Determining the Optimal Time on X-Ray Analysis for Transportation Security Officers

Ann Speed(✉), Austin Silva, Derek Trumbo, David Stracuzzi, Christina Warrender, Michael Trumbo, and Kristin Divis

Sandia National Laboratories, Albuquerque, NM 87185, USA
aespeed@sandia.gov

Abstract. The Transportation Security Administration has a large workforce of Transportation Security Officers, most of whom perform interrogation of x-ray images at the passenger checkpoint. To date, TSOs on the x-ray have been limited to a 30-min session at a time, however, it is unclear where this limit originated. The current paper outlines methods for empirically determining if that 30-min duty cycle is optimal and if there are differences between individual TSOs. This work can inform scheduling TSOs at the checkpoint and can also inform whether TSOs should continue to be cross-trained (i.e., performing all 6 checkpoint duties) or whether specialization makes more sense.

Keywords: Visual search · Fatigue · Vigilance · Transportation security agency

1 Introduction

One of the key tasks that Transportation Security Officers (TSOs) must perform is interrogation of passenger bags at the airport security checkpoint using an x-ray system. This is a difficult task for a number of reasons, and requires the TSO to sustain focused attention for a period of time. Currently, TSOs perform x-ray duties for 30 min at a time, with approximately an hour break in between each x-ray session.

The Transportation Security Administration (TSA) is currently examining how they staff security checkpoints and, as such, is curious about how long TSOs can actually spend performing x-ray imagery analysis and whether this differs from TSO to TSO. Such information can impact scheduling and can inform any questions around specializing TSOs for particular duties.

This paper details experimental methods aimed at answering these questions. First is a review of relevant peer-reviewed literature on vigilance, visual search, and inspection. Then, the experimental design and methods for the current study are presented.

1.1 Vigilance

Research on the ability for individuals to sustain attention on a task for an extended period of time started with Mackworth's [1] seminal study of Royal Air Force radar

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 825–834, 2015.
DOI: 10.1007/978-3-319-20816-9_79

operators using his "clock" paradigm in which RAF operators monitored the movement of the hands of a clock-type apparatus for low probability jumps as the hands moved around the dial. He found that after about 10–15 min, subjects were increasingly likely to miss such jumps during the first 30 min and that performance bottomed out after that. He also found that short breaks restored performance, that response to an event inhibited response to a subsequent event proximal in time, that an additional monitoring task (e.g., monitoring for a phone call) also further inhibited performance, and that a small dose of amphetamine helped performance.

Similar studies have found similar results – for example Molloy and Parasuraman [2] found that performance degraded during the course of a 30-min vigil of a monitoring task, especially when that task involved attending to multiple complex tasks. However, even a single, complex task, which had the highest performance, still demonstrated some decrement in detection accuracy.

When the number of events to be detected is of sufficient number, the results are often reported in terms of signal detection theory metrics such as d' and a measure of response bias (c or A'). In these cases, a dissociation is often seen between changes in d' and changes in response bias over time, with d' generally declining over time (the *sensitivity decrement*) and positive response bias (i.e., the tendency to say "no target") increasing over time *[the bias increment; 3]*.

Such a sensitivity decrement has been shown repeatedly in numerous paradigms over the course of several decades. See, Howe [4] performed a meta-analysis of 42 vigilance studies and found that the sensitivity decrement is substantially impacted by both the type of discrimination being made (a *successive* one that requires the subject to recall the target from memory or one that simply requires the subject to compare two *simultaneously* occurring stimuli) as well as the rate of the event (higher rates of events lead to greater decrements). They also found that distinguishing between primarily sensory stimuli (e.g., lines of different lengths) and primarily cognitive stimuli (e.g., letters and numbers) also made a difference in the sensitivity decrement. Specifically, type of stimuli (sensory or cognitive) interacted with event rate and type of discrimination (simultaneous or successive). For simultaneous tasks that were also sensory, the sensitivity decrement decreased as event rate increased while the opposite was true for simultaneous cognitive tasks. When the discrimination was successive, that is when subjects had to recall from memory the target stimulus they were maintaining a vigil for, event rate increased the sensitivity decrement for sensory tasks and only slightly decreased it for cognitive tasks.

One key difference between what will be called the "traditional" vigilance literature in the current review and the tasks TSOs face at the checkpoint is the transience of the signal in the traditional vigilance tasks. In the See, Howe [4] meta-analysis all of the studies included in the analysis had a stimulus duration measured in milliseconds – the longest such stimulus was 1.5 s. Thus, as Wolfe, Horowitz [5] point out, a temporary distraction drawing the subject's attention from the display can cause them to miss an event. This is not true in the TSO's case – even in the case of a continuously moving x-ray belt.

1.2 Visual Search

Turning now to another relevant literature, the work on visual search does not often consider performance over time as a critical variable. Wolfe and colleagues [5, 6] reported data for d' and response bias as a function of time and failed to see definitive effects, although Wolfe, Horowitz [5] reported one instance of a bias increment.

There are a few sources of information about domain-specific visual search performance over time in the TSA environment and what literature does exist is sensitive. However it does bear mentioning that those results are mixed with some resulting in a decrement in d' or response bias, some an increment in d' or response bias, and some showing no change in performance over time.

1.3 Inspection

There is a separate literature from the Human Factors world on inspection that also informs the current study. Often, this work has been done with aircraft inspectors looking for fuselage defects (i.e., much of the work Colin Drury and colleagues have done over the years), but some of the work also includes inspection in manufacturing, food inspection, printed circuits, medical, and nuclear weapons domains [7]. In the studies of inspection that don't utilize a dynamic search task, the findings of performance decrements as a function of time in field studies are mixed. Some, such as Hartley, Arnold [8] found significant decrements in airborne inspectors looking for noxious weeds over the course of several hours in the field. Others, such as Spencer and Schurman [9] found no differences in Pd for a four-hour, self-paced eddy current inspection task [also see 3 for a review].

In a review of the literature at the time, Craig [10] offers an interesting observation several of the articles included in his review make – that although inspection tasks in the field can last for quite a long time, inspectors often only perform sustained inspection for short periods – between 5 and 15 min at a time. Presumably this is because inspection tasks in operational environments involve more than just visual inspection. In the TSA domain, this also seems to be the case. In addition to searching x-ray images for threats or prohibited items, the TSO performing x-ray analysis also has to indicate bag checks, can ask for advice from other TSOs, might have to wait for passengers to remove their bags from the exit of the x-ray before moving the belt along, etc.

1.4 Summary

Thus, while there may be an impact of time on task on performance, the TSA x-ray screening task is most likely not cognitively identical to a traditional vigilance task of the Mackworth or radar operator variety. Table 1 outlines some other key differences between these types of tasks.

Ultimately, the question at hand is not whether there is a performance decrement – it would be hard to believe that there would not be after some period of time, which may differ from individual to individual [11]. What is currently at issue is the time-course of this decrement and whether any of the factors that might be under the TSA's

Table 1. Comparison between traditional vigilance tasks (e.g., Mackworth, 1948) and x-ray screening or inspection tasks.

Traditional Visual Search or Inspection tasks	Traditional Vigilance tasks
Complex scenes under the control of the searcher (i.e., the TSO can stop the moving x-ray belt at will to examine a potential target)	Complex dynamic scene not under the control of the searcher
Can have multiple targets/classes of targets simultaneously	Usually has only one event at a time (e.g., appearance of a radar return, clock hand jump, etc.) that is transient target
In TSA it is self-paced in terms of how long the TSO spends making his decision about whether the scene contains a target or not– and the scene doesn't change until the TSO advances the belt	The task is usually task-paced in that targets appear and disappear as a function of the task timing, not as a function of the observer's decision process
Momentary lapse of attention won't result in an error with the Standard SOP (stopping the belt for every bag), but even in SOPs allowing a continuous moving belt, the TSO has the option to reverse the belt and stop it on a bag, thus a momentary lapse of attention can be corrected. Length of signal presence is measured in seconds and is under the control of the TSO	Momentary lapse of attention can result in a miss error that is not correctable, stimulus durations typically measured in milliseconds [e.g., in the 4 meta-analysis, stimulus durations ranged from 2 to 1500 ms]

control could have a mitigating effect on this decrement over a large number of TSOs (e.g., whether continuous belt results in a more dramatic performance decrement over time than does the static belt).

In studies employing a dynamic task in which subjects have to constantly monitor some screen or device for a transient signal, decrements often show up in the first 10–15 min [1, 2, 4]. For tasks with a less dynamic nature, such as inspection or bag screening, decrements sometimes show up that fast [11], sometimes they are slower to appear [7], and sometimes they don't appear at all [5, re-analysis of 12–15]. Thus, while decrements do occur, the timecourse is often different – and timecourse is of preeminent importance for the TSA operational environment for scheduling and personnel specialization reasons. Whether a longer-term decrement that might occur in the x-ray task is a result of the same underlying cognitive or physiological processes as that in the more traditional vigilance tasks is an interesting question – qualitative differences might point to different mitigations, for example – however, this question is not empirically addressed in the current work.

Rather, the goal of this work is to first determine whether there is a reliable decrement in performance during a 2-h task, measured in increments of 5 min [cf. 16 who present data in hour-long increments], whether there are reliable individual differences in the presence and timecourse of this decrement, and whether there are indicators of a change in performance that precede changes in Pd, Pfa, d' or c (such as patterns in eye movements, search patterns, patterns in image manipulation tool use, etc.).

2 Methods

Overall, ~ 240 TSOs from 6 airports will interrogate x-ray images of mock carryon items using software designed to emulate the look, feel, and a good proportion of the key functionality of one of the x-ray systems currently in use at the TSA checkpoint. Mock items were constructed by the TSA's Transportation Security Integration Facility (TSIF), located at Reagan International Airport in the Washington DC area. They will interrogate these images for 2 h without breaks. After the 2 h bag search task, they then will perform a 45-min battery of general visual cognition tasks (e.g., attention beam, a variant of Raven's progressive matrices, a version of the T and L task, and a pop-out task). The details of this cognitive battery are provided in Matzen [17] and won't be discussed further here. Each aspect of the bag search task is presented in the sections below.

2.1 Experimental Design

This will be a 6x2x2x2x24 mixed experimental design, illustrated in Table 2. The factors include:

- Airport – between subjects, 6 total airports
- SOP – 2 (PreCheck,[1] Standard) – between subjects
- Belt condition – 2 (static, continuous) – between subjects
- Threat Type – 2 (Clear, explosive) – within subjects
- 5-min epoch – 24, (with the possibility of increasing time epoch to 10 min in the event that we have too few observations per subject per 5-min epoch).

The rationale for including static belt on the PreCheck is to simply measure the separate effects of the static belt on TSOs' accuracy/response bias/decision time within the PreCheck context at the actual checkpoint. The rationale for including continuous belt in the Standard checkpoint lane is to explore the effects of this condition on our various dependent measures in anticipation of the inclusion of continuous belt in Standard lanes.

It is estimated that we will collect data from approximately 240 TSOs: approximately 40 TSOs from each of the 6 airports. The only requirements TSOs must meet is that the must have been a TSO for at least a year, they must regularly perform the x-ray task at the checkpoint, and they must be certified to function in both PreCheck and Standard lanes. Otherwise, no attempt will be made to select for gender, age, education, or other such demographic variables although this information will be collected in order to include it in the statistical analysis.

[1] PreCheck and Standard refer to the two different screening lanes currently in use at American airports. Passengers who are screened via PreCheck lanes have either registered with the TSA or have been opted in via airline status programs, because of their military service, or because they are a known crew member. In PreCheck lanes, the belt on the x-ray is run continuously unless the TSO sees something he or she wants to inspect futhrer, whereas in the Standard lanes, the TSO is required to stop the belt on every passenger item.

Table 2. Experimental design

Experimental Design		
SOP/Belt (Between Ss)	Threat Type (Within Ss)	5-min Epoch (Within Ss)
PreCheck - static	Clear	1–24
	Threat	1–24
PreCheck - continuous	Clear	1–24
	Threat	1–24
Standard - static	Clear	1–24
	Threat	1–24
Standard - continuous	Clear	1–24
	Threat	1–24

Dependent variables include (all calculated overall in addition to as a function of 5-min epoch):

- Image manipulation tool use (zoom, color changes, etc.)
- Probability of Detection (Pd), Probability of False Alarm (Pfa)
- Decision Time
- Eye Tracking – time to first fixation, number of fixations, types of errors, etc.
- Calculated variables – including d', c, search time consistency [18]
- Image product use (e.g., order of image manipulation tools, which tools selected, eye tracking patterns associated with each bag, etc.).

2.2 Stimuli

Stimulus Creation. All images were created using a Rapiscan AT-2 system. Threat prevalence was set at 10 %. We assumed TSOs would spend 10 s per image on average, thus a 120 min study will require images for at least 720 bags. However, in order to ensure all TSOs spent the entire 120 min interrogating bags we created image sets containing 1,000 unique items.

Stimuli were created by using an existing set of 500 items (including briefcases, rollerboards, car seats, etc.) from the Transportation Security Integration Facility (TSIF). Those items were imaged using the Rapiscan AT-2, then were re-packed adding and subtracting benign prohibited items and threats. Because of the differences in stream of commerce between PreCheck and Standard lanes at the airport, a total of 1,327 item images were created in order to have the stimulus sets closely mirror the expected stream of commerce for the two different types of lanes. The raw X-ray data from the Rapiscan imaging system were then transferred to an emulator that provided a full suite of X-ray data interpretation algorithms.

Each image was then manipulated using 31 manipulation tools on the emulator (32 in the event the emulator imposed target detection boxes around suspect items in the bag), and those image products were then saved as PNG files and transferred to a

standard computer system. Each item was also imaged from the top and side, thus each item had either 62 or 64 images associated with it, for a total of 83,624 PNG images generated. Note that this is not the complete set of images available to TSO's at the checkpoint, as the full set would have been prohibitive to collect. However, it does represent the most frequently used subset of image products (determined by a survey conducted of 10 airports from around the country).

In terms of the characteristics of the 1,327 items, there are three general classes of items:

- Threat items - 133 will have IEDs or IED components
- Cleared threat items - Those same 133 items will be re-imaged without the IED/component
- Clear items - The remaining 1061 items will not have threats

Note that each subject will only be able to interrogate 1,000 bags and that the additional bags were generated to reflect differences between what is permitted in bags in the PreCheck and Standard lanes. Thus, for each set of 1,000 bags, 99 contain threat items and those same 99 are cleared threat items. Thus, the threat prevalence rate in a given TSO's stimulus set is 9.9 %.

Across these three classes of bags, 3 % contain benign prohibited items (e.g., water bottles with water in them). 10 % of the items with benign prohibited items contain two benign prohibited items; the rest contain only one benign prohibited item. Overall, an effort was made to match the stimulus set characteristics to what TSA knows about the stream of commerce (in terms of numbers of laptops, iPads, types of bags, etc.).

Target detection boxes were included on images when they were generated by the Rapiscan emulator, and subjects were asked to clear them as a group (rather than individually). In this way, collection and analysis of dependent measures becomes fairly straightforward because for bags with target detection boxes there were two decisions (decision about the target detection boxes, decision about the overall bag) and for those bags without target detection boxes, there was one decision.

The presence and number of target detection boxes was not manipulated as a separate independent variable, but the number of target detection boxes (0 through N) will be entered as a covariate in the statistical model built to analyze the data. Target detection boxes for threat bags and for clear bags were those generated by the Rapiscan AT-2 system used to generate the images. For the cleared threat bags (i.e., threat bags with the threat removed and re-imaged), we re-generated target detection boxes around the same areas of the bags as the Rapiscan system generated for the threat versions of each bag, in order to hold that aspect of those bags constant. In this way, the cleared threat bags were as close to identical to the threat bags, with the exception of the threat presence, as possible, making an ideal stimulus set for the calculation of signal detection measures (d', c).

Stimulus Validation. Because we wound up actually generating 83,624 separate image files, we needed to validate that we had not duplicated any of the images and that we had generated all 31 or 32 image products for each separate X-ray data file. We did this in three ways.

First, we developed a naming convention that created a unique label for each image file that indicated which X-ray data file it belonged to, what view it was (top or side), and which of the 31 or 32 image manipulation tools was applied. This enabled us to validate that we had, at least in filename, all of the image files we expected to have for each X-ray file.

Second, we used hash values to determine whether we had any duplicate images. Each image has a unique hash value, which is essentially a character string that uniquely describes a file's content. Hash values are unique to images in much the same way that fingerprints are unique to individual humans.

Finally, we compared the pixel values of the PNG versions of the original false color image (provided by TSA alongside the corresponding X-ray files) with the pixel values for the normal false color images captured via the emulator by Sandia to determine if we had, in fact, matched all of the original files provided by TSA. After several iterations through the errors we found in these three ways, we reduced the error rate to zero. While this does not guarantee that there are no errors in the stimulus set, it does provide a certain amount of confidence that errors are rare.

2.3 Stimulus Presentation and Data Collection Platform

In the spirit of attempting to measure TSO performance in an environment as similar as possible to a live checkpoint, we developed custom x-ray emulation software that enabled us to collect a wide range of dependent measures. This software emulated the look and feel of the Rapiscan AT-2, and used a keyboard that was mocked up to look and operate like the Rapiscan operator control panel (OCP). While users did not have access to all of the image manipulation tools available on the actual Rapiscan system, a poll conducted of TSOs at several airports yielded the most common image manipulation tools used, and those were included in the functionality of the software.

The emulator software is designed to present each subject with a demographic questionnaire, instructions, practice bag images, and finally the experiment bag images. During both the practice images and experiment images the subject may perform similar operations to the Rapiscan AT-2 system, including: zoom in/out, start/stop image motion, change image variant (CC, BW, etc.), toggle target detection boxes, and reset the view. For each image the subject can indicate whether or not they believe the bag contains a threat. They are also prompted to count the number of threat and prohibited items they believe were in the bag.

The emulator collects detailed information about experiment state and subject behavior during the experiment. The system keeps track of which image is currently being displayed and its current position on the screen at every moment. Each command the subject issues to during the experiment is time-stamped in nanoseconds and associated with the bag and image being displayed at that time. Each of the subject's responses are also collected alongside subject demographics and experiment conditions such as subject ID, SOP and belt mode.

The emulator software is platform independent and can be installed on a machine with any operating system. It can operate with as few as two monitors, though a third monitor can greatly assist the operator with a detailed diagnostics display showing the

experiment's state at any point in time. The monitors used were the same monitors currently in use at the TSA checkpoints and operated at a 1280x768 resolution.

Integrated into the x-ray emulation software is a three-camera eye tracking system created by SmartEye. The system operates at 60 Hz. Data collected from the system includes screen intersection (x- and y- coordinates of gaze), fixation locations, and duration. A screen recorder also captures the video that is displayed to each subject so that the eye tracking information can be imposed on top. The software used for screen capture is EyesDX's Video Streamer and Record Manager.

2.4 Procedure

After providing informed consent, TSOs who volunteered to participate are seated at one of two instrumented stations. Before beginning the simulated task, the eye tracking cameras re repositioned and adjusted to ensure the subject is fully within the frame of the camera. The system is then further calibrated using an eight-point gaze calibration where subjects stare at eight strategically placed visual markers places across the two screens.

With the eye trackers and the display screen capture software recording, the subjects then answer an electronic demographic questionnaire. Upon completion, the TSOs re guided through instructions on the display regarding the task and information regarding the keyboard layout. Following the guided instructions, the subjects complete a ten-bag practice session where the experimenter walks them through the first few bags to ensure they understand the task and controls. Before beginning the main 2-h session,, the subjects have one last time to ask questions regarding the task. If no further questions remain, the main task begins.

After the participants complete the 2-h bag search task, are given a 15-min break during which they can leave the instrumentation station. Subjects then return to the stations for a visual cognitive battery with seven tasks. These tasks are detailed in Matzen [17].

3 Results

Statistical Analysis of the Behavioral Data. The data will be analyzed using multilevel growth modeling because of the hierarchical structure of the design and because of the interest in subject behaviors over time. Specifically, for all behavioral independent measures (Pd, Pfa, d', response bias, decision time), except for calculating signal detection measures (d', c), the cleared threat bags will be treated as clear bags in the analysis.

In addition to these more traditional statistical analyses, several machine learning techniques will be applied to the data in order to best exploit the millisecond-level eye tracking and state change logs being collected. These methods are detailed in Stracuzzi, Speed [19].

Acknowledgements. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. This research was funded by an Interagency Agreement between the Transportation Security Administration and the Department of Energy.

References

1. Mackworth, N.H.: The breakdown of vigilance during prolonged visual search. Q. J. Exp. Psychol. **1**, 6–21 (1948)
2. Molloy, R., Parasuraman, R.: Monitoring an automated system for a single failure: Vigilance and task complexity effects. Hum. Factors **38**, 311–322 (1996)
3. Drury, C.G., M. Saran, Schultz, J.: Effect of fatigue / vigilance / environment on inspectors performing floursecent penetrant and/or magnetic particle inspection: Year 1 interim report, F.A. Administration, Editor (2004)
4. See, J.E., et al.: Meta-analysis of the sensitivity decrement in vigilance. Psychol. Bull. **117** (2), 230–249 (1995)
5. Wolfe, J.M., et al.: Low target prevalence is a stubborn source of errors in visual search tasks. J. Exp. Psychol. Gen. **136**(4), 623–638 (2007)
6. Wolfe, J.M., Van Wert, M.J.: Varying target prevalence reveals two dissociable decision criteria in visual search. Curr. Biol. **20**, 121–124 (2010)
7. See, J.E.: Visual Inspection: A Review of the Literature. S.N. Laboratories, Mexico (2012)
8. Hartley, L.R., et al.: Vigilance, visual search and attention in an agricultural task. Appl. Ergonomics **20**(1), 9–16 (1989)
9. Spencer, F., Schurman, D.: Reliability assessment at airline inspection facilities Vol III: Results of an eddy current inspection reliability experiment. F.A. Administration, Editor (1995)
10. Craig, A.: Field studies of human inspection: the application of vigilance research. In: Folkard, S., Monk, T.H. (eds.) Hours of Work: Temporal Factors in Work Scheduling, pp. 133–145. Wiley, New York (1985)
11. Washburn, D.A., et al.: Individual differences in sustained attention and threat detection. Cogn. Technol. **9**(2), 30–33 (2004)
12. Speed, A.: The possible effects of judge-advisor groups on local baggage screener accuracy. Transportation Security Administration (2011)
13. Speed, A., Gaspelin, N., Ruthruff, E.: The effects of social pressure and image resolution on x-ray screener accuracy. Transportation Security Administration (2012)
14. Smith, V.J.: Test and evaluation report for the screener workload fatigue study. Transportation Security Administration (2005)
15. Haass, M.J., et al.: Understanding expert visual search (2014)
16. Ghylin, K.M., et al.: Temporal effects in a security inspection task: breakdown of performance components. In: Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting (2007)
17. Matzen, L.E.: Effects of professional visual search experience on domain-general and domain-specific visual. In: HCI International. Los Angeles, CA (2015)
18. Biggs, A.T., et al.: Assessing visual search performance differences between transportation security administration officers and non-professional visual searchers. Vis. Cogn. **21**, 330–352 (2013)
19. Stracuzzi, D.S., et al.: Exploratory analysis of visual search data. In: HCI International, Los Angeles, CA (2015)