Research Article

Visual Skills in Airport-Security Screening

Jason S. McCarley,[1] Arthur F. Kramer,[1] Christopher D. Wickens,[1,2] Eric D. Vidoni,[1] and Walter R. Boot[1]

[1]*Beckman Institute and* [2]*Aviation Research Laboratory, University of Illinois at Urbana-Champaign*

ABSTRACT—*An experiment examined visual performance in a simulated luggage-screening task. Observers participated in five sessions of a task requiring them to search for knives hidden in x-ray images of cluttered bags. Sensitivity and response times improved reliably as a result of practice. Eye movement data revealed that sensitivity increases were produced entirely by changes in observers' ability to recognize target objects, and not by changes in the effectiveness of visual scanning. Moreover, recognition skills were in part stimulus-specific, such that performance was degraded by the introduction of unfamiliar target objects. Implications for screener training are discussed.*

Recent concern over aviation security has focused interest on the role of airport-security screeners in keeping weapons and other potential threats off aircraft. The job of these screeners is to examine x-ray images of carry-on luggage to detect the presence of suspicious or threatening objects. Unfortunately, screeners often work under conditions characterized by high levels of noise and time stress. More problematically, their task—detecting weak and infrequent visual signals among high levels of background clutter—is inherently difficult, straining the perceptual and cognitive capacities of the typical human observer (Harris, 2002). Efforts to optimize screener training and redesign the screeners' task, bringing it more in line with the limits of human perception and cognition, are therefore a priority in aviation security.

To date, relatively little research appears to have focused directly on human performance in the task of aviation security screening (for an exception, see Gale, Mugglestone, Purdy, & McClumpha, 2000). Knowledge of visual search in other domains, however, is likely to provide a strong foundation for understanding the security screeners' task. Study of applied visual search in medical image inspection and other real-world domains has led to a multistage model of search performance (Kundel, Nodine, & Carmody, 1978; Nodine, Krupinski, & Kundel, 1993). The first stage consists of a rapid, global assessment of the stimulus image, during which general spatial layout is determined, familiar structures or features are identified, and potential target regions are noted. This process has been termed *orientation*, and corresponds roughly to what has been referred to in the basic research literature as preattentive (Wolfe, 1994) or distributed attentional (Bravo & Nakayama, 1992) processing. Only targets that are highly salient or poorly camouflaged may be acquired during orientation. Targets that are less conspicuous or are embedded in visual noise generally demand further and more effortful processing, with the observer scanning the image to fixate and inspect potential target regions. Moreover, successful acquisition of a camouflaged target requires that observers not only inspect the appropriate region of the display, but also recognize the target once they have looked at it, parsing it from the embedding background and matching it to the target template. When the target is of low contrast or is otherwise near sensory threshold, this aspect of performance may be limited by low-level visual sensitivity. When the target is well above sensory threshold, recognition is more likely to be a test of perceptual organization, that is, of the ability to group image regions belonging to the target object into an accurate perceptual representation. Failure to find a target in a cluttered or naturalistic display can result from a lapse of either scanning or recognition; observers searching cluttered images sometimes fail to fixate the region of the image containing the target, but can also fail to acquire a target even after they have gazed directly at it.

In medical image reading, learning can affect both scanning and recognition. Nodine, Mello-Thoms, Kundel, and Weinstein (2002), for example, found that experienced radiologists were more likely than interns to fixate abnormal regions of a mammogram. Such increased effectiveness of visual scanning could reflect strategic expertise in planning scan paths (Kundel & La Follette, 1972) or perceptual expertise in noticing and guiding the eyes toward peripherally viewed targets (Kundel, Nodine, & Toto, 1991).

Experienced medical image readers are also more likely than others to recognize abnormalities, a skill that appears to reflect both increased low-level visual sensitivity (Sowden, Davies, & Roling, 2000) and increased knowledge about the likelihood of particular abnormalities in light of patients' clinical histories (Norman, Brooks, Coblentz, & Babcook, 1992). It is not obvious, though, what skills might contribute to expertise in security x-ray inspection. Although between-patient consistencies in anatomical structure and clinical characteristics allow experienced medical image readers to direct scanning toward the regions of a stimulus most likely to contain an abnormal feature, security images in which distractor objects are chosen randomly and target and distractor placement is unconstrained provide little if any spatial or conceptual regularity to guide search. Similarly, development of target recognition skills in security

Address correspondence to Jason McCarley, Department of Psychology, Mississippi State University, P.O. Box 6161, Mississippi State, MS 39762; e-mail: jmccarley@psychology.msstate.edu.

screening may be hindered by the fact that the set of potential target objects—that is, all potential weapons or threatening objects—is infinitely heterogeneous in appearance. Evidence indicates that the ability to perceptually organize and identify shape representations from degraded imagery is based largely on observers' ability to retrieve stored mental representations of particular familiar objects (Furmanski & Engel, 2000; Kundel & Nodine, 1983; Moore & Cavanagh, 1998), and similarly, that the assignment of figure-ground relations in ambiguous imagery is guided by stimulus recognizability (Peterson & Gibson, 1994). Therefore, it is possible that improvements in observers' ability to recognize targets in security imagery will transfer poorly when novel targets are introduced.

Thus, we had two aims in the present research. The first was to determine whether practice performing a simulated security x-ray screening task improves search, recognition, or both. The second was to assess the degree to which the search and recognition skills that a subject might develop with practice are specific to the target stimuli employed during training. The finding that only one set of skills or the other is amenable to practice, or that either set of skills is stimulus-specific, would entail potentially important implications for the design of training methods and materials.

METHOD

Observers
Observers were 16 young adults (mean age = 21 years, 12 female). All had normal or corrected-to-normal visual acuity and normal color vision.

Apparatus and Stimuli
Stimuli were presented on a 19-in. monitor with a resolution of 800 × 600 pixels and an 85-Hz refresh rate. Eye movements were recorded with an Eyelink eye tracker (SR Research, Ltd., Mississauga, Ontario, Canada) with temporal resolution of 250 Hz and spatial resolution of $0.2°$. An eye movement was classified as a saccade either when its distance exceeded $0.2°$ and its velocity reached $30°/s$ or when its distance exceeded $0.2°$ and its acceleration reached $9500°/s^2$. Observers viewed displays from a distance of 91 cm, with viewing distance controlled by a chin rest.

Stimuli were produced from chromatic x-ray images provided by the Federal Aviation Administration. Images of 89 bags served as backgrounds. All bags were moderately to densely cluttered with a variety of everyday objects (e.g., clothes, hair dryers, pill bottles), and could be presented in any of four orientations. Eight knives served as target objects. These items were chosen from a larger set on the basis of a pilot study in which observers rated the visual similarity between all pair-wise combinations of 25 knives. Two sets of 4 knives each, referred to hereafter as Sets 1 and 2, were chosen from the full set of 25 such that the rated similarity of the items within each set was higher than the rated similarity of items between sets. All knives were imaged with their flat side perpendicular to the line of sight. As measured by the maximum distance from edge to edge horizontally and vertically, bags ranged in size from $10.34° × 8.34°$ to $15.34° × 13.30°$; knives ranged from approximately $2.66° × 0.60°$ to $6.14° × 1.21°$.

Target-present stimuli (see Fig. 1 for an example) were generated by digitally inserting images of knives into images of bags at random locations and at randomly chosen orientations of $0°$, $45°$, $90°$, $135°$,

Fig. 1. Example of a target-present stimulus image. The target (the blade and shaft show up in dark blue, and the handle in orange) is a little above the center of the image, to the left of the toy airplane.

$180°$, $225°$, $270°$, or $315°$ in the picture plane. No more than one target was inserted into each image. Stimuli for target-absent trials were images of luggage with no weapons inserted. Bags appeared centered against a white background. Four sets of 300 images were generated for each set of target items. A target was present within 20% of the images in each set of 300.

Procedure
The observers' task was to search the stimulus images for the presence of a knife. Across several days, each observer completed five experimental sessions of 60 target-present trials and 240 target-absent trials each. Within a session, all targets were drawn from the same set of four items. During Sessions 1 through 4, all targets for a given observer were drawn from one set of targets or the other. During Session 5, all targets for that observer were drawn from the alternative set. Session 5 thus provided a test of the degree to which skills developed in the earlier sessions transferred to facilitate search for unfamiliar targets. Half of all observers searched for targets from Set 1 during Sessions 1 through 4 and targets from Set 2 during the transfer session. The remaining observers searched for targets from Set 2 during Sessions 1 through 4 and from Set 1 during the transfer session. The order in which sets of images were presented was counterbalanced such that the number of observers who saw a given set in Session 1 was equal to the number of observers who saw that set in Session 4 and Session 5.

Before beginning the first session, observers were given written instructions explaining their task. The instructions asked observers to imagine that they were workers at an airport-security station, and that their job was to search for hidden knives in images of luggage.

Observers were instructed to stop a bag if they believed that it contained a target, and to pass the bag if they believed it contained no target. "Stop bag" (i.e., target present) responses were made by pressing the "F" key on the experimental computer's keyboard. "Pass bag" (i.e., target absent) responses were made by pressing the "J" key. Observers were instructed that they should emphasize accuracy in their responses, but without using any more time than necessary to produce each response. Before beginning the transfer session, observers were given instructions explaining that the target knives they would now encounter would look different from those they had seen earlier, but that their task would otherwise be the same as it had been.

Observers initiated each trial by gazing at a central fixation mark and pressing the space bar on the experimental computer's keyboard. Thereafter, a stimulus image appeared and remained visible until the observer's response. Text feedback was given after each response.

RESULTS

We selected dependent variables to assess various aspects of performance. To measure general task performance, we calculated a signal detection measure of sensitivity, A_z, along with mean reaction times (RTs) for accurate target-present and target-absent responses. To measure visual scanning performance, we calculated the probability that the observer fixated the target item, when it was present, at least once within the course of trial, along with the number of saccades executed prior to the first fixation on the target object for those trials ending with successful target acquisition. To measure target recognition, we calculated hit rate (i.e., the probability of target acquisition) for trials on which the observer fixated the target, along with the mean number of oculomotor dwells on the target preceding the successful response, and hit rate for trials on which the target was never fixated. Finally, we calculated false alarm rate as a control variable for examining target recognition data. A fixation was classified as being on the target if it fell inside or within $2°$ of visual angle of the smallest rectangle that could be drawn around the target object.

Our presentation is organized as follows. First, we examine changes between the first and last sessions of training (Session 1 vs. 4). This analysis provides insight as to how performance was affected by practice, holding target set constant within observer. Second, we compare performance during the last session of training with performance during the transfer session (Session 4 vs. 5). This analysis indicates whether the skills developed by practice were in part specific to the target set employed during practice. Finally, we compare performance during the first session of practice with performance during the transfer session (Session 1 vs. 5). This analysis provides evidence as to whether any transfer of skill was possible when practiced observers were required to search for unfamiliar target shapes. Means and standard errors for all variables are presented in Table 1.

Changes Across Blocks of Practice

All dependent variables were submitted to two-way mixed analyses of variance (ANOVAs) with session (1 vs. 4) as a within-subjects factor and stimulus set (Set 1 vs. Set 2) as a between-subjects variable. Given that there were no theoretical reasons to expect performance differences between the two target sets, stimulus set was included in these analyses only as a control variable to reduce error variance (Keppel, 1991). Data are collapsed across stimulus set for presentation in Table 1, and we do not discuss effects involving stimulus set.

Analysis indicated that A_z increased reliably between Sessions 1 and 4, $F(1, 14) = 14.255$, $p = .002$, reflecting an increase in overall hit rate from .71 to .80. Concurrently, RTs declined for both target-present responses, $F(1, 14) = 59.504$, $p < .001$, and target-absent responses, $F(1, 14) = 66.396$, $p < .001$. Thus, general task performance improved significantly as a result of practice.

Further analyses pointed to the sources of these improvements. Oculomotor data revealed that the mean number of saccades preceding a target fixation decreased between the first and last sessions of practice, $F(1, 14) = 22.516$, $p < .001$. Interestingly, however, the data gave no evidence of a concomitant change in the probability with which target fixations occurred, $F(1, 14) = 1.633$, $p = 0.221$. In other words, observers were quicker to localize and fixate the target after practice, but were not more likely to do so. Improvements in sensitivity, rather, were produced by changes in observers' ability to recognize targets. Hit rates for trials on which the target was fixated once or more increased reliably between Sessions 1 and 4, $F(1, 14) = 31.544$, $p < .001$, such that the proportion of erroneous target-absent responses resulting from failure to recognize a fixated target declined from 54% to 38%. There was no concomitant change in false alarm

TABLE 1

Mean Performance for Sessions 1, 4, and 5 and Results of Statistical Comparisons Between Sessions

Measure	Session 1 (first practice)	Session 4 (last practice)	Session 5 (transfer)	Session 1 vs. 4	Session 4 vs. 5	Session 1 vs. 5
Sensitivity (A_z)	.94 (.01)	.97 (.01)	.96 (.01)	+	+	+
RT (ms), target-present trials	1,874 (101)	1,147 (51)	1,343 (58)	+	+	+
RT (ms), target-absent trials	3,922 (234)	2,137 (180)	2,508 (208)	+	+	+
Probability of a target fixation	.69 (.03)	.66 (.03)	.69 (.03)	−	−	−
Saccades preceding first target fixation	2.83 (0.15)	2.01 (0.10)	2.31 (0.10)	+	+	+
Hit rate following a fixation on target	.77 (.02)	.89 (.02)	.85 (.01)	+	+	+
Hit rate following no fixation on target	.54 (.03)	.61 (.04)	.56 (.04)	−	−	−
False alarm rate	.05 (.01)	.04 (.01)	.04 (.01)	−	−	−
Dwells on target preceding recognition	1.24 (0.04)	1.08 (0.02)	1.14 (0.02)	+	+	+

Note. Standard errors are in parentheses. In the columns presenting results of statistical comparisons, a plus sign indicates that the difference between sessions was statistically reliable at the level of $p = .05$ or better, and a minus sign indicates that it was not. RT = reaction time.

rate, $F < 1$, indicating that improvements in hit rate for trials involving a target fixation were not produced by changes in response bias. Hit rate for trials on which gaze never fell within 2° of the target did not change across sessions, $F(1, 14) = 2.104$, $p = .169$. After practice, observers were also able to recognize targets more readily, showing a decrease in the mean number of target dwells necessary for successful recognition, $F(1, 14) = 14.985$, $p = .002$.

Stimulus-Specific Benefits of Practice

To determine whether the benefits of practice with a restricted target set were wholly generalized or were in part specific to the familiar target stimuli, we compared performance during the final session of practice with performance in the transfer session. Dependent variables were submitted to two-way mixed ANOVAs with session as a within-subjects factor and stimulus set used during training as a between-subjects variable. There were again no theoretical reasons to expect performance differences between the two target sets. Furthermore, the current design did not allow any method for determining whether practice with either stimulus set produced more robust skill development than practice with the other (e.g., performance might decline more severely during the transfer session for one group than for the other either because the former group developed less generalizable skills during practice or because the stimulus set that group used during practice was easier than the set it used during the transfer session). Therefore, stimulus set was again included only as a control variable.

Sensitivity was reliably lower during the transfer session than during the final session of practice, $F(1, 14) = 6.677$, $p = .022$, the result of a decline in overall hit rate from .80 to .76. Furthermore, RTs for the transfer session were reliably longer than those for Session 4, $F(1, 14) = 28.334$, $p < .001$, for target-present responses and $F(1, 14) = 11.734$, $p = .004$, for target-absent responses. Analysis of oculomotor data found no reliable change between sessions in the probability of target fixation, $F < 1$. However, the introduction of unfamiliar targets did produce an increase in the number of saccades preceding a target fixation, $F(1, 14) = 10.599$, $p = .006$; a decrease in the probability of a hit following a target fixation, $F(1, 14) = 5.574$, $p = .033$; and an increase in the number of dwells on the target preceding recognition, $F(1, 14) = 8.356$, $p = .012$. False alarm rate did not change reliably between Sessions 4 and 5, $F < 1$, nor did hit rate for trials with no target fixation, $F(1, 14) = 1.131$, $p = .305$.

Generalized Benefits of Practice

The analyses discussed thus far indicate that the target recognition skills developed during practice were at least partially stimulus-specific. To determine whether practice with one set of target shapes produced *any* benefits to performance with the alternative set, we compared performance during Session 1 of practice with performance during the transfer session. Dependent variables were again submitted to two-way mixed ANOVAs with session as a within-subjects factor and stimulus set as a between-subjects control variable.

Sensitivity was reliably higher in the transfer session than in Session 1 of practice, $F(1, 14) = 5.406$, $p = .036$, and RTs were reliably shorter, $F(1, 14) = 40.503$, $p < .001$, for target-present responses and $F(1, 14) = 35.893$, $p < .001$, for target-absent responses. The oculomotor results were consistent with the results obtained in the analyses described earlier: Analysis found no reliable difference between Sessions 1 and 5 in the probability of a target fixation, $F < 1$, but did indicate that the number of saccades preceding the first target fixation was lower in the transfer session than in the first session of practice, $F(1, 14) = 14.483$, $p = .002$; the likelihood of a hit following a target fixation was higher, $F(1, 14) = 21.595$, $p < .001$; and the number of dwells on the target preceding successful recognition was lower, $F(1, 14) = 7.983$, $p = .013$. False alarm rate did not differ between Sessions 1 and 5, $F < 1$, nor did the hit rate for trials on which the target was not fixated, $F < 1$.

DISCUSSION

In this study, we examined the development of visual scanning and target detection-recognition skills in a simulated airport-security inspection task. As expected, both sensitivity and RT improved significantly with practice. More important, oculomotor data illuminated the bases of these improvements; after practice, observers were faster to fixate the target region of an image, and were both faster and more likely to recognize the target once they had fixated on or near it.[1] These improvements in performance were in part stimulus-specific, being attenuated by the introduction of unfamiliar target objects.

A surprising aspect of these results is the finding that observers were quicker to fixate the target region of an image as a result of practice, but were not more likely to do so. In other words, scanning became more efficient with practice, but not more effective. A further dissociation was seen in that scanning efficiency was reduced when unfamiliar target shapes were introduced following practice, whereas effectiveness was not. A likely explanation for this result is that decreases in the number of saccades needed to localize the target region were produced by changes in general scanning behavior, rather than by improvements in observers' specific ability to guide the eyes toward a target. For example, familiarity with task and stimuli may have led observers to adopt less meticulous or less redundant scanning strategies, allowing them to sample each image to a criterion level of confidence in a smaller number of fixations. This would have ensured that the target region was fixated sooner in the course of a trial even if scanning skills did not improve. Consistent with this speculation is the finding that practice reduced search time not just for hits but for correct target-absent responses as well (and indeed, though data were not presented here, for misses and false alarms), indicating that changes in the speed of target acquisition were not in themselves entirely responsible for decreases in RT. The results suggest, in any case, that practice had little effect on observers' ability to locate targets through "visual foraging" (Klein & MacInnes, 1999). This may reflect the fact, noted earlier, that imaged luggage provides little if any trial-to-trial regularity to guide scanning.

---

[1]An alternative interpretation is that recognition of fixated targets did not change, but that recognition of targets outside of fixation improved. Increases in hit rate for trials involving a target fixation might then have resulted from a tendency for observers to saccade toward peripherally detected targets for confirmatory inspection. This hypothesis, however, suggests that training should have made observers more likely to fixate the target. This did not happen. Thus, although observers were clearly capable of peripheral target detection on some trials (as can be seen by comparing false alarm rates and hit rates for trials with no target fixation), it is improbable that an improvement in peripheral detection produced the observed changes in hit rates for trials on which the target in fact was fixated.

In contrast, the data indicate clearly that practice did sharpen observers' ability to recognize targets. Interestingly, this effect was significant only for trials involving a target fixation, suggesting that improvements occurred primarily for cases in which the target was well camouflaged and recognition therefore demanded foveation. It seems likely that the effect of practice was to hone observers' ability to organize the fragments of a camouflaged object into a veridical representation of a target shape, disambiguated from background clutter. The finding that these skills were in part specific to familiar target objects is consistent with evidence that stimulus familiarity guides the perceptual organization and recognition of degraded or ambiguous displays (Furmanski & Engel, 2000; Kundel & Nodine, 1983; Moore & Cavanagh, 1998; Peterson & Gibson, 1994). The stimulus-specific benefits of practice in the current experiment were smaller than the stimulus-invariant benefits; hit rate for fixated targets declined by only about 4% following the introduction of unfamiliar targets, and remained about 8% higher than during the initial block of practice. This suggests that the recognition skills developed by practice were largely generalized. The present experiment, however, is likely to have optimized transfer of skill from familiar to unfamiliar targets. All targets used were drawn from the same class of objects (knives), and thus did not differ dramatically in appearance. Additionally, observers were warned prior to the introduction of novel target shapes, and may therefore have adopted a performance strategy that facilitated recognition of the unfamiliar items. Stimulus unfamiliarity could thus be more detrimental to performance in real-world circumstances than it was here.

These results carry at least two implications for the training of security screeners. The first concerns the goals and design of training. The current data gave little evidence that practice naturally improves the effectiveness of screeners' visual scanning. Furthermore, past research has found that it may be difficult to inculcate artificial scanning strategies in naturalistic tasks (e.g., Carmody, Kundel, & Toto, 1984). Together, these findings imply that training should not be designed to modify the screeners' scanning behavior, but should focus instead on developing their ability to perceptually organize and recognize objects in security imagery. The second implication of the present results concerns the selection of training materials. The recognition skills developed by practice in the current task were to a degree stimulus-specific. Thus, the data suggest that the target materials employed during training should be maximally heterogeneous so as to ensure skill generalization.

REFERENCES

Bravo, M.J., & Nakayama, K. (1992). The role of attention in different visual-search tasks. *Perception & Psychophysics*, *51*, 465–472.

Carmody, D.P., Kundel, H.L., & Toto, L.C. (1984). Comparison scans while reading chest images: Taught, but not practiced. *Investigative Radiology*, *19*, 462–466.

Furmanski, C.S., & Engel, S.A. (2000). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research*, *40*, 473–484.

Gale, A.G., Mugglestone, M.D., Purdy, K.J., & McClumpha, A. (2000). Is airport baggage inspection just another medical image? *Proceedings of the SPIE*, *3981*, 184–192.

Harris, D.H. (2002). How to really improve airport security. *Ergonomics in Design*, *10*, 17–22.

Keppel, G. (1991). *Design and analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Klein, R.M., & MacInnes, W.J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, *10*, 346–352.

Kundel, H.L., & La Follette, P.S. (1972). Visual search patterns and experience with radiological images. *Radiology*, *103*, 523–528.

Kundel, H.L., & Nodine, C.F. (1983). A visual concept shapes image perception. *Radiology*, *146*, 363–368.

Kundel, H.L., Nodine, C.F., & Carmody, D. (1978). Visual scanning, pattern recognition, and decision-making in pulmonary nodule detection. *Investigative Radiology*, *13*, 175–181.

Kundel, H.L., Nodine, C.F., & Toto, L. (1991). Searching for lung nodules: The guidance of visual scanning. *Investigative Radiology*, *26*, 777–781.

Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition*, *67*, 45–71.

Nodine, C.F., Krupinski, E.A., & Kundel, H.L. (1993). Visual processing and decision making in search and recognition of targets embedded in pictorial scenes. In D. Brogan & A.G. Gale (Eds.), *Visual search* (pp. 239–249). Philadelphia: Taylor & Francis.

Nodine, C.F., Mello-Thoms, C., Kundel, H.L., & Weinstein, S.P. (2002). Time course of perception and decision making during mammographic interpretation. *American Journal of Roentgenology*, *179*, 917–923.

Norman, G.R., Brooks, L.R., Coblentz, C.L., & Babcook, C.J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Memory & Cognition*, *20*, 344–355.

Peterson, M.A., & Gibson, B.S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science*, *5*, 253–259.

Sowden, P.T., Davies, I.R.L., & Roling, R. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 379–390.

Wolfe, J.M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238.