bags and a wide array of isolated objects such as sunglasses, toys, containers, keys, and clothing. Also included were 100 images of guns and 100 images of knives that were used as targets. Because detecting bombs (or IEDs: improvised explosive devices) requires substantially more training than detecting guns and knives, they were not included in these studies. Colors in these images code material type, with blue indicating metal, orange indicating organic material (plastics, clothing, food), and green indicating materials of intermediate density.

Using Matlab and the Psychophysics Toolbox (Brainard, 1997) running on Macintosh G4 computers, we generated artificial packed bags. For each trial, we took an image of an empty bag and added  $N$  objects, where  $N$  was the set size. Objects could overlap in a transparent manner, as in real X-ray images (though our algorithm for superimposing objects does not perfectly capture the effects of having X rays passing through multiple objects). No effort was made to model normal packing. Objects were assigned and placed at random.

Bags and individual objects varied in size in a realistic manner (e.g., a shaver would be bigger than a quarter). At the 57-cm viewing distance, bag images varied in size from  $9.5^\circ$  visual angle in height  $\times$   $16.0^\circ$  in width to  $20.0^\circ \times 21.5^\circ$ . Eight pieces of clothing were added to each bag but were not counted in the set size. Clothing in these images adds diffuse orange noise to the image. The resulting bags, although not perfectly realistic, are a good approximation of the stimuli presented to screeners at the airport. Figure 1 shows a representative example bag. Eye movements were unconstrained.

## Observers

In all experiments reported in this article, observers reported no history of eye or muscle disorders and were screened for color blindness and normal or corrected-to-normal visual acuity (no worse than 20/25). Informed consent was obtained from observers, and each was paid \$10/hr for his or her time. In Experiment 1a, 24 observers were tested on all conditions (ages 19–53 years,  $M$  age = 28.5 years,  $SD$  = 9.7 years; 14 women, 10 men). Observers were paired to form 12 pairs of observers. Each observer saw the same stimuli as his or her partner. Different sequences of trials were generated for each pair of observers.

## Procedure

To familiarize observers with the task, they were first briefly shown isolated examples of potential target weapons.

On each trial, a fixation cross and audible click were followed after 200 ms by the appearance of the stimulus. The stimulus remained visible until the observer responded. A 500-ms blank interval preceded the start of the next trial. Observers were instructed to indicate as quickly and accurately as possible whether a target was present or absent. If a target was present and the observer responded, "Yes," the target was outlined on the screen and the observer was presented with the message "Good for you. You found the target. Take a look and then press a key to continue." If the target was missed, it was then outlined on the screen, and the message read, "You missed the target. Take a look and then press a key to continue." Thus, observers were always given full feedback about target items and were required to make a second keypress to move to the next trial. Feedback was given for correct and incorrect responses to target-absent trials, but no second keypress was required. Feedback also included a point system designed to emphasize the importance of finding the target:

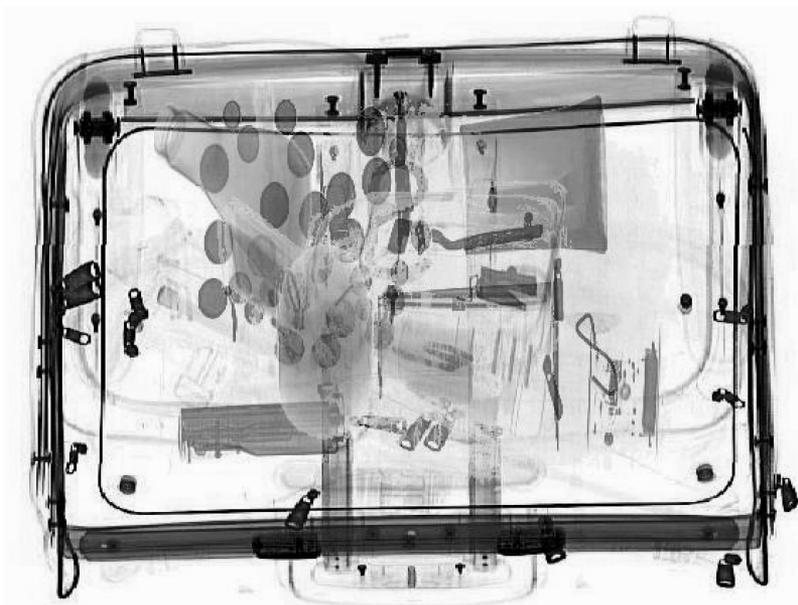


Figure 1. Sample of stimuli used in Experiment 1. This is a target-absent trial.

75 points subtracted for a false alarm and 150 points subtracted for a miss error, 25 points added for a hit and 5 points added for a correct rejection of a blank trial. In addition, observers lost 1 point per second to encourage speedy responses.

For each pair of observers, a single sequence of 1,299 stimuli was pregenerated. Observers were given 99 practice trials at 50% prevalence then tested for 200 experimental trials at 50% prevalence and 1,000 experimental trials at 2% prevalence. Experimental block order was counterbalanced across observers. The computer enforced breaks every 50 trials. These breaks had no fixed duration. Observers could get up and leave the testing room during these breaks, but this was not required. Every 200 trials, a minimum 2-min break was enforced. Six of the pairs of observers knew that another person would see exactly the same stimuli that they saw. The other six pairs did not know about the pairing.

### Data Analysis

In all experiments, except as noted, we removed as outliers all RTs > 10,000 ms or < 200 ms. Error rates were arcsine transformed before analysis; we report the back-transformed means. Analyses of variance (ANOVAs) were conducted in SPSS 11 for MacOS X. We report partial eta squared ( $\hat{\eta}^2$ ) as a measure of effect size.

### Results

Trimming outliers resulted in the removal of 1.4% of trials. Only 0.2% of trials were too long (43 of 24,000). Of the 369 RTs less than 200 ms, 269 were committed by two observers. The pattern of errors was the same with and without the data from those observers. Removing these outlier RTs decreased noise in the RT analysis.

Knowledge of the pairing manipulation made no difference to the results. At 50% prevalence, the mean error rate for observers who were aware that another observer would see the same images was 0.44, compared with 0.48 for observers who were not aware. The corresponding rates for 50% prevalence were 0.20 and 0.21. These rates did not significantly differ at either prevalence level, both  $t(11) < 1.00$ ,  $p > .50$ .

### Replication of the Prevalence Effect

Our previous results had shown that miss error rates were elevated at low target prevalence. This experiment replicated that effect with more realistic stimuli and clarified its underpinnings. The error data are shown in Figure 2 as a function of prevalence and set size. Note that, because of the vagaries of randomization, five observers had zero target-present trials at least one set size in the low-prevalence condition, so their miss rates could not be estimated; they were removed from the ANOVA. Performance for these observers did not differ systematically from the included observers. The miss errors clearly replicated the basic prevalence effect finding. Miss errors increased from an average of 0.20 ( $SD = 0.06$ ) at 50% prevalence to 0.46 (0.15) at 2% prevalence. The main effect of prevalence was statistically significant,  $F(1, 18) = 26.1$ ,  $p < .001$ ,  $\hat{\eta}^2 = .59$ , and this was true for each set size analyzed separately, all paired  $t$  tests,  $t(18) > 2.2$ ,  $p < .05$ .

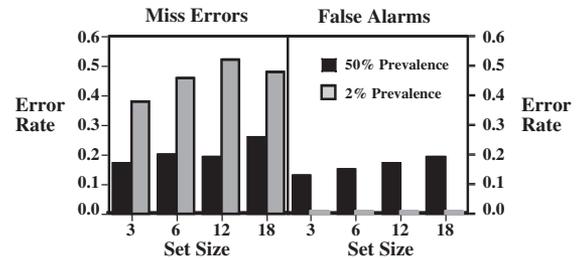


Figure 2. Miss and false-alarm errors averaged over the 24 observers in Experiment 1. Notice that miss errors are elevated at low prevalence, whereas false alarms are elevated at high prevalence. Note that within-subjects confidence intervals are not visible at this scale (about 1% for miss errors and less than that for false alarms).

In our prior work, false-alarm rates were very low. With this harder task, substantial false-alarm rates were generated but only at 50% prevalence ( $M = 0.19$ ,  $SD = 0.16$ ). False alarms were very rare at 2% prevalence ( $M = 0.01$ ,  $SD = 0.01$ ),  $F(1, 23) = 72.0$ ,  $p < .001$ ,  $\hat{\eta}^2 = .76$ . Similar results were obtained when we restricted analysis to the 18 observers included in the miss error analysis.

The presence of false alarms made it possible to compute  $d'$  for low and high prevalence. Pooling data across set size,  $d'$  was 1.97 (0.53) at 50% prevalence and 2.55 (0.48) at 2% prevalence. This difference was statistically significant,  $t(23) = 9.5$ ,  $p < .001$ , but may have been a side effect of the very low false-alarm rates at low prevalence. We discuss this more extensively later. The safe conclusion seems to be that elevated miss errors at low prevalence should not be attributed to a loss of sensitivity. Observers did not simply become sloppy under low-prevalence conditions.

If observers were not losing sensitivity at low prevalence, the alternative in a decision theory context is that they were shifting criterion (Green & Swets, 1967). Criteria ( $c$ ) were calculated as  $(\text{norminv}[\text{hit}] + \text{norminv}[\text{false alarm}]) / -2$  (Macmillan & Creelman, 2005). As can be inferred from the rough equality of miss and false-alarm errors,  $c$  was near zero in this task at 50% prevalence ( $M = 0.13$ ,  $SD = 0.32$ ). It was strongly skewed toward saying no at low prevalence ( $M = 1.14$ ,  $SD = 0.32$ ). The difference was significant,  $t(23) = 10.6$ ,  $p < .001$ .

### Other Effects

There were also main effects of set size for both types of errors—misses,  $F(3, 54) = 3.3$ ,  $p = .027$ ,  $\hat{\eta}^2 = .16$ ; false alarms,  $F(3, 69) = 5.8$ ,  $p = .001$ ,  $\hat{\eta}^2 = .20$ —and significant interactions between prevalence and set size for false alarms,  $F(3, 69) = 3.1$ ,  $p = .03$ ,  $\hat{\eta}^2 = .12$ , but not miss errors,  $F(3, 54) = 2.7$ ,  $p = .056$ ,  $\hat{\eta}^2 = .13$ . Again, the pattern of results did not change when the false-alarm analysis was restricted to the 18 observers analyzed in the miss error ANOVA, with the exception that the Prevalence  $\times$  Set Size interaction was no longer statistically significant ( $p = .089$ ).

Figure 3 shows RT as a function of set size. As reported in our earlier work, the main effect of target prevalence was on the RTs for target-absent responses, which were markedly faster at low prevalence; for correct true-negative trials,  $F(1, 24) = 35.6$ ,  $p < .001$ ,  $\hat{\eta}^2 = .60$ , and for misses,  $F(1, 13) = 40.1$ ,  $p < .001$ ,  $\hat{\eta}^2 =$

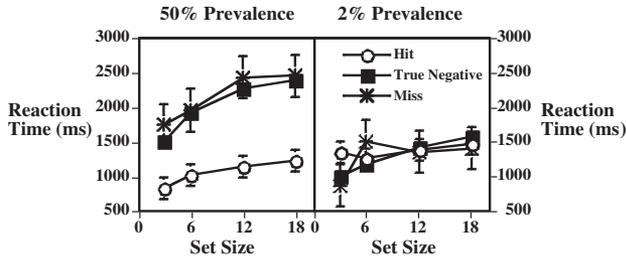


Figure 3. Reaction time as a function of set size for 50% and 2% prevalence. Error bars are within-subjects confidence intervals based on the comparison between high and low prevalence for each trial type.

.76. Hit RTs were reliably slower at low prevalence,  $F(1, 15) = 12.8, p < .001, \eta^2 = .46$  (note that  $N$  varied across these three analyses because of empty cells in the ANOVAs).

Paired Errors

Experiment 1a serves as a clear replication of the prevalence effect. The main purpose of the experiment, however, was to determine if two observers, viewing the same set of stimuli, made substantially fewer errors than either observer alone. At worst, the joint error rate for two observers can be no higher than the error rate for the better of the two. At best, errors might be completely independent in two observers. In that case, the error rate would be the product of the individual error rates. For example, the best that could be expected from a combination of observers with independent error rates of 0.40 and 0.30 would be a combined error rate of  $0.40 \times 0.30 = 0.12$ , whereas the worst would be 0.40 (paired performance could exceed this prediction if errors were somehow negatively correlated).

Figure 4 plots paired error (those trials missed by both observers) against the independence prediction (the product of Error 1 and Error 2). If the error rates of the two observers were independent, then the data in Figure 4 would lie on the line of unit slope. To the extent that the data lie above that line, errors were corre-

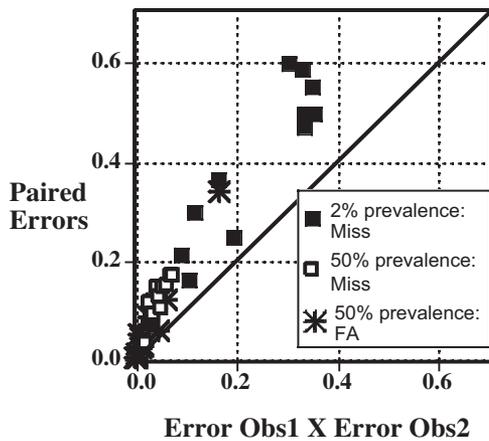


Figure 4. Paired errors—those targets missed by both observers—are greater than predicted by the product of the error rates of the individual observers. FA = false alarm; Obs = observer.

lated between observers. Data for false alarms at 2% prevalence are not shown because those errors were too rare. The data fall consistently above the independence line. In fact, the paired error is only slightly lower than the better of the two observers. At 50% prevalence, the advantage was 0.05 for miss errors and 0.05 for false alarms. Had the errors been independent, the advantage would have been 0.13 for miss errors and 0.09 for false alarms. At 2% prevalence, the case of greater interest, the improvement was only 0.01. Had the errors been independent, the improvement would have been 0.18, more than 10 times greater. If the two observers with high numbers of anticipation errors (RTs < 200 ms) are removed, the improvement for the remaining 10 pairs of observers is still only 0.02.

Errors in this task are strikingly correlated. At 2% prevalence, a total of 240 targets were presented. Each one was presented to a pair of observers. Summing the errors of the more accurate member of each pair, those observers missed 92 of the 240 targets. Of those 92 missed targets, 83 were missed by the other observer as well.

In practice, paired miss errors would be different than paired false alarms. If two screeners were examining baggage, the bag would be referred if either observer thought there was a target. Thus, paired misses require Observer 1 and Observer 2 to miss, whereas a paired false-alarm error occurs when either Observer 1 or Observer 2 commit a false alarm. With this definition of errors, we can compute  $d'$  and  $c$  for each pair of observers and compare that with the individual performance. Essentially,  $d'$  was unchanged relative to individual performance (50% prevalence: 1.97 individual vs. 1.94 paired; 2% prevalence: 2.55 individual vs. 2.57 paired; unpaired  $t$  tests,  $ns$ ). Criteria became slightly but reliably more liberal (50% prevalence: 0.13 individual vs.  $-0.28$  paired; 2% prevalence: 1.14 individual vs. 0.83 paired; both unpaired  $t$  tests,  $p < .025$ ). The shift in criterion is not surprising because the mechanics of pairing make it easier to make a false alarm than a miss. Nevertheless, the effect on miss rate is very modest.

Time Course

Returning to the individual data, plotting miss rates as a function of epoch illustrates the development of the prevalence effect (Figure 5, left panel). Error rates were computed for each 250 low-prevalence trials. For comparison, the error rate for the 50%-prevalence condition is shown as a dashed line.

Miss errors at 2% prevalence are comparable to the 50%-prevalence level over the first 250 trials, then the miss rate rises to

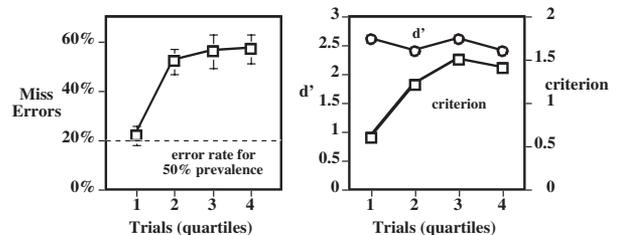


Figure 5. Left panel shows miss errors at 2% prevalence as a function of time (binned into 250 trial quartiles). Dashed line represents miss rate at 50% prevalence. Right panel shows the corresponding  $d'$  and criterion values. Error bars are within-subjects confidence intervals.

over 0.50 for the remainder of the experiment. The right panel of Figure 5 shows that the increase in miss errors over quartiles was due entirely to changes in criterion, rather than reduction in sensitivity; observers became less willing to say yes. This slow onset of the prevalence effect was not replicated in our later experiments, in which the criterion shift occurred more rapidly.

### Discussion

Three aspects of the present results discount the theory that observers simply became careless at low prevalence. First, if observers were becoming careless,  $d'$  should have been reduced at low prevalence. In fact, it increased. That increase should not be overinterpreted. The  $d'$  estimates at low prevalence were derived from a condition yielding very few false alarms, and as discussed in the introduction, the rise in  $d'$  may derive from an asymmetrical ROC curve (see below). Nevertheless, we can reject the hypothesis that  $d'$  declines at low prevalence in this experiment.

Second, simple lapses of attention or motor errors would predict a reasonable degree of independence in the errors, but that is not what happened. Errors were highly correlated between paired observers. It strains credulity to imagine that a pair of observers, viewing the same stimuli independently, would be independently careless on the same trials.

Finally, the signal detection measures shown in the right panel of Figure 5 indicate that the increase in miss errors was caused by a criterion shift and not a loss of sensitivity. Presumably, the correlation within pairs occurred because an image that is hard to interpret by one observer will be hard to interpret by another. Indeed, many errors involved guns and knives presented in non-canonical orientations (e.g., edge on) in addition to canonical ones (e.g., a gun laying flat on one side), though the current experiment lacked the power to determine precisely how these stimulus features affect performance. Note that, if observers were simply missing hard targets that they might have caught at high prevalence when those targets were presented at low prevalence, then  $d'$  should have declined. Instead, it appears that both members of a pair shifted criterion and that the shift affected each observer's assessments of difficult stimuli in the same way. At low prevalence, they missed more ambiguous targets. At high prevalence, both members of a pair false-alarmed for more ambiguous non-targets.

What drives observers to shift their criterion? In the General Discussion, we present a model proposing that observers try to equalize the numbers of misses and false alarms, a method that will produce an optimal criterion at 50% prevalence. Evidence for this hypothesis can be seen in the raw counts of errors. At 50% prevalence, observers made an aggregate 388 false alarms and 483 miss errors. At 2% prevalence, observers made an aggregate 245 false alarms and 214 miss errors. Of course, the false-alarm rate was much lower than the miss rate at 2% prevalence because of the vastly greater number of target-absent trials. Paired  $t$  tests show that the absolute number of miss errors was not different from the absolute number of false alarms at low prevalence ( $p = .56$ ) or high prevalence ( $p = .22$ ).

#### Experiment 1b: Two Observers Working Together

Suppose that two observers worked together rather than viewing stimuli separately. Are two heads better than one?

### Method

#### Observers

Twenty observers between the ages of 18 and 43 years participated ( $M$  age = 24.7 years,  $SD = 7.4$  years; 13 women, 7 men). Four observers who participated in Experiment 1a also participated in this experiment. For these observers, at least one month elapsed between experiments.

#### Procedure

Observers were divided into 10 pairs. Six pairs were composed of observers paired randomly with whomever was able to participate during the next available time slot. Four pairs were acquaintances who arranged to arrive together. The task was similar to Experiment 1a except that each pair of observers sat together, viewed the stimuli simultaneously, and needed to concur on the response before it was finalized. As before, they were asked to detect guns and knives. Images were projected on the wall using an LCD projector. At an approximate viewing distance of 2.9 m, stimulus (i.e., bag) size varied from  $80.0^\circ \times 47.5^\circ$  to  $107.5^\circ \times 100.0^\circ$ . Observers were instructed to discuss stimuli together and to each make an independent response on each trial (one pressed computer keys, the other clicked a mouse button). If responses were different after initial discussion, the computer asked observers to reconcile their responses through further discussion before proceeding. Once a consensus was reached, one of the observers (designated at the beginning of the experiment) made the final response for the pair, and the next trial was presented. RTs were collected for each observer's response and for the concurring response, if any. Observers were tested for 1,000 trials at 2% prevalence and 200 trials at 50% prevalence. Methods were otherwise similar to Experiment 1a.

### Results and Discussion

The average error results are shown in Figure 6. The solid bars show miss and false-alarm errors for Experiment 1b. The patterned bars in the background show the equivalent data for the 24 individual observers of Experiment 1a (there is no equivalent single-

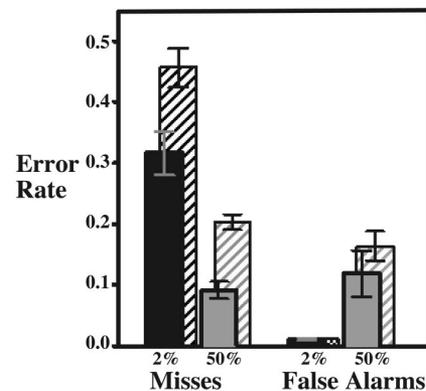


Figure 6. Error rates for 2% and 50% prevalence. Solid bars are for Experiment 1b. Patterned bars are equivalent data from individual observers in Experiment 1a. Error bars are within-subjects confidence intervals.

observer data for this experiment because observers were always sitting together). There is a clear prevalence effect. Observers made 0.09 (0.02) miss errors at 50% prevalence. This rose to 0.32 (0.04) miss errors at 2% prevalence. False alarms fell from 0.12 (0.04) to 0.01 (0.00). Both of these changes are statistically significant: miss,  $t(9) = 5.7, p = .0003$ ; false alarms,  $t(9) = 3.6, p = .006$ . As in Experiment 1a, this represents a change in criterion (0.11–0.93) and not  $d'$  (2.98–2.85). The change in criterion is significant,  $t(9) = 4.8, p = .001$ . Clearly, having two observers looking at the same display at the same time does not eliminate the prevalence effect.

As shown in Figure 6, performance was somewhat better with two observers than with one. The improvement from 0.45 miss errors to 0.31 at 2% prevalence is respectable. However, there is less here than meets the eye. First, if one analyzes the data from Experiment 1a, one sees a similar improvement. The average error for individual observers in that experiment was 0.45. The average paired error was 0.35. Recall, however, that the discouraging aspect of this result was that the paired result was only 0.01 better than the better of the two observers in a pair. In Experiment 1b, we cannot know who would have been the better observer. Moreover, the improvement in miss errors comes at a substantial cost in speed. At 2% prevalence, average correct target-absent responses (the bulk of the responses) rise from about 1,300 ms in Experiment 1a to about 2,800 ms in Experiment 1b. At 50% prevalence, the rise is from about 2,000 ms to over 7,000 ms.

How one regards these results depends on how much one is willing to pay for the elimination of about a third of the miss errors. Experiment 1a showed that most of the improvement could be achieved by using the better screeners. This assumes that the observer who is better today would also be better tomorrow—something one does not know. If one cannot identify the better observers, then pairing them, either independently or working together, will produce a reasonable reduction in miss errors. In principle, however, it should be possible to do much better. The improvement from 0.45 miss errors to 0.31–0.35 miss errors does not get those errors down to the 0.20 rate of miss errors seen at 50% prevalence in this task. In the remaining experiments in this article, we focus on changing the performance of individual screeners. We seek to reduce miss errors by shifting the low-prevalence criterion to a value closer to that seen for high prevalence.

### Experiment 2: Slowing the Response

In both versions of Experiment 1 and in our previous work, observers responded, “No,” much more quickly at 2% prevalence than at 50%. If observers were missing targets because they are saying no too quickly, then perhaps we could reduce the miss errors by persuading observers to respond more slowly. Fleck and Mitroff (in press) reported that the prevalence effect was eliminated in one experiment when they gave observers the option to correct responses after the fact. They argued that the rise in miss errors at low prevalence is primarily a rise in motor errors. Li, Li, Gao, Chen, and Lin (2006) also reported that they could reduce the prevalence effect if they changed the response in a manner designed to interfere with the motor habit of making quick no responses. In our Experiment 2, we slowed observers by warning them when they were going too quickly.

### Method

#### Observers

Twelve observers between the ages of 22 and 53 years participated ( $M$  age = 30.1 years,  $SD = 10.2$  years; 5 women, 7 men).

#### Procedure

The apparatus and stimuli were the same as in Experiment 1. Observers were first tested for 99 practice trials at 50% prevalence with set sizes of 3, 6, 12, and 18 items. This was followed by 200 experimental trials at 50% prevalence. From those 200 trials,  $RT \times$  Set Size functions were computed. The speed limit was defined as 1.3 times the target-present RT at each set size. Observers were then run for 1,000 experimental trials at 2% prevalence. If an observer made a target-absent response with an RT faster than the speed limit for that set size, the observer was asked to slow down by an annoying series of beeps and by text on the screen. We were not concerned with the speed of target-present trials. Observers were required to make a second response to any trial that was too fast. Note that these speeding tickets were handed out without regard to the accuracy of the response. In all other respects, this experiment was similar to Experiment 1.

### Results and Discussion

Speeding tickets were handed out on 9.3% of target-absent trials. As Figure 7 shows, this had the desired effect of slowing responses.

In Figure 7, thick lines are Experiment 2 data for those trials on which observers were not issued speeding tickets. Thin lines show Experiment 1a data, replotted from Figure 3. The left panel shows that the performance at 50% prevalence was identical in Experiments 1a and 2. On the right, the slowing effect of the speeding tickets is easily seen. At low prevalence, the correct target-absent responses were about 1 s slower in Experiment 2 than in the comparable condition of Experiment 1a. This made low-prevalence RTs comparable to high-prevalence RTs. However, Figure 8 shows that this massive slowing had no appreciable effect on errors.

The error rates in Experiment 2 were very similar to those for Experiment 1a (compare Figure 2). Critically, there was no evidence for a reduction in the overall miss error rates. The main

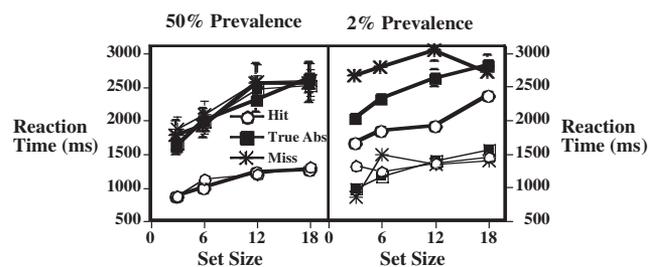


Figure 7. Reaction time as a function of set size. Thick lines show data from Experiment 2. Thin lines reproduce data from Experiment 1a. Note that low-prevalence reaction times were markedly slowed by the speeding tickets handed out at low prevalence. Error bars are  $\pm 1 SE$ . True Abs = true target-absent responses.

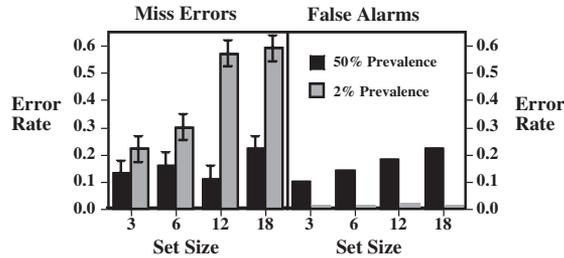


Figure 8. Error rates for trials without speeding tickets in Experiment 2. Note the similarity to Experiment 1a (Figure 2). Error bars are within-subject confidence intervals (not visible for false alarms).

effect of prevalence on miss errors was statistically significant,  $F(1, 10) = 23.6, p = .001, \hat{\eta}^2 = .70$ . The main effect of set size on miss errors was not statistically significant,  $F(3, 30) = 1.7, p = .18, \hat{\eta}^2 = .15$ , nor was the Set Size  $\times$  Prevalence interaction,  $F(3, 30) = 1.3, p = .28, \hat{\eta}^2 = .12$ . Turning to the false alarms, the main effect of prevalence on false alarms was statistically significant,  $F(1, 10) = 134.3, p < .001, \hat{\eta}^2 = .92$ . The main effect of set size on false alarms was statistically significant,  $F(3, 30) = 5.4, p = .004, \hat{\eta}^2 = .33$ , as was the Set Size  $\times$  Prevalence interaction,  $F(3, 30) = 4.2, p = .012, \hat{\eta}^2 = .28$ .

The speeding tickets did have an effect on the trials in which they occurred. The miss error rate for ticketed trials was 0.12, compared with 0.44 for all other trials. This agrees with Fleck and Mitroff's (in press) finding that allowing a second response improves performance. However, a speeding ticket on one trial did not convince the observers to take that second look on subsequent trials.

$d'$  increased from 2.01 at 50% prevalence to 2.43 at 2% prevalence. This was statistically significant,  $t(11) = 3.5, p = .01$ , but may be an artifact of a non-unit-slope ROC (see below). Criterion ( $c$ ) shifted from neutral ( $-0.02$ ) at 50% prevalence to strikingly conservative (1.07) at 2% prevalence,  $t(11) = 10.9, p < .0001$ . The effects of epoch were less dramatic than in Experiment 1a. Still, there was evidence for an increase in criterion after the first quartile of trials and no evidence for a decrease in  $d'$ .

This result is rather mysterious. Observers were taking more time with the stimulus but accrued no accuracy benefit from that additional time. It cannot be that the targets were intrinsically difficult to detect because the same targets were more frequently detected when prevalence was 50%. Apparently, the additional time improved neither sensitivity nor criterion. The results of Experiments 1 and 2 indicated that we should be trying to shift observers' criteria back toward zero in low-prevalence search. In Experiments 3–6, we tried to accomplish this by having observers search for more than one target.

### Experiment 3: Saying Yes Frequently Might Help

Low-prevalence search produces high miss error rates. High-prevalence search produces lower rates. This suggests that miss errors might be reduced by artificially increasing the prevalence rate. There are a number of ways to do this. If one imagines a set of luggage images or mammograms, one could double the size of the set by adding 50% known target-present images from a library of such images, though this would double the workload of the

screener. Experiment 3 took a different approach to increasing effective target prevalence. In this experiment, observers searched for four targets of differing prevalence. In screening at an airport checkpoint, this might be akin to searching for threats and hair dryers. Does the high prevalence of hair dryers counteract the low-prevalence effect and permit observers to find more threats?

## Method

### Observers

Ten observers between the ages of 18 and 55 years participated ( $M$  age = 35.4 years,  $SD = 12.8$  years; 8 women, 2 men).

### Procedure

Stimuli were 60 photorealistic black-and-white pictures of objects taken from the Hemera Photo-Objects Collections, previously used in Wolfe et al. (2005). The objects were shown 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 semitransparent (opacity = 40%) to simulate airport X-ray images. This allowed entire objects to be seen even when objects overlapped. Objects were drawn from five categories: toys, birds, fruit, clothing, and tools. Six examples of each category were used. In this experiment, observers searched for tools. Figure 9 shows an example of a search stimulus from this experiment.

On each trial, 3, 6, 12, or 18 objects were presented at a viewing distance of 57.4 cm, the background subtended  $23.3^\circ \times 23.3^\circ$ . Each object was constrained to fit in a virtual  $4.5^\circ \times 4.5^\circ$  box.

In this study, there were three conditions: blocked 50%, blocked 1%, and mixed.

**Blocked 50%.** The 10 observers searched for a single tool (e.g., a hammer) that was present on 50% of trials. Observers were tested for 50 trials on this task as practice and then for 100 trials. If present, the target tool was the only tool.

**Blocked 1%.** The 10 observers searched for a single tool that was present on 1% of trials. Seven observers were tested for 4,000 trials on this task. Three observers were tested for just 2,000 trials. There is no evidence that the number of trials affected the pattern

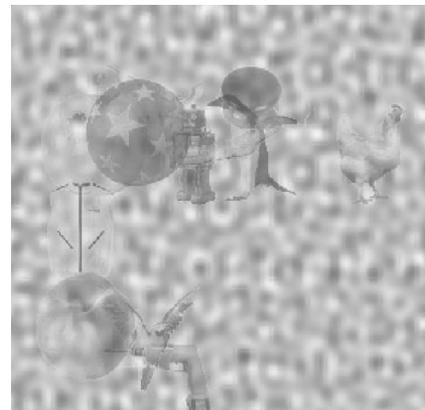


Figure 9. Sample stimulus from Experiment 3. The target is the drill at the bottom of the figure.

of results. Average results represent the means of mean results for each observer.

*Mixed.* In the critical mixed condition, the 10 observers searched for four tools, each with a different prevalence. Observers were told to search for tools and were not given information about the specific frequencies of different tools. Prevalence levels were assigned randomly to tools for each observer. For a given observer, the prevalence structure might be pliers on 34% of trials, axe on 10%, drill on 5%, and hammer on 1%. No tool was presented on the remaining 50% of trials. Thus, the overall probability of a tool was 50%, but hammers might appear only 1% of the time in this condition, as in the blocked-1% condition. Seven observers who were tested for 4,000 blocked trials were tested for 4,000 mixed trials, and the other three observers were tested for 2,000 trials. Again, there is no evidence for a difference across these groups.

Two set sizes were tested, 3 and 18 items. Observers were asked to take breaks after every 50 trials.

### Results and Discussion

Trimming RT outliers removed 1% of the data. Figure 10 shows the critical miss error data.

The manipulation of prevalence in Experiment 3 was a partial success. The blocked-1% condition replicated the now familiar low-prevalence effect, producing a 0.39 rate of miss errors against only 0.06 miss errors at high prevalence. In the mixed condition, the 1%-prevalence targets were missed at a rate of 0.21. This is significantly better than the low-prevalence rate,  $t(9) = 4.0$ ,  $p = .003$ . The success of the manipulation was only partial because the miss error rate remained significantly elevated when compared with the 50%-blocked case,  $t(9) = 6.9$ ,  $p < .001$ . However, a nearly 50% reduction in the miss errors is quite encouraging. The improvement was more marked for the smaller set size (0.36–0.13) than at the larger set size (0.43–0.30). Unsurprisingly, in the mixed condition, it can be seen that performance improved steadily as target prevalence increased. Observers were more successful in finding more common tools.

With these stimuli, false-alarm errors were very rare (1% overall, with many 0% cells). This made it unwise to use signal detection methods to analyze the results.

This is an encouraging result, suggesting that if screeners were asked to look for threats and hair dryers, they might find more threats than if they search for threats alone. However, we must

sound at least three cautionary notes. First, this experiment used single items as targets. That is, all trials with a hammer target used the same hammer. There is evidence that observers learn specific targets and that this learning does not generalize well to other members of the same category (J. D. Smith, Redford, Washburn, & Tagliatalata, 2005). Second, more critically, the three different targets were all from the tools category. It might be that observers were able to search for the category, making this effectively a search for high-prevalence targets. Recent evidence shows that observers pay a cost when searching for items from several categories (J. D. Smith, Redford, Gent, & Washburn, 2005). On the other hand, if searching within a category were better than searching across categories, then one could enrich the search stimuli with added examples from the target category. In both airport security and radiology, it is possible to introduce artificial targets into the real image being screened. This, however, introduces the third issue. In Experiment 3, a trial could have either one target or no targets. If artificial targets are added to displays that might also contain real targets, then it becomes possible to have more than one target per display. This raises the possibility that the real target would be found less frequently in the presence of the added target because observers would find the artificial target and be satisfied enough to end the search. In the radiology literature, this is known as the “satisfaction of search” problem (Berbaum, Franken, Dorfman, Caldwell, & Krupinski, 2000; Berbaum et al., 1990; Nodine, Krupinski, Kundel, Toto, & Herman, 1992; Samuel, Kundel, Nodine, & Toto, 1995). The next two experiments explored these issues.

### Experiment 4: Saying Yes Frequently Might Not Help

In Experiment 4, we addressed the limitations of Experiment 3 by having observers search for three different categories of target rather than four instances of the same category. Each category had a different prevalence rate, and the probability of each category appearing on a given trial was independent of all the others.

### Method

#### Observers

Eleven observers between the ages of 19 and 49 years participated ( $M$  age = 33.1 years,  $SD = 11.3$  years; 4 women, 7 men).

#### Procedure

Using stimuli like those in Experiment 3 (see Figure 9), observers searched for members of three new target categories: guns, knives, and clocks. Each category was assigned a prevalence rate. Prevalence rates were 1%, 10%, and 44%. With the appearance of each category independent of the others, this meant that there was a 50.1% chance of at least one target being present on each trial: The probability of just one target was 45.2%. Two targets were present on 4.8% of trials and three targets on a very rare 0.04%. The mapping of prevalence to specific target types was random across observer. Observers were told that the prevalence of different categories would be very different. Guns and knives were chosen as categories because of the obvious relevance to airport security. Clocks were used as a stand-in for IEDs.

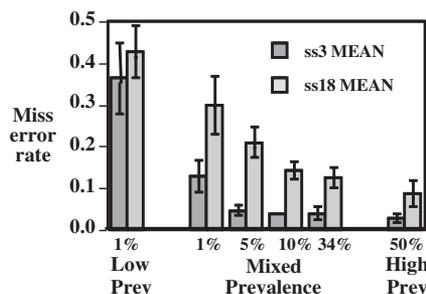


Figure 10. Miss error rates as a function of condition and target prevalence in Experiment 4. Dark bars show set size (ss) 3 data. Light bars show ss 18. Error bars are standard errors of the mean.

Observers were tested at a single set size of 18 items. Testing was divided into two visits on different days. Observers performed 100 practice trials and 1,250 experimental trials on the first visit, followed by 50 practice and another 1,250 experimental trials on a second visit. Six observers performed both sessions one day apart, and an average of 5.2 days elapsed between the two visits for the remaining five observers.

The response method was changed for this experiment. Observers were given separate response keys for guns, knives, and clocks. If a target was detected, observers responded with the appropriate key. A fourth key was used to terminate a trial. That is, when an observer decided that he or she had found all the targets that he or she was going to find, the observer pressed the fourth key to continue. Different types of trials, therefore, had different numbers of responses. For example, on a successful two-target trial, the observer would press one key for each target as it was discovered and make a third, final keypress to indicate that search was complete for that trial. RTs were recorded but are not particularly informative in this experimental design. The primary dependent variable is the miss error rate. Methods were otherwise similar to previous experiments.

### Results and Discussion

Figure 11 shows the miss error rates for each target prevalence for trials with one or two targets present (independent probabilities for three types of targets created three-target trials, but these were too rare to be of interest). False alarms were rare (0.5% of target-absent trials) and may represent responses to one target with the key assigned to another.

The most obvious finding is that the prevalence effect reappeared here in strength. Observers missed an average of 0.52 (0.07) of the rare, 1%-prevalence targets even though they were responding yes on 50% of trials. They missed 0.25 (0.05) of the 10%-prevalence targets and 0.11 (0.02) of the common, 44%-prevalence targets. The main effect of prevalence is significant,  $F(2, 22) = 10.2$ ,  $p = .001$ ,  $\hat{\eta}^2 = .48$ . Post hoc tests show that errors at 1% prevalence are reliably greater than those at 10% and 44% prevalence (Tukey honestly significant difference [HSD],  $p < .001$ ). The difference between 10% and 44% prevalence is marginal (Tukey HSD,  $p = .067$ ).

Because the assignment of category (guns, knives, clocks) to prevalence (low, medium, or high) varied across observers, we

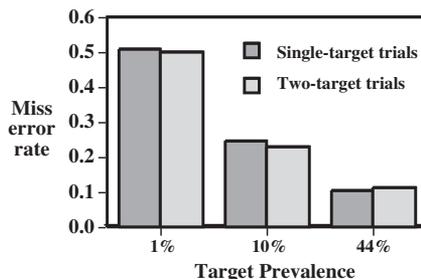


Figure 11. Miss error rates as a function of target prevalence and of the number of targets in a display. Error rates are back-transformed from arcsine-transformed errors. Within-subject confidence interval error bars are not visible at this scale.

know that the low-prevalence items were not missed simply because they were hard to recognize. The same items were found on about 0.90 of the trials when they were high-prevalence items.

Though the satisfaction of search problem is a factor in some tasks in radiology, Figure 11 suggests that it is not the issue here. The error rates were similar on one- and two-target trials. An ANOVA with target prevalence and number of targets as factors showed no effect of number of targets,  $F(1, 10) = 0.019$ ,  $p = .89$ ,  $\hat{\eta}^2 = .002$ , and no Number of Targets  $\times$  Prevalence interaction,  $F(2, 20) = 0.38$ ,  $p = .69$ ,  $\hat{\eta}^2 = .037$ . Observers seem to have found a second target as effectively as the first.

In Experiment 3, observers could look for tools as a category. In Experiment 4, it is more likely that observers had to search for three distinct categories of target. It appears that observers could maintain three separate criteria, and as in the other low-prevalence tasks, the criterion for the low-prevalence category in this experiment was set at a level that produced a very high miss rate. In practical terms, this means that asking airport screeners to look for hair driers and threats would be unlikely to increase the hit rate for threats. On the other hand, it remains possible that adding artificial threats might make observers more likely to find real threats even if a real and an artificial target happened to appear in the same display (Rubenstein, 2001).

It would be interesting to know if the sensitivity to rare targets is the same as the sensitivity to common targets in this task. As noted, the false alarms in this experiment were rare and might have been related to confusions about which key mapped to which category. With those caveats, it is still possible to compute a  $d'$  measure for the data pooled over observers for each of the three levels of prevalence. Sensitivity was similar for 1% ( $d' = 3.95$ ) and 44% prevalence ( $d' = 4.00$ ). Because these values are near ceiling due to very low false-alarm rates, they should not be given much weight. Nevertheless, we find no support for the idea that observers were less sensitive to low-prevalence targets.

There was also no evidence for a time-on-task effect in this experiment. Over the first 250 trials, observers missed 0.52 of the 1%-prevalence targets, 0.27 of the 10%-prevalence targets, and 0.11 of the 44%-prevalence targets. For the last 250 trials on the second day, those error rates were 0.56, 0.27, and 0.11 respectively.

### Experiment 5: Multiple-Target Trials

In a replication of Experiment 4 designed to produce more multiple-target trials, we used prevalences of 50%, 25%, and 10%. This means that two targets were now present on 16.2% of trials. We tested 11 observers between the ages of 18 and 29 years ( $M$  age = 23.4 years,  $SD = 4.1$  years; 6 women, 5 men) for 100 practice and 1,000 experimental trials.

We replicated the basic prevalence effect, but the critical finding here is that miss error rates were about 2% higher on two-target than on one-target trials,  $F(1, 10) = 5.5$ ,  $p = .039$ ,  $\hat{\eta}^2 = .33$ . Thus, the addition of false targets might reduce, albeit modestly, the detection of real targets if two or more targets could appear on the same trial.

### Experiment 6: Bursts of Higher Prevalence

As discussed above, criterion shift is an important part of the prevalence effect problem. If we could persuade observers to hold

a high-prevalence criterion during low-prevalence search, we would expect to move the miss error rate back down to the rate found with high-prevalence targets. In the remaining experiments reported in this article, we adopted a recalibration strategy. We intermittently showed observers targets under high-prevalence conditions in the hope that they would recalibrate and that the high-prevalence criterion would carry over into low-prevalence search. With Experiment 6, we illustrate a method that did not work. With Experiment 7, we show a method that appears more promising.

In Experiment 6, bursts of higher prevalence were inserted into a low-prevalence search task. On the assumption that low prevalence caused observers to adopt maladaptive search criteria, we hypothesized that occasional intervals of higher prevalence would recalibrate observers, allowing them to maintain a more liberal criterion setting, thus reducing the miss rate at low prevalence.

### Method

#### Observers

Ten participants between the ages of 18 and 51 years participated ( $M$  age = 26.3 years,  $SD$  = 10.4 years; 3 women, 7 men).

#### Procedure

Stimuli were same as those used in Experiments 3–5. As in Experiment 3, observers searched for tools. Observers were tested for 2,000 trials at 1% prevalence. Into those 2,000 trials, we inserted six 60-trial bursts for a total of 2,360 trials. Over the course of each burst, the probability of target presence rose from 1% to 50% and then fell back to 1% in a sinusoidal manner. Start times of bursts were randomized. Because the appearance of a target was probabilistically determined on each trial, it was sometimes possible to get an extended run of target-present trials during a burst. Figure 12 shows a typical block of 2,360 trials for one observer. Target prevalence is calculated as the percentage of target-present trials in the preceding 20 trials. Open symbols indicate miss errors. Black ovals are hits. A symbol's vertical position indicates the prevalence over the last 20 trials prior to that target-present trial.

Observers were given feedback after each trial. Methods were otherwise similar to Experiments 1 and 2.

### Results and Discussion

Trials with RTs greater than 8,000 ms and less than 200 ms were removed from analysis. Two observers were excluded from further

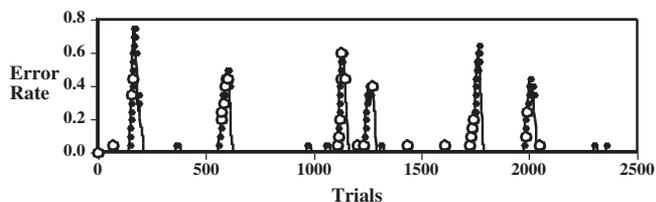


Figure 12. Target prevalence (averaged over the preceding 20 trials) as a function of trial number for a typical observer in Experiment 6. Open symbols indicate miss errors. Filled ovals are hits.

analysis because they committed 7% and 13% of these types of errors, respectively. After removing those two observers, only 0.7% of trials were either too long or too short. False-alarm errors were exceedingly rare in this version of the task (17 out of 17,654 target-absent trials). Observers simply did not mistake these objects for one another. Accordingly, it is not possible to calculate  $d'$  or  $c$  values with any confidence.

Figure 13 shows the miss error rates for trials in and out of a high-prevalence burst.

The manipulation of prevalence had a strong effect. During periods of elevated target prevalence, miss error rates dropped to an average of 0.18 (0.02). Regrettably, this did not eliminate the prevalence effect seen in previous experiments. Miss error rates at low prevalence averaged 0.41 (0.06), comparable to low-prevalence results in other studies with these stimuli. An ANOVA with set size and prevalence (in or out of a high-prevalence burst) showed main effects of prevalence,  $F(1, 7) = 18.6, p = .003, \eta^2 = .73$ , and set size,  $F(1, 3) = 4.6, p = .012, \eta^2 = .40$ , but no interaction,  $F(3, 21) = 0.9, p = .44, \eta^2 = .12$ .

Though we cannot reliably compute signal detection measures in the absence of false alarms, there is no reason to imagine that this version of the task is fundamentally different from the others. Presumably, observers were changing criteria in response to target prevalence. Figure 14 shows that observers made those changes very quickly. The figure plots average RT on correct target-absent trials as well as miss error rate in the last 20 trials as a function of target prevalence computed over 20 trial bins.

As recent prevalence went up, RT increased by about 500 ms on correct target-absent trials. At the same time, the chance of making an error dropped from almost 0.50 to less than 0.15. This would be a standard speed-accuracy trade-off were it not for the suspicion that, as in Experiments 1 and 2, the speeding of responses probably did not produce a loss of sensitivity. If we could measure  $d'$  in this case, we would expect that it would remain high. In any case, there is no evidence of a beneficial carryover from periods of high prevalence into periods of low prevalence.

#### Experiment 7: “Curing” the Prevalence Effect—The Role of Feedback

The results of Experiment 6 indicated that observers rapidly adjust criterion as prevalence changes. How does an observer know that target prevalence has changed? In the experiments reported so far, observers could estimate prevalence from the feedback given after each trial. In real-world screening situations (baggage, medical, etc.), feedback is much less reliable. In pilot experiments, we tested observers entirely without feedback. This did not appear to produce the desired reduction in miss errors (and was rather frustrating for our observers). The burst paradigm of Experiment 6 suggested a different approach. If we conceptualized our experiments in terms of a real-world screening task, we could think of low-prevalence blocks as real screening trials and high-prevalence bursts as retraining episodes. The paradigm of Experiment 7 was to provide no explicit feedback during low-prevalence trials (mimicking the real-world situation) but full feedback during high-prevalence bursts (which could be implemented during retraining). Additionally, we added measures of subjective alertness and objective vigilance to determine to what extent sleepiness or attentional lapses might be contributing to the prevalence effect.

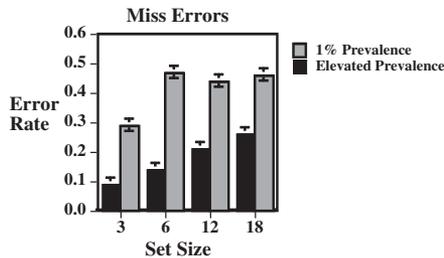


Figure 13. Miss error rates as a function of set size for low (1%) prevalence and higher prevalence periods in Experiment 6. Error rates are back-transformed from arcsine-transformed errors. Error bars represent within-subject confidence intervals.

### Method

#### Observers

Fourteen observers between the ages of 21 and 47 years participated ( $M$  age = 32.6 years,  $SD$  = 9.1 years; 7 women, 7 men).

#### Procedure

The experiment used the realistic baggage stimuli introduced in Experiment 1. Some observers were tested with the same set of bag images used in Experiment 1. Others used a newly generated set. There was no apparent difference in performance as a function of stimulus set. In the basic screening task, as in previous experiments, observers were instructed to indicate as quickly and accurately as possible whether a target was present or absent. Each observer was tested on a total of 1,700 trials. First came 300 training trials at 50% prevalence with full feedback. The main portion of the experiment consisted of five repetitions of a sequence of 200 trials at 2% prevalence with no feedback, followed by a burst of 40 trials at 50% prevalence with full feedback. This yielded a sequence of 1,400 trials with an overall prevalence of 8.9%. Observers were informed that they would receive no feedback on the majority of trials but that there would be occasional short blocks of trials during which they would receive feedback. They were instructed to think of these periods as though they might be retraining or evaluative episodes inserted by airport administration to ensure screening quality. Every 240 trials, a minimum 2-min break was enforced.

This experiment did not include the motivational point system described for Experiment 1. After making a response on a no-feedback trial, observers saw the word “Thanks,” as well as the trial number. On feedback trials, if a target was present and the observer responded, “Yes,” the target was outlined on the screen, and the observer was presented with the message “Good for you. You found the target. Take a look and then press a key to continue.” If the target was missed, it was then outlined on the screen, and the message read, “You missed the target. Take a look and then press a key to continue.” If observers correctly identified a target as absent, they were presented with the message “Correct—No target present,” and if observers incorrectly identified a target as present, they were presented with the message “False alarm—No target present.”

### Alertness Rating and Vigilance Tasks

This experiment measured subject alertness with a rating task and vigilance with the PVT (Dinges et al., 1994). The subjective alertness task asked observers to use the mouse to click somewhere on a horizontal line (15.54 cm in length) representing a continuum of their overall subjective feeling of alertness. On the left extreme of the continuum was the word “Alert,” and on the right extreme was the word “Sleepy.” Observers were instructed to click on the line at a point representing their subjective alertness at that moment. The alertness rating task was not timed, and observers could correct their response if they felt they had inadvertently clicked in the wrong place. Observers had to click on a button at the bottom of the screen to continue once they were satisfied with their response.

After giving an alertness rating, observers performed the PVT. They monitored a 2° white-outlined square on a black background in the center of the computer monitor and were instructed to press the space bar of the computer keyboard as soon as the square filled with white (the probe). After response, the square reverted to its unfilled state, and observers were informed about their RT (in ms) for that particular keypress for 500 ms. Probes were separated by random intervals between 1,000 and 9,000 ms. If observers responded when there was no probe present, the computer beeped and the word “Anticipation” flashed on the screen in red for 500 ms. The PVT lasted for 3 min.

The alertness rating and PVT sequence was performed seven times during the experiment: once at the beginning of the experiment, once at the end of the high-prevalence block, and once prior to each burst (i.e., five times during the low-prevalence block). The enforced 2-min breaks followed each of the sequences. After each sequence and break, the main baggage screening task resumed.

### Results

In this experiment, we succeeded in curing the prevalence effect. As shown in Figure 15, there were no significant effects of prevalence. We divided the data into the five 240-trial epochs and performed an ANOVA with factors of epoch and prevalence. There were no significant effects. For miss errors, the main effects of prevalence and epoch and the interaction all yielded  $F$ s < 1.00,  $p$ s > .5, and  $\eta^2$ s < .04. Within-subject confidence intervals for

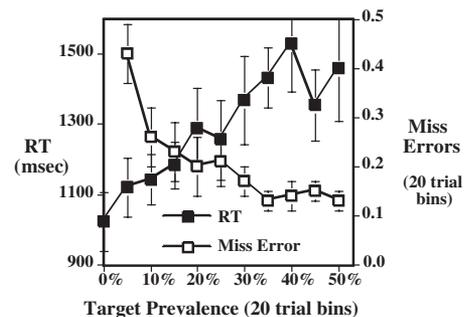


Figure 14. Reaction time (RT) for correct rejections (solid symbols) and miss error rate (open symbols) as a function of target prevalence (computed over 20 trial bins) in Experiment 6. Error bars are  $\pm 1$  SEM.

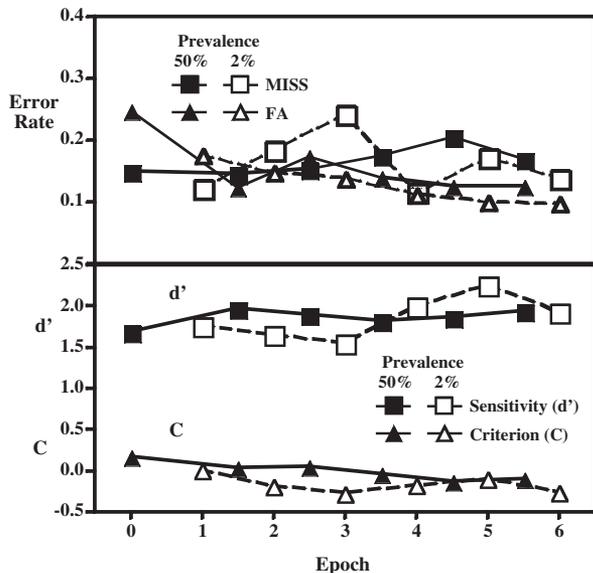


Figure 15. Results for Experiment 7. Top panel shows miss and false-alarm (FA) rates as a function of time. Epoch 0 is the initial practice block. Each subsequent epoch consists of 200 low-prevalence trials (open symbols), followed by 40 high-prevalence trials (filled symbols). Note that there were no significant effects of prevalence in this case. Bottom panel shows that  $d'$  and  $c$  are similar for low- and high-prevalence trials. Within-subjects confidence intervals for the miss error analysis of variance were  $\pm .034$  and  $\pm .006$  for the false alarms.

this ANOVA were  $\pm 0.034$ . For false alarms, the main effects of prevalence and epoch and the interaction all yielded  $F$ s  $< 1.25$ ,  $p$ s  $> .25$ , and  $\hat{\eta}^2$ s  $< .09$ . Within-subject confidence intervals for this ANOVA were  $\pm 0.006$ .

Pooling across epochs, the average miss rate for low prevalence was 0.21 versus 0.18 for high prevalence. The difference was not significant,  $t(13) = 1.05$ ,  $p = .312$ . Similarly, the difference in average false-alarm rates was not significant (low prevalence = 0.13, high prevalence = 0.15),  $t(13) = 1.32$ ,  $p = .21$ .

We can compare these error rates with the error rates in Experiment 1a, a similar study with very similar stimuli, albeit with different observers. This comparison is shown in Figure 16 for average error rates.

The high-prevalence results were essentially the same in the two experiments—*independent sample t* tests: miss errors,  $t(36) = 1.00$ ,  $p = .32$ ; false alarms,  $t(36) = 0.22$ ,  $p = .83$ . The low-prevalence results were dramatically different. Miss error rates were much reduced in Experiment 7,  $t(36) = 4.68$ ,  $p = .00$ , and false alarms were correspondingly increased,  $t(36) = 11.49$ ,  $p = .00$ . The insertion of high-prevalence retraining bursts allowed observers in Experiment 7 to maintain the high-prevalence criterion under low-prevalence conditions. This can be seen in Figure 17. Unlike other experiments reported in this article, in Experiment 7, criterion remained about the same across epoch and independent of target prevalence. Criterion and sensitivity measures in Figure 15 were derived from data pooled over observers because low prevalence produced too many extreme values in individual observer data (e.g., cells where the miss rate is 0.00 or 1.00). Thus, statistical tests are not appropriate. The statistical conclusions can be derived from analysis of error rates.

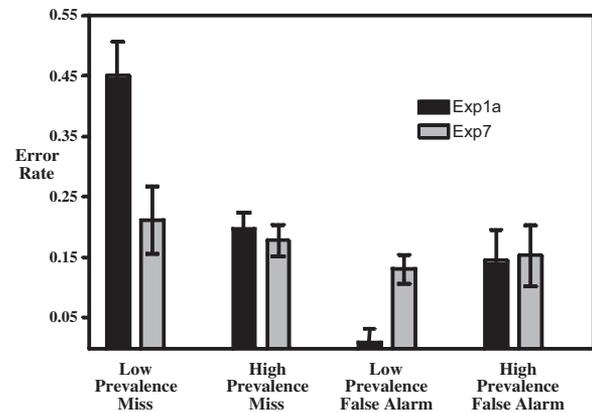


Figure 16. Average miss and false-alarm rates for Experiments 1a and 7. Error bars are  $\pm 1$  SEM of the difference.

Briefly turning to the vigilance and sleepiness measures, there was no significant change in RT on the vigilance task over trials,  $F(6, 78) = 1.272$ ,  $p = .280$ ,  $\hat{\eta}^2 = .089$ , nor was there an effect on number of errors of anticipation,  $F(6, 78) = 0.769$ ,  $p = .597$ ,  $\hat{\eta}^2 = .056$ . There was a decline in self-rated alertness from high alertness (90 out of 100) to middling alertness (50 out of 100) over the course of the experiment,  $F(6, 78) = 9.958$ ,  $p < .0001$ ,  $\hat{\eta}^2 = .434$ . The decline in subjective alertness was not accompanied by a decline in search performance. Undoubtedly, fatigue would modulate performance, but there is no evidence that it was a factor in this experiment.

### Discussion

In Experiment 7, unlike in previous experiments, our effort to counteract the prevalence effect was a success. Providing high-prevalence trials with feedback amidst low-prevalence trials without feedback appears to induce observers to maintain a high-prevalence criterion during extended periods of low prevalence. From a practical vantage point, this is very encouraging because these feedback conditions are ecologically plausible. In baggage screening or medical screening, the basic task is a low-prevalence search with little or no feedback. Periodic retraining at high prevalence with full feedback is easy to implement because retraining would involve stimuli where ground truth is known.

This single experiment cannot specify the ideal length of a high-prevalence burst or the duration of the therapeutic effects of that burst. However, it does appear to identify a method that could be fine-tuned for specific tasks.

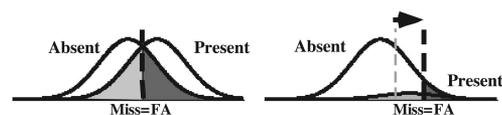


Figure 17. Criterion and prevalence: If observers try to equate the number (not the proportions) of miss and false-alarm (FA) errors, they set a more liberal criterion at low prevalence (right side) than at high prevalence (left side).

## General Discussion

Three broad conclusions can be drawn from the seven experiments reported here. First, the prevalence effect described by Wolfe et al. (2005) is robust and persistent. It occurs under a wide range of different laboratory conditions. In other work, we have reproduced the effect with much simpler stimuli, such as a classic visual search for a T among Ls (Rich, Hidalgo-Sotelo, Kunar, Van Wert, & Wolfe, 2006). In the work presented here, strong prevalence effects were seen in all experiments except Experiment 7. Second, like similar target frequency effects elsewhere, the prevalence effect in visual search is accompanied by a shift in decision criterion rather than a loss in sensitivity. There is evidence in other work (e.g., Fleck & Mitroff, in press) of a speed-accuracy trade-off account of the prevalence effect, but it is hard to see how this could explain the present results, notably, the failure to find a sensitivity change and the failure to improve matters by slowing responses (Experiment 2). Third, manipulating the feedback available to the observer (Experiment 7) seems to be an effective way to manipulate criterion in search tasks. The prevalence effect is of both theoretical and practical interest, and we end this article with a few comments about each aspect.

### Modeling the Prevalence Effect

We have created a model of the effects of prevalence on errors based on three assumptions.

#### Assumption One: Changes in Target Prevalence Cause Changes in Criterion and Not in Sensitivity

The present experiments show that this classic decision theory finding is true in visual search. This is not a trivial finding because visual search is not a simple two-alternative forced-choice task. In visual search, at least for inefficient searches with long exposure durations, the observer is actually faced with a series of three-alternative choice points for each object or region attended: the observer can respond “target present” or “target absent” or choose to continue searching. In this context, it is interesting to note that low prevalence is accompanied by a speeding of responses. Apparently, observers are less inclined to chose the *continue search* option. Surprisingly, this speeding of the response is not accompanied by a loss of sensitivity.

#### Assumption Two: To Set a Criterion, Observers Try to Equate the Number of Miss and False-Alarm Errors

How did observers set criterion in these experiments? As noted earlier, the classic definition of optimal setting of the criterion is based on the ratio of target-present to target-absent trials. However, especially at very low prevalence, this predicts too many errors (Green & Swets, 1967; Maddox, 2002). Moreover, observers did not have direct access to that ratio information in these experiments. What observers could access was implicit or explicit feedback about their performance (*implicit feedback* here refers to the self-generated feedback that one might get when, e.g., one knows one has found a target even if no one provides explicit confirmation of that fact).

Suppose observers tried to equalize the number (not the proportion) of miss and false-alarm errors. At 50% prevalence, that

would produce a neutral criterion, illustrated on the left of Figure 17.

If we take Figure 17 as a histogram rather than a probability density function, low prevalence can be illustrated by the smaller target-present distribution in the right-hand panel. Equating counts of misses and false alarms now requires a liberal criterion that will result in a high proportion of rare targets being missed.

#### Assumption Three: Observers Are Operating on an ROC Function With a Slope Less Than 1.0

Like the ratio model, a model that equates the number of miss and false-alarm errors predicts too many miss errors at low prevalence. However, the data provide one more useful constraint. As noted earlier, for those experiments for which sensitivity in the form of  $d'$  can be calculated,  $d'$  tends to rise as prevalence falls. It seems unlikely that sensitivity is actually higher at low prevalence. When Kundel (2000) obtained similar results in his retrospective survey of chest X-ray data, he worked from the assumption that prevalence was shifting criterion and thus shifting performance along an ROC curve with the same sensitivity at low and high prevalence. If we make the same assumption about the present data, then the low- and high-prevalence data points lie on one ROC. ROCs form straight lines when plotted in a  $z$ -transformed space. The results of Experiments 1 and 2, plotted in that manner, are shown in Figure 18.

Except for the case of two observers viewing the display together (Experiment 1b), the results for different experiments all seem to lie on an ROC with a slope of about 0.6 in  $z$ -transformed space. The standard case of equal variance signal and noise distributions (as shown in Figure 17a) yields an ROC with a slope of unity in this space. In the usual analysis, slopes less than unity are assumed to arise when the variance of the noise distribution is less than that of the signal distribution. It is not obvious what this means in the case of X-ray images of baggage. However, it is clear that this is a reproducible result. Interestingly, the difference between the Experiment 1b data (gray squares in Figure 18) and the results of the other experiments is largely due to a shift in the high-prevalence data point.

Perhaps the nonunit slope of the ROC is the result of the complex nature of the search from a signal detection point of view. Previously, we described a visual search trial as a series of three-

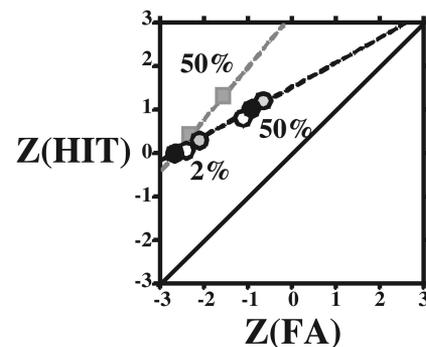


Figure 18.  $z$ -transformed average data from Experiments 1a—individual (open circle), 1a—paired (gray filled circle), 1b—shared (gray squares), and 2a (black circles). FA = false alarm.

alternative forced-choice tasks. Another way to think about the task is as a pair of two-alternative forced-choice tasks. One repeated two-alternative forced-choice decision has *target-present* as one alternative and *keep searching* as the other. The second decision is a search termination decision with *keep searching* as one alternative and *quit* (typically with a target-absent response) as the other. Many researchers have applied signal detection models to data from briefly presented search stimuli where search termination is not an issue (Baldassi & Verghese, 2002; Davis, Kramer, & Graham, 1983; Palmer, 1995; Palmer, Verghese, & Pavel, 2000). Other work has modeled search termination (Chun & Wolfe, 1996; Cousineau & Shiffrin, 2004; Hong, 2005; Zenger & Fahle, 1997). We are currently working on the details of a model that combines those two decisions into a single RT model (see also Thornton & Gilden, 2007).

Figure 19 presents the results of a simulation embodying the three assumptions laid out above: Errors are based on criterion position, criterion position is based on an effort to equate miss and false-alarm numbers, and the underlying ROC has a slope of about 0.6. Note that the error rates are very similar to those shown in, for example, Figure 16 for Experiment 1a. At low prevalence, miss error rates were above 0.40, falling to less than 0.20 at 50% prevalence. At the same time, false alarms were very rare at low prevalence and rose to equal the miss error rate at 50% prevalence. Thus, a model based on these three assumptions can capture the basic prevalence effect.

### Practical Considerations

One of the most striking findings in these data is that the high miss rates at low prevalence were not the product of careless or lazy observers. Observers were doing the task we set for them. If one wanted to reduce miss errors due to the prevalence effect, the path to the cure would be to shift criterion back to a less extreme position. A standard way to change criterion is to change the payoff matrix. As we noted in the introduction, payoff manipulations are less effective at shifting criterion than even a change to 25% prevalence (Maddox, 2002), so finding a payoff sufficient to offset 1%–2% prevalence may not be possible in the laboratory. In the field, it is possible that threats of job loss may be sufficient incentive to move criterion (Adrian Schwanger, personal communication, November 29, 2006).

As described in Experiment 7, our most successful effort to manipulate criterion involved bursts of high prevalence with feed-

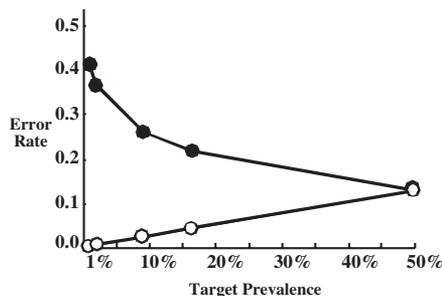


Figure 19. Results of a simulation of the effects of target prevalence on errors based on the three assumptions outlined in the text. Open circles represent false alarms. Closed circles represent miss errors.

back interspersed with low-prevalence trials without feedback. This suggests that it might be useful to insert a short period of high-prevalence retraining prior to each longer period performing a low-prevalence screening task. Experiments involving professionals performing these low-prevalence tasks will be needed before it can be stated with any certainty that the prevalence effect is a problem in real-world settings and whether our proposed retraining intervention (or any other intervention) would be effective.

### References

- Baldassi, S., & Verghese, P. (2002). Comparing integration rules in visual search. *Journal of Vision*, 2, 559–570.
- Beck, L. H., Bransome, E. D., Jr., Mirsky, A. F., Rosvold, H. E., & Sarason, I. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, 20, 343–350.
- Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Caldwell, R. T., & Krupinski, E. A. (2000). Role of faulty decision making in the satisfaction of search effect in chest radiography. *Academic Radiology*, 7, 1098–1106.
- Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., et al. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology*, 25, 133–140.
- Bond, A. B., & Kamil, A. C. (2002, February 7). Visual predators select for crypticity and polymorphism in virtual prey. *Nature*, 415, 609–613.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 443–446.
- Broadbent, D. E. (1964). Vigilance. *British Medical Bulletin*, 20, 17–20.
- Broadbent, D. E., & Gregory, M. (1965). Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Human Factors*, 7, 155–162.
- Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, 30, 39–78.
- Colquhoun, W. P. (1961). The effect of “unwanted” signals on performance in a vigilance task. *Ergonomics*, 4, 41–51.
- Cousineau, D., & Shiffrin, R. M. (2004). Termination of a visual search with large display size effects. *Spatial Vision*, 17, 327–352.
- Davis, E., Kramer, P., & Graham, N. (1983). Uncertainty about spatial frequency, spatial position, or contrast of visual patterns. *Perception & Psychophysics*, 33, 20–28.
- Dinges, D. F., Gillen, K. A., Powell, J. W., Carlin, M., Ott, G. E., Orne, E. C., et al. (1994). Discriminating sleepiness by fatigability on a psychomotor vigilance task. *Sleep Research*, 23, 129.
- Eggin, T. K., & Feinstein, A. R. (1996). Context bias: A problem in diagnostic radiology. *JAMA*, 276, 1752–1755.
- Ethell, S. C., & Manning, D. (2001). Effects of prevalence on visual search and decision making in fracture detection. In E. A. Krupinski & D. P. Chakraborty (Eds.), *Medical Imaging 2001: Image perception and performance* (pp. 249–257). Bellingham, WA: SPIE.
- Fleck, M., & Mitroff, S. (in press). Rare targets rarely missed in correctable search. *Psychological Science*.
- Green, D. M., & Swets, J. A. (1967). *Signal detection theory and psychophysics*. New York: Wiley.
- Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., et al. (2003). Prevalence effect in a laboratory environment. *Radiology*, 228, 10–14.
- Gur, D., Sumkin, J. H., Rockette, H. E., Ganott, M., Hakim, C., Hardesty, L., et al. (2004). Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *Journal of the National Cancer Institute*, 96, 185–190.
- Healy, A. F., & Kubory, 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, 344–354.
- Hong, S.-K. (2005). Human stopping strategies in multiple-target search. *International Journal of Industrial Ergonomics*, 35, 1–12.
- Hsieh, P. C., Chu, C. L., Yang, Y. K., Yang, Y. C., Yeh, T. L., Lee, I. H., et al. (2005). Norms of performance of sustained attention among a community sample: Continuous performance test study. *Psychiatry and Clinical Neurosciences*, 59, 170–176.
- Kribbs, N. B., & Dinges, D. F. (1994). Vigilance decrement and sleepiness. In R. D. Ogilvie & J. R. Harsh (Eds.), *Sleep onset: Normal and abnormal processes* (pp. 113–125). Washington, DC: American Psychological Association.
- Kundel, H. L. (2000). Disease prevalence and the index of detectability: A survey of studies of lung cancer detection by chest radiography. In E. A. Krupinski (Ed.), *Medical imaging 2000: Image perception and performance* (pp. 135–144). Bellingham, WA: SPIE.
- Li, H., Li, F., Gao, H. H., Chen, A., & Lin, C. (2006). *Appropriate responding can reduce miss errors in visual search*. Unpublished manuscript.
- Mackworth, J. (1970). *Vigilance and attention: A signal detection approach*. Harmondsworth, England: Penguin Books.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory*. Mahwah, NJ: Erlbaum.
- Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, 78, 567–595.
- Nodine, C. F., Krupinski, E. A., Kundel, H. L., Toto, L., & Herman, G. T. (1992). Satisfaction of search (SOS). *Investigative Radiology*, 27, 571–573.
- Olendorf, R., Rodd, F. H., Punzalan, D., Houde, A. E., Hurt, C., Reznick, D. N., et al. (2006, June 1). Frequency-dependent survival in natural guppy populations. *Nature*, 441, 633–636.
- Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set size effect. *Current Directions in Psychological Science*, 4, 118–123.
- Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, 40, 1227–1268.
- Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. R. Boff, L. Kaufmann, & J. P. Thomas (Eds.), *Handbook of human perception and performance: Vol. 2. Cognitive processes and performance* (1–39). New York: Wiley.
- Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., et al. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, 353, 1773–1783.
- Rich, A. N., Hidalgo-Sotelo, B., Kunar, M. A., Van Wert, M. J., & Wolfe, J. M. (2006, May). *What happens during search for rare targets? Eye movements in low prevalence visual search*. Paper presented at the annual meeting of the Vision Sciences Society, Sarasota, FL.
- Rubenstein, J. (2001). *Test and evaluation plan: X-Ray Image Screener Selection Test* (No. DOT/FAA/AR-01/47). Washington, DC: Office of Aviation Research.
- Samuel, S., Kundel, H. L., Nodine, C. F., & Toto, L. C. (1995). Mechanism of satisfaction of search: Eye position recordings in the reading of chest radiographs. *Radiology*, 194, 895–902.
- Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General*, 134, 443–460.
- Smith, J. D., Redford, J. S., Washburn, D. A., & Tagliabue, L. A. (2005). Specific-token effects in screening tasks: Possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1171–1185.
- Smith, P. A., & Turnbull, L. S. (1997). Small cell and “pale” dyskaryosis. *Cytopathology*, 8, 3–8.
- Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review*, 114, 71–103.
- Treisman, M. (1984). A theory of criterion setting: An alternative to the attention band and response ratio hypotheses in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: General*, 113, 443–463.
- Warm, J. S. (1993). Vigilance and target detection. In B. M. Huey & C. D. Wicken (Eds.), *Workload transition: Implications for individual and team performance* (pp. 139–170). Washington, DC: National Academy Press.
- Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005, May 26). Rare items often missed in visual searches. *Nature*, 435, 439–440.
- Zenger, B., & Fahle, M. (1997). Missed targets are more frequent than false alarms: A model for error rates in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1783–1791.

Received July 17, 2006

Revision received March 8, 2007

Accepted March 12, 2007 ■