Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too

Jeremy M. Wolfe
Visual Attention Lab, Brigham and Women's Hospital, Cambridge, MA, USA
Harvard Medical School, Boston, MA, USA

David N. Brunelli
Chelsea Regional Training Center, Transportation Security Administration, Chelsea, MA, USA

Joshua Rubinstein
Army Research Lab – HRED, ARDEC Field Element, Picatinny Arsenal, NJ, USA

Todd S. Horowitz
Visual Attention Lab, Brigham and Women's Hospital, Cambridge, MA, USA
Harvard Medical School, Boston, MA, USA

Many socially important search tasks are characterized by low target prevalence, meaning that targets are rarely encountered. For example, transportation security officers (TSOs) at airport checkpoints encounter very few actual threats in carry-on bags. In laboratory-based visual search experiments, low prevalence reduces the probability of detecting targets (Wolfe, Horowitz, & Kenner, 2005). In the lab, this "prevalence effect" is caused by changes in decision and response criteria (Wolfe & Van Wert, 2010) and can be mitigated by presenting a burst of high-prevalence search with feedback (Wolfe et al., 2007). The goal of this study was to see if these effects could be replicated in the field with TSOs. A total of 125 newly trained TSOs participated in one of two experiments as part of their final evaluation following training. They searched for threats in simulated bags across five blocks. The first three blocks were low prevalence (target prevalence $\leq$ .05) with no feedback; the fourth block was high prevalence (.50) with full feedback; and the final block was, again, low prevalence. We found that newly trained TSOs were better at detecting targets at high compared to low prevalence, replicating the prevalence effect. Furthermore, performance was better (and response criterion was more "liberal") in the low-prevalence block that took place after the high-prevalence block than in the initial three low-prevalence blocks, suggesting that a burst of high-prevalence trials may help alleviate the prevalence effect in the field.

Introduction

The search for threats in carry-on luggage is an important aspect of airport security. It is also a very difficult visual search task. In typical visual search tasks in the laboratory, observers look for a target item among distractor items. After several decades of research, the factors that modulate search difficulty are well established (Kundel, 2004; Wolfe, 2010; Wolfe & Reynolds, 2008). Search is easiest if the target is defined by a unique basic feature like color, size, or orientation (Egeth, Jonides, & Wall, 1972; Neisser, 1963). As the difference between the target and distractors gets smaller (or less salient), the search gets harder (Nagy & Sanchez, 1990; Nothdurft, 2000). As distractors become more heterogeneous, the search gets harder (Duncan & Humphreys, 1989). As the target definition becomes more abstract, search gets harder (J. Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004). When the target is not defined by a unique feature (e.g., when it is an arbitrary object), search is inefficient (Vickery, King, & Jiang, 2005). It is harder to find the target if the scene is cluttered and/or the neighborhood around the target is cluttered (Beck, Lohrenz, & Trafton, 2010; Rosenholtz, Li, & Nakano, 2007). Finally, it is harder to find an object if the clutter is random, compared to being a

Citation: Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, *13*(3):33, 1–9, http://www.journalofvision.org/content/13/3/33, doi:10.1167/13.3.33.

doi: 10.1167/13.3.33 Received February 4, 2013; published December 2, 2013 ISSN 1534-7362

feature of a meaningful scene (Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinsky, 2006).

On all of these counts, the x-ray screener task (XRST) is hard. The target is broadly defined (and changes over time). Target items are not always marked by unique features. Distractors are very diverse and, in some cases, very similar in appearance to possible targets. Objects do not appear in well-structured "scenes" but in cluttered luggage x-rays. Packed bags follow some rules, but they do not provide the sort of guidance that a natural scene or even a medical x-ray provides. Moreover, x-rays add the challenges of object transparency and overlap.

As if these factors did not render the task challenging enough, transportation security officers (TSOs) are looking for targets that are very rare. The probability of finding a real threat would be vanishingly rare for an average TSO on an average day although TSOs do confront a higher prevalence of prohibited items, like water bottles, and other items requiring action, such as opaque regions that could hide a threat. The "prevalence" of threat images is increased because target items are projected into the image by the checkpoint x-ray itself. This is a quality-control measure known as threat image projection (TIP). The target prevalence rate for a TSO performing his task would be the TIP rate plus the very small true threat rate. The rate with which TIP images are inserted into images is security-sensitive information, but it is fair to say that it is well below the 0.5 target prevalence rate that would be common in a typical laboratory study.

In prior research with volunteer, non-TSO observers, it was found that prevalence has a profound effect on the pattern of errors in search tasks. Specifically, miss errors are much more common at low prevalence (Wolfe et al., 2005). In the original study, miss errors rose from about 7% at 0.5 prevalence to 30% at 0.02 prevalence. Fleck and Mitroff (2007) argued that these were essentially motor errors with which observers, who had gotten into the habit of responding "no," mistakenly responded "no" when they meant "yes." In their experiment, the prevalence effect was abolished when observers were given the chance to "call back" an error that they noticed. If true, this would render the prevalence effect uninteresting in an airport setting. If a TSO sees a gun but erroneously pushes the "clear" button, she would still have the opportunity to correct her error and stop the bag from leaving the x-ray chamber. Such motor errors do occur. However, studies have found that if the task is reasonably hard (e.g., a search for guns and knives in realistic baggage stimuli), the prevalence effect does not disappear if a "call back" option is provided (Van Wert, Horowitz, & Wolfe, 2009).

With tasks that produce false-alarm errors, prevalence effects appear to be primarily *criterion shifts*, to use the terms of signal-detection theory (Green & Swets, 1967). That is, while miss errors increase at low prevalence, false-alarm errors decrease. Sensitivity (measured by signal-detection theory parameter d') remains roughly constant while criterion, indicating the likelihood of a "target present" response, changes significantly (Wolfe et al., 2007). Change in criterion with target or signal frequency has been seen repeatedly in nonsearch tasks (Colquhoun, 1961; Healy & Kubovy, 1981). The fact that miss and false-alarm errors trade off is not particularly comforting if miss errors are far more "expensive" than false alarms. Observers will respond to differential payoffs for different types of errors (Maddox, 2002; Navalpakkam, Koch, & Perona, 2009). Nevertheless, target prevalence has a powerful effect on search behavior, at least in laboratory-based search tasks. Moreover, knowledge of the probability of a target in the next bag is not enough to suppress the prevalence effect (Ishibashi, Kita, & Wolfe, 2012; Lau & Huang, 2010).

A similar phenomenon has been observed in the vigilance literature under the heading of "signal-probability effects." It has been widely demonstrated that low signal probability reduces hit rates in classical low event-rate, low cognitive-load vigilance tasks by shifting criterion rather than by decreasing sensitivity (Baddeley & Colquhoun, 1969; Davies & Parasuraman, 1982; Parasuraman & Davies, 1976; Williges, 1973), and these effects are accompanied by a slowing of "yes" reaction times and speeding of "no" reaction times (Parasuraman & Davies, 1976), similar to the pattern we observe in visual search (Wolfe et al., 2005; Van Wert, Horowitz, & Wolfe, 2009). Thus, prevalence effects may be a more general phenomenon, applying across many cognitive domains. However, the prevalence effects we observe in the search context are unlikely to be simple examples of signal-probability effects in vigilance. In general, in vigilance paradigms, as the task becomes more complex and demanding, signal probability affects sensitivity rather than criterion (See, Howe, Warm, & Dember, 1995). Factors that make a vigilance task more demanding include, inter alia, memory load (e.g., having to compare a stimulus with a standard in memory) and perceptual degradation of the stimulus (e.g., Nuechterlein, Parasuraman, & Jiang, 1983). For example, Matthews (1996) found sensitivity effects, rather than effects of signal probability, on criterion in a task that required participants to respond only to the digit "0" among other digits with 30% added pixel noise. It is reasonable to suppose that the memory demands of the XRST are at least an order of magnitude greater than those in Matthews' experiment. When we also consider the perceptual difficulty, a purely vigilance-based account of prevalence effects would predict an effect on sensitivity rather than criterion in the XRST, which is the opposite of what we

have observed. Moreover, it is important to remember that in the XRST the observer must actively dismiss each stimulus rather than, for example, failing to notice it as it passes by.

Prevalence effects are scientifically interesting, giving us insight into how observers adjust their quitting criteria. However, for practical application, the critical question is whether prevalence would have a similar effect on professional searchers performing in their domain of expertise. Would TSOs, looking at baggage, miss more threats at low target prevalence than at higher prevalence? The answer to that question is "yes." In this study, newly trained TSOs examined test images of bags with and without inserted threat images. Threats were inserted at low or high prevalence rates. Lower prevalence produced higher miss error rates. This being the case, one would want to ameliorate the problem. Laboratory tests have found that a period of high-prevalence testing with good feedback to the observer reduces the miss rate for a subsequent period of low prevalence with poor feedback (Wolfe et al., 2007). This study finds some evidence that this might work with expert populations as well.

Method

These data were collected at the TSA training facility at Chelsea, Massachusetts. Observers were trainees at the end of their training to be TSOs. Training equipment presents images of luggage with and without threats inserted in them. At the end of training, students are given various tests of their proficiency. To accommodate this study, students performed five blocks of 80 (or, in some cases, 100) bags per block (one bag = one trial). Blocks 1, 2, 3, and 5 were low-prevalence blocks. Block 4 was a high-prevalence block. Three sets of bag images were created for each of the five blocks. An observer saw one of those three sets. The order of blocks—*low, low, low, high, low*—was fixed. However, the order of the low-prevalence blocks was varied so that different sets of low-prevalence bags could be shown in the final block. Low-prevalence blocks contained three or four targets (prevalence = 0.0375 or 0.05). High-prevalence blocks had prevalence rates of 0.5. One type of high-prevalence block had 100 rather than 80 bags (50 targets). Within a block, the order of bags was randomized. Thus, while all observers who saw block X would have seen the same bags, they would not have seen them in the same order.

In the low-prevalence blocks, targets were rare, and no feedback was given. In the high-prevalence block, targets were frequent, and feedback was provided on every trial. The primary purpose of this design was to test the hypothesis that high-prevalence training with feedback would reduce miss errors on subsequent low-prevalence trials as has been shown in the laboratory (Wolfe et al., 2007).

A secondary consideration of the design was ecological validity or representativeness. The low-prevalence blocks mimicked working conditions at the airport. Two caveats should be kept in mind concerning the low-prevalence blocks: (a) The specific value of prevalence was not intended to match prevalence at the checkpoint. Targets were merely rare. (b) At the checkpoint, TSOs would get feedback about TIP trials but not about the vast set of other bags. In this study, no feedback was given in the low-prevalence blocks. Meanwhile, the high-prevalence blocks mimicked typical training conditions in which prevalence would be high relative to airport conditions, and feedback would be present. The present design does not allow us to independently assess the effects of manipulating prevalence and feedback. While it would be valuable to test the effects of each factor separately, that was not possible in this context. Note that, in addition to feedback about accuracy, observers could also utilize a "bag file" feature to get a list of the bag's contents along with a diagram with arrows pointing out certain elements. On each trial that contained a threat, they could also access a magnified image of the threat all by itself, both as a photograph and as an x-ray image. Observers completed all blocks in a single day, generally spending about three hours on the five blocks.

Observers

Data were obtained from 102 TSO trainees as part of their training. The test blocks reported here were deemed to fall into the category of quality-assurance and quality-improvement measures that do not require formal consent procedures beyond the consent to be trained as a TSO. Nevertheless, data were de-identified for the analysis reported here, and no individual observer results were reported from this analysis to TSA staff (although, of course, TSA has access to the original results that they collected as part of the observers' training). Twenty-three of the data sets were unusable (incomplete data, incorrect order of testing), leaving 79 observers. All observers met the basic vision and color vision qualifications for TSO training. TSO applicants were aged 18–70. All of the observers had passed a basic computer-based x-ray assessment test prior to any initial training. This is a requirement for entrance into TSO training. All of the participants in the study had then successfully completed their classroom training and passed a certification x-ray assessment. After the classroom training, all the participants in the study then proceeded to the airport to complete approximately three weeks of on-the-job

training. At the end of the on-the-job training, potential officers need to pass another x-ray assessment to become certified. The present testing was performed at that time.

These observers could be classified, somewhat oxymoronically, as "novice experts." That is, by virtue of their training, they were qualified to do baggage screening tasks that untrained novices would not be qualified to do. At the same time, these observers were still rookies, and their performance would, no doubt, change with experiment. Biggs, Cain, Clark, Darling, and Mitroff (2013) have shown differences in TSO performance as a function of experience, and our observers did not have that experience. That said, our particular interest is whether prevalence effects can be seen in trained observers performing the task that they were trained to do. These observers fall into that category.

Stimuli

The x-ray training program uses a TIP-ready x-ray (TRX) simulator that can present images and simulate all image manipulation functionality as if operating an actual fielded TSA x-ray system. For our purposes, there are two types of bag images: *clear* images and *threat* images. A clear image is an entire packed bag (hard or soft suitcase, shoulder bag, computer bag, roller bag, garment bag, etc.) containing varying amounts of clutter in the form of different packed objects. To create a threat bag, one manually selects a clear bag and a threat item and then chooses the location and orientation for the threat item inside the bag. The TRX simulator can also perform this task by randomly choosing clear bags and threat items and splicing them together. The program has a reservoir of over 1,000 clear x-ray images of bags and over 450 specific threat images.

Every clear bag and threat bag is given a difficulty rating by the computer simulation. This is based on the clutter inside the bag, defined by the various nonthreat items packed in the bag (e.g., clothing, electronic devices). For threat bags, the simulator also takes into account the orientation and placement of the threat item. The threat bags for the low-prevalence blocks were created by hand, and the simulator chose the clear bags to fill out the blocks. For the high-prevalence blocks, the computer created the threat bags and chose the clear bags to fill in for the rest of the block. All the blocks were set to the same difficulty level as assessed by the program. The experiment was not designed to assess the interaction of prevalence with bag difficulty. As noted, within a block, bags were presented in random order.

Data analysis

Because TSO error rates are considered to be security-sensitive information, we are presenting normalized versions of the data. These are transformations of the raw error data and signal-detection measures designed to preserve the relative differences between blocks and the statistical measures of those differences without presenting the actual error rates. Thus, Figure 1a shows the probability of correct detection of the target (hit rate) as a function of block, normalized by subtracting and dividing by the grand mean. Statistical analyses were conducted on the raw data, and analyzing the normalized data leads to identical results.

Accuracy data (hits and false alarms) were arc-sin transformed before analysis. Accuracy data were also transformed into the signal-detection theory parameters d' (sensitivity) and c (criterion). Because these parameters cannot be computed if either hit rate or false-alarm rate is 0% or 100%, those cells were corrected by adding half a correct response or half an incorrect response as needed (Macmillan & Creelman, 2005).

We report generalized eta-squared (ges) (Bakeman, 2005) as a measure of effect size. Analyses were conducted in R 2.15.0 (R Development Core Team, 2011), using the "ez" package (Lawrence, 2010).

Experiment 1: Results

Figure 1a shows the (normalized) probability of correct detection of the target (hit rate) as a function of the block. It is obvious that the normalized hit rates (and thus the true hit rates) were higher in the high-prevalence block. Observers missed more targets at low prevalence than at high prevalence. This is borne out statistically. The hit rate was higher in the high-prevalence fourth block than in the first three low-prevalence blocks, $F(1, 78) = 98.1$, $p < 0.0001$, ges = 0.30. As noted above, laboratory studies have shown that performing a high-prevalence task with feedback produced effects that persisted into a subsequent low-prevalence, no-feedback block (Wolfe et al., 2007). There is evidence for this effect here. The hit rate was higher in Block 5 than in the first three low-prevalence blocks, $F(1, 78) = 24.8$, $p < 0.0001$, ges = 0.12, and the hit rate in low-prevalence Block 5 was not significantly lower than the hit rate in the high-prevalence block, $F(1, 78) = 1.0$, $p = 0.31$, ges = 0.01. Hit rate appeared to decline over the first three low-prevalence blocks as reflected in a main effect of block over those three blocks, $F(2, 156) = 4.3$, $p = 0.016$, ges = 0.02.

These data clearly show that misses are elevated at low prevalence. This is not a subtle effect. The average

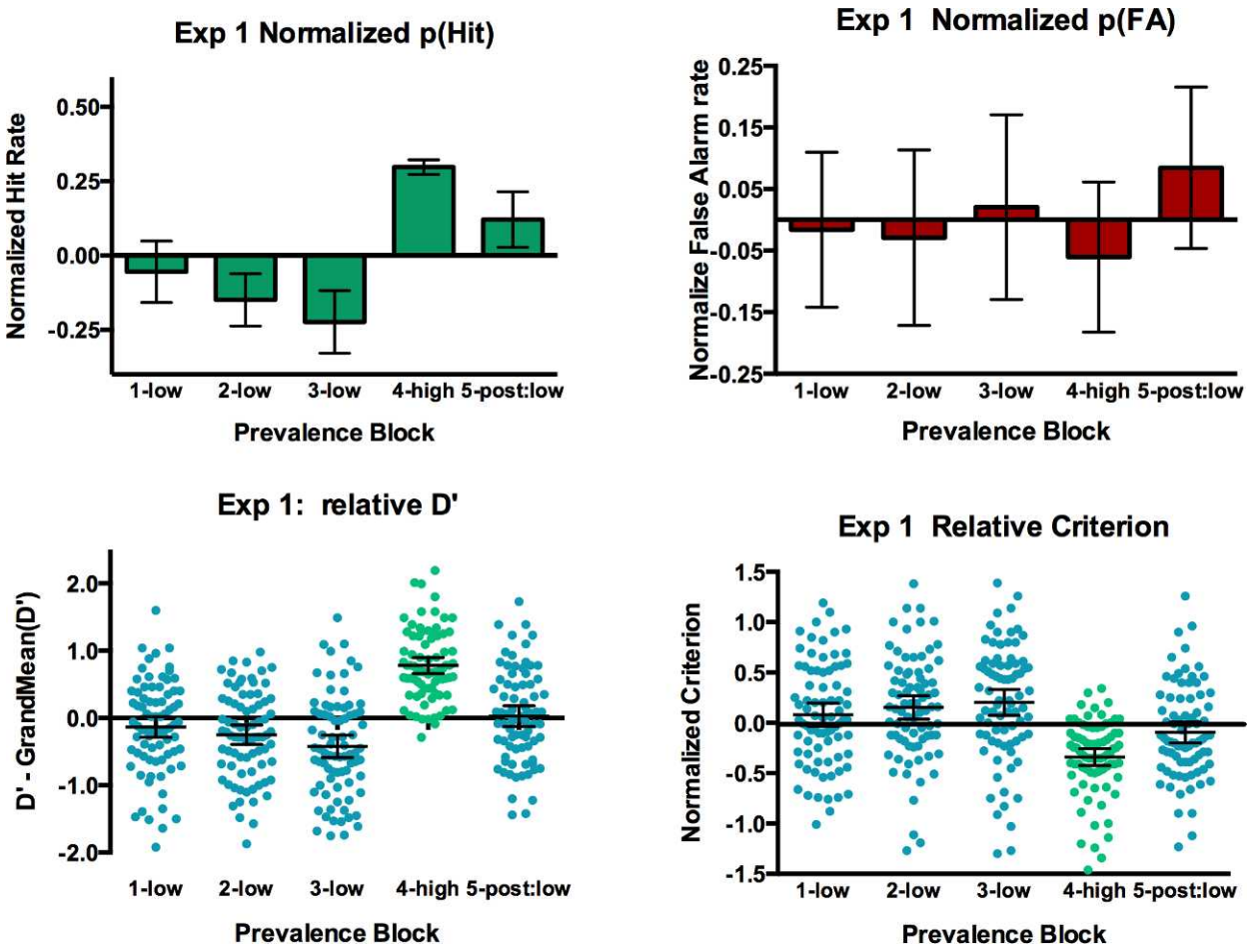

Figure 1. Normalized hit rate, false alarm, D', and criterion as functions of prevalence block in Experiment 1. Error bars show 95% confidence limits around the mean. Scatter plots show one data point per observer.

hit rate for the first three low-prevalence blocks was lower than the hit rate at high prevalence for 74 of 79 observers.

Figure 1b shows the normalized false-alarm data. The false-alarm rate was not higher in the high-prevalence block than in the first three low-prevalence blocks, $F(1, 78) = 1.5$, $p = 0.22$, ges = 0.00. The false-alarm rate in the final low-prevalence block was greater than the false-alarm rate in both the high-prevalence block, $F(1, 78) = 10.1$, $p = 0.002$, ges = 0.12, and the first three low-prevalence blocks, $F(1, 78) = 5.7$, $p = 0.019$, ges = 0.01.

Figure 1c shows the (normalized) signal-detection parameter, d', a measure of the ability to discriminate target-present from target-absent trials. In laboratory work with baggage stimuli, d' tends not to change with prevalence (Gur et al., 2003; Wolfe et al., 2007). In the case of these TSO observers, however, d' did change with prevalence. The d' at high prevalence was significantly greater than in the three initial low-prevalence blocks, $F(1, 78) = 304.6$, $p < 0.0001$, ges = 0.51. As in the hit data, there was some preservation of this improvement in the final low-prevalence Block 5. The d' was higher in that block than in the first three low-prevalence blocks, $F(1, 78) = 13.5$, $p = 0.0004$, ges = 0.06. The d' in Block 5 was lower than in the high-prevalence Block 4, $F(1, 78) = 70.2$, $p < 0.0001$, ges = 0.28. Finally, like the hit rate, d' fell over the first three low-prevalence blocks, $F(1, 78) = 4.7$, $p = 0.0107$, ges = 0.03.

Finally, Figure 1d shows the (normalized) signal-detection criterion parameter c, reflecting the level of bias toward responding target present or target absent. A lower criterion is considered more "liberal" and indicates a greater likelihood of a target-present response, and a higher criterion is considered more "conservative" and reflects a greater likelihood of a target-absent response.

Consistent with what has been found in laboratory studies of prevalence effects (Wolfe & Van Wert, 2010), criterion was significantly lower at high prevalence than during the first three low-prevalence blocks, $F(1, 78) = 191.4$, $p < 0.0001$, ges = 0.25. This shift was partially preserved in the last low-prevalence block, in which criterion remained more liberal than during the first three low-prevalence blocks, $F(1, 78) = 25.0$, $p < 0.0001$, ges = 0.06, although it was

significantly more conservative than criterion in the high-prevalence block, $F(1, 78) = 33.5$, $p < 0.0001$, ges $= 0.08$. High prevalence makes observers more likely to say "yes," and that bias persists into a subsequent low-prevalence regime.

Experiment 2: Replication

The basic result from these data is quite clear. Miss-error rates are markedly higher when newly trained TSOs are tested at low prevalence without feedback than when they are tested at high prevalence with feedback. Anecdotally, it is reported that TSO performance is lower on the job than it is at the end of training. It is possible that this reflects the transition from high prevalence in training to low prevalence on the job. However, there is a potential problem with these data. As noted in the Method section, threat items in the low-prevalence conditions were chosen by hand, and threat items in the high-prevalence condition were chosen by the TRX software. Although the displays had equivalent difficulty ratings from the software, it is possible that the low-prevalence targets were systematically more difficult to find, independent of prevalence. Accordingly, the study was rerun with both the low- and high-prevalence targets chosen by the TRX software. Forty-six newly trained TSOs were tested. Methods were otherwise identical to the preceding experiment.

Results are shown in Figure 2. It is clear that the basic pattern of results was replicated although the effects were somewhat weaker statistically. This appears to be due, in part, to a restriction of range in the accuracy data. The basic pattern is seen most clearly in d' and criterion analyses. The d' was higher at high prevalence than in the first three low-prevalence blocks, $F(1, 45) = 32.2$, $p < 0.0001$, ges $= 0.25$, and criterion was more liberal, $F(1, 45) = 14.5$, $p = 0.0004$, ges $= 0.06$, even though the hit and false-alarm rates did not differ significantly, both $F(1, 45) < 2.5$, $p > 0.1$.

Evidence for the beneficial effects of high prevalence on subsequent low-prevalence blocks is weaker than in Experiment 1. The hit rate was greater in Block 5 than it was in Blocks 1 through 3, $F(1, 45) = 7.3$, $p = 0.0095$, ges $= 0.06$. However, the false-alarm rate was also greater, $F(1, 45) = 4.4$, $p = 0.0421$, ges $= 0.02$, leading to no significant change in d', $F(1, 45) = 0.8$, $p = 0.37$.

Discussion

This paper addresses two questions: First, does the prevalence effect seen in laboratory settings occur when trained observers (albeit newly trained) perform search for rare targets in their area of expertise? Second, if so, is there anything we can do about it?

The answer to the first question is "yes" at least for newly trained TSOs. Miss-error rates were higher at low prevalence than at high prevalence. There are two caveats here. First, TSOs were not at the airport checkpoint but were carrying out a simulated test. Second, the high prevalence was confounded with feedback. This confound was deliberate: In our previous work, we found that a "burst" of high-prevalence trials with no feedback was insufficient to alter low-prevalence behavior (Wolfe et al., 2007), presumably because observers could not necessarily perceive the prevalence of targets without feedback.

As to the first caveat, we suspect that we would get the same results if we were able to carry out the study at the checkpoint (although high prevalence would never be realistic). Keep in mind that the TSOs were highly motivated because the experiment was part of their performance evaluation. Furthermore, we have converging data from another expert domain. Breast cancer screening is also a difficult, low-prevalence search task. In an experiment in which we slipped cases into the regular workflow of a breast cancer screening practice, we found that miss-error rates at low prevalence were over twice as high as miss-error rates for the same cases read at 50% prevalence in a nonclinical setting (Evans, Birdwell, & Wolfe, 2013). Similar effects were found with expert readers of Pap tests for cervical cancer (Evans, Tambouret, Wilbur, Evered, & Wolfe, 2011). We see no reason to expect that TSOs at work would be any more immune to the pressures of low prevalence than radiologists. This research also speaks to the second caveat: It is possible that it was the feedback rather than the prevalence manipulation that changed the TSOs' behavior in the fourth block. However, this would be counter to the behavior of both naive observers in the laboratory and highly trained medical observers in the clinic. Feedback tends to magnify the effects of prevalence by giving the searcher more effective information about the prevalence. It does not appear to change criterion by itself.

Given a positive answer to the question of whether prevalence alters the behavior of TSOs, the second question is whether prevalence effects can be counteracted in a manner that might be practical in the field. In laboratory situations, we have found the effect to be quite robust and resistant to a variety of "cures" (Wolfe et al., 2007). One obvious cure would be to raise the prevalence rate to 50%. This could be done by adding TIP targets to half of the bags. However, it is not clear that this is practical. The time required to clear a bag that includes a TIP is greater than the time required to clear an otherwise unproblematic bag. The bag needs to be rescanned without the TIP to make sure that the TIP

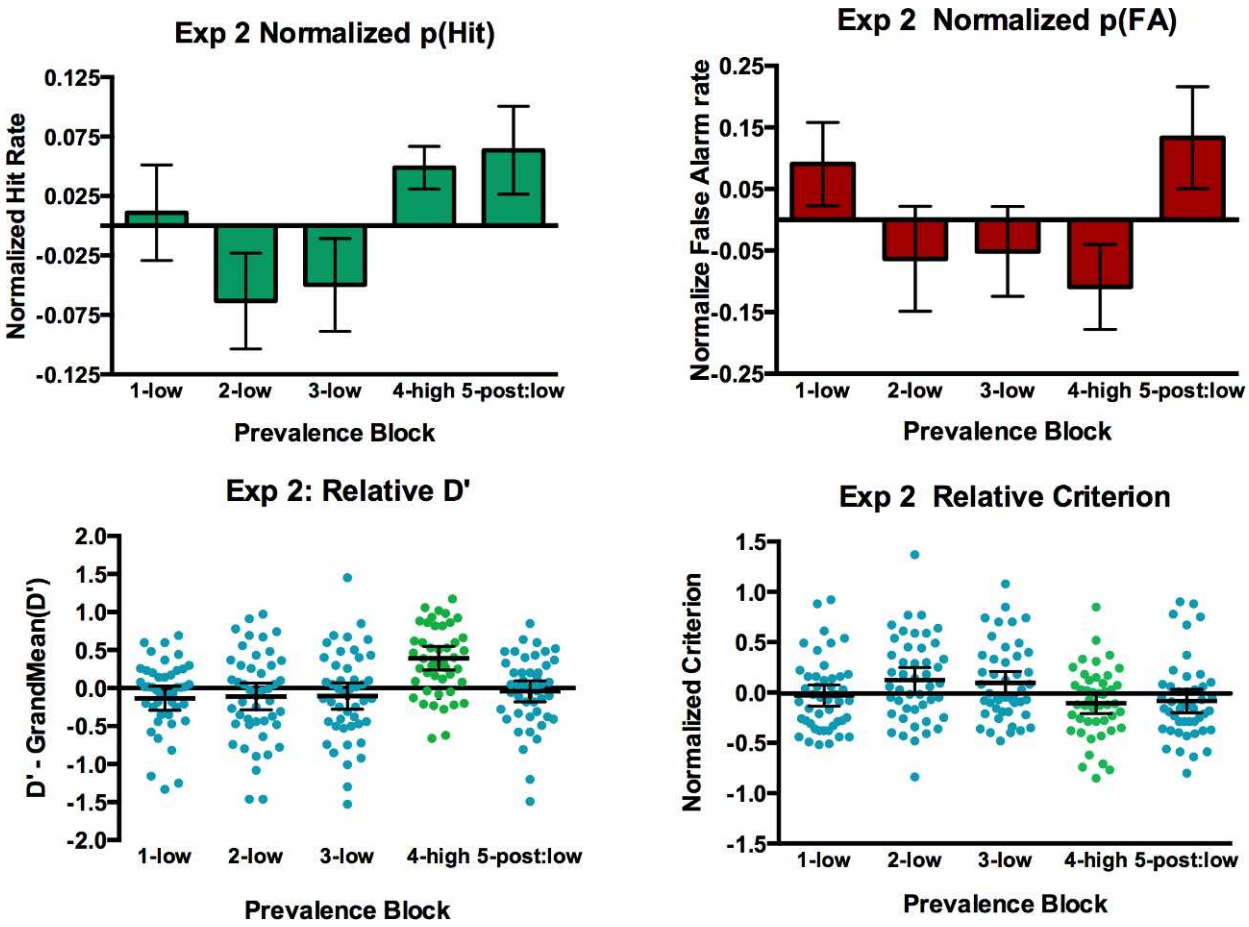

Figure 2. Hit and false-alarm rates as well as signal-detection parameters (D' and c) for the replication of the experiment. Results are normalized as in Panels 1 through 4.

did not obscure a real target. At 50% TIP rates, the lines at the checkpoint would probably grow unacceptably long. It would also be important to determine if the presence of a TIP reduces the probability of detecting a true target (even on the second viewing of the bag). This would be a version of what is known as "satisfaction of search" in the radiology literature (Berbaum et al., 1990; Cain, Dunsmoor, LaBar, & Mitroff, 2011; Nodine, Krupinski, Kundel, Toto, & Herman, 1992; Samuel, Kundel, Nodine, & Toto, 1995).

Under laboratory conditions, we have found that brief epochs of high-prevalence search with feedback reduce miss errors in a subsequent epoch of low-prevalence search (Wolfe et al., 2007). The present study was designed to assess whether this might work in the field. There is an indication that it might. The results are encouraging, if not unequivocal. In Experiment 1, miss-error rates were lower in the final block and d' was higher. In Experiment 2, the desired reduction in miss errors was found. The d' was not improved because false-alarm errors rose in the final block. This latter pattern, characteristic of a criterion shift, is the pattern that has been seen in laboratory studies. If the primary goal is to reduce miss errors, it appears that exposure to high prevalence, with feedback, has the desired effect. Note that this is unlikely to be a simple practice effect because the effects over the first three blocks were generally in the wrong direction. However, it is possible that the trend would have reversed and that the changes on the final block would have occurred without the high-prevalence block. It would be desirable to perform a control experiment with no high-prevalence block to directly test this hypothesis. Thus, it seems possible that a regimen of a brief high-prevalence block just prior to going to work at the checkpoint might be worth investigating as a countermeasure to the prevalence effect.

Keywords: visual search, prevalence effects, airport security, visual attention, error rates, criterion shift

Acknowledgments

JMW and TSH were supported by funding from the U.S. Department of Homeland Security Science and

Technology Directorate and the Transportation Security Administration and by NIH grant EY017001.

Commercial relationships: none.
Corresponding author: Jeremy M. Wolfe.
Email: wolfe@search.bwh.harvard.edu.
Address: Visual Attention Lab, Brigham and Women's Hospital, Cambridge, MA, USA.

References

Baddeley, A. D., & Colquhoun, W. P. (1969). Signal probability and vigilance: A reappraisal of the 'signal-rate' effect. *British Journal of Psychology*, *60*(2), 169–178.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: global and local clutter. *Journal of Experimental Psychology: Applied*, *16*(3), 238–250, doi: 2010-19027-002 [pii]10.1037/a0019633.

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J... Montgomery, W. J. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology*, *25*(2), 133–140.

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., & Mitroff, S. R. (2013). Assessing visual search performance differences between Transportation Security Administration Officers and non-professional visual searchers. *Visual Cognition*, *21*(3), 330–352.

Cain, M. S., Dunsmoor, J. E., LaBar, K. S., & Mitroff, S. R. (2011). Anticipatory Anxiety Hinders Detection of a Second Target in Dual-Target Search. *Psychological Science*, *22*(7), 866–871, doi: 10.1177/0956797611412393.

Colquhoun, W. P. (1961). The effect of 'unwanted' signals on performance in a vigilance task. *Ergonomics*, *4*, 41–51.

Davies, D. R., & Parasuraman, R. (1982). The psychology of vigilance. London: Academic Press.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*, 433–458.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980.

Egeth, H., Jonides, J., & Wall, S. (1972). Parallel processing of multielement displays. *Cognitive Psychology*, *3*, 674–698.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don't find it often, you often don't find it: Why some cancers are missed in breast cancer screening. *PLoS ONE*, *8*(5), e64366.

Evans, K. K., Tambouret, R., Wilbur, D. C., Evered, A., & Wolfe, J. M. (2011). Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. *Archives of Pathology & Laboratory Medicine*, *135*(12), 1557–1560, doi: 10.5858/arpa.2010-0739-OA.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947.

Green, D. M., & Swets, J. A. (1967). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.

Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., ... Warfel, T. E. (2003). Prevalence effect in a laboratory environment. *Radiology*, *228*(1), 10–14, doi: 10.1148/radiol.2281020709228/1/10 [pii].

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

Ishibashi, K., Kita, S., & Wolfe, J. M. (2012). The effects of local prevalence and explicit expectations on search termination times. *Attention, Perception, & Psychophysics*, *74*(1), 115–123.

Kundel, H. L. (2004). Reader error, object recognition, and visual search. Paper presented at the Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment, San Diego, CA, USA.

Lau, J. S., & Huang, L. (2010). The prevalence effect is determined by past experience, not future prospects. *Vision Research*, *50*(15), 1469–1474, doi:10.1016/j.visres.2010.04.020.

Lawrence, M. A. (2010). ez: Easy analysis and visualization of factorial experiments (Version R package version 2.1-0). Retrieved from http://CRAN.R-project.org/package=ez.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78*(3), 567–595.

Matthews, G. (1996). Signal probability effects on

high-workload vigilance tasks. *Psychonomic Bulletin and Review*, *3*(3), 339–343.

Nagy, A. L., & Sanchez, R. R. (1990). Critical color differences determined with a visual search task. *Journal of the Optical Society of America - A*, *7*(7), 1209–1217.

Navalpakkam, V., Koch, C., & Perona, P. (2009). Homo economicus in visual search. *Journal of Vision*, *9*(1):13, 1–16, http://www.journalofvision.org/content/9/1/31, doi:10.1167/9.1.31. [PubMed] [Article]

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621.

Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology*, *76*, 376–385.

Nodine, C. F., Krupinski, E. A., Kundel, H. L., Toto, L., & Herman, G. T. (1992). Satisfaction of search (SOS). *Investigative Radiology*, *27*(7), 571–573.

Nothdurft, H.-C. (2000). Salience from feature contrast: additivity across dimensions. *Vision Research*, *40*, 1183–1201.

Nuechterlein, K. H., Parasuraman, R., & Jiang, Q. (1983). Visual sustained attention: image degradation produces rapid sensitivity decrement over time. *Science*, *220*(4594), 327–329.

Parasuraman, R., & Davies, D. R. (1976). Decision theory analysis of response latencies in vigilance *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 578–590.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2):17, 1–22, http://www.journalofvision.org/content/7/2/17, doi:10.1167/7.2.17. [PubMed] [Article]

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*(3), 895–902.

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-Analysis of the Sensitivity Decrement in Vigilance. *Psychological Bulletin*, *117*(2), 230–249.

Van Wert, M. J., Wolfe, J. M., & Horowitz , T. S. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception & Psychophysics*, *71*(3), 541–553, http://www.pubmedcentral.gov/articlerender.fcgi?artid=2701252.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*(1):8, 81–92, http://www.journalofvision.org/content/5/1/8, doi:10.1167/5.1.8. [PubMed] [Article]

Williges, R. C. (1973). Manipulating the response criterion in visual monitoring. *Human Factors*, *15*, 179–185.

Wolfe, J., Horowitz, T., Kenner, N. M., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, *44*(12), 1411–1426.

Wolfe, J. M. (2010). Visual search. *Current Biology*, *20*(8), R346–R349, doi:10.1016/j.cub.2010.02.016.

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

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

Wolfe, J. M., & Reynolds, J. H. (2008). Visual search. In A. I. Basbaum, A. Kaneko, G. M. Shepherd, & G. Westheimer (Eds.), *The senses: A comprehensive reference* (Vol. 2, pp. 275–280). San Diego: Academic Press.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, *20*(2), 121–124, doi:10.1016/j.cub.2009.11.066.