Temporal Effects in a Security Inspection Task: Breakdown of Performance Components

Ghylin, K.M., Drury, C.G., Batta, R and Lin, L.
University at Buffalo: SUNY, Department of Industrial & Systems Engineering
435 Bell Hall, Buffalo, NY 14260

Data from certified screeners performing an x-ray inspection task for 4 hours, or 1000 images, were analyzed to identify the nature of the vigilance decrement. The expected vigilance decrement was found, with performance measured by probability of detection (PoD) and probability of false alarm [P(FA)] decreasing from hour 1 to hour 4. Correlations between PoD and P(FA) indicate that sensitivity between hours remained the same, however a shift in criterion (Beta) occurred. Significant decreases in both detection and stopping time were found from the first hour to the second, third, and fourth hour. Evidence of changes in the search component of the time per item was found to account for part of the vigilance decrement. As the task continued, participants spent less time actively searching the image, as opposed to other activities. Evidence is provided for truncation of active search as security inspection continues.

INTRODUCTION

It is well accepted that maintaining vigilance within a security screening task is critical to assuring the security of the system. Naturally occurring decreases in performance by x-ray screeners inspecting luggage in the airport could result in devastating consequences if a prohibited threat object were missed. Historically, performance decreases across time in tasks requiring sustained attention or effort (Mackworth, 1970). A vigilance decrement, resulting in a reduction in fault detection performance, is expected in any task where the signal to noise ratio is unpredictable and automatic processing is not possible (Fisk & Schneider, 1981). This is the case in x-ray security screening where the probability of occurrence of a threat item is low and random, and each image is unique. Vigilance decrements in security inspection have yet to be fully explored, with recent studies showing vigilance relationships with high-performing x-ray screeners (McCallum, Bittner, Rubinstein, Brown, Richman, & Taylor, 2005). Over 1000 studies and theoretical papers have been published investigating human vigilance and the dynamics of the visual process (Mackie, 1987). However, the true nature of the vigilance decrement is still not clear and many variables can account for the changes in performance across time.

Changes in perceptual sensitivity (measured by A' or d'), or a shift in observer's criterion (measured by β), or a combination of both have been postulated to account for the loss of vigilance (Mackworth, 1970). Shifts in behavior can be grouped into a sensitivity decrement or a bias increment (Wickens and Hollands, 2000). As signals become more infrequent, and harder to discriminate, operators shift their response bias towards the non-signal events, resulting in a sensitivity decrement (Craig, 1987). Other theories, which believe the vigilance decrement is a combination of exceeding the information processing capabilities and inadequate training causing a reduction in effort across time (Williams, 1986), also contain properties of a sensitivity decrement. Vigilance is in fact an effortful activity (Finomore et al, 2006). These theories postulate that the decline in performance in vigilance tasks is due to the requirement for continuous and repetitive processing of visual information while consciously controlling the search and decision-making processes (Fisk & Schneider, 1981).

A second group of theories argues that the vigilance decrement is due to a bias increment, or change in operator criterion. Characteristics of the conditions under which this arises generally include a low probability of an event signal combined with low levels of feedback (Wickens and Hollands, 2000). These changes can be thought of as a decreased expectancy of a signal, for example between training and on-line tasks.

The first documented study of vigilance was performed during World War II and considered a watch keepers ability to sustain attention during 30-minute increments across a four-hour task (Mackworth, 1948). Mackworth found marked decreases in performance for every 30-minute's on the task (the smallest measure of time collected). These results have since been challenged as other studies have shown vigilance drops about half-way after the first 15-minutes of inspection (Tiechner, 1974), with multiple studies showing large drops during the first 30-minutes (Teichner, 1974, Craig, 1985) with little reduction of performance beyond 30 minutes (Teichner, 1974).

Various studies of both laboratory inspection tasks, and field inspection processes not directly related to security inspection, have found significant decrements in performance over time time. Beginning with laboratory tasks, Thackery (1992) and Gramopadhye (1992) both found small decrements (1%-5% decrease) in performance between the first and second halves of 60 and 90 minute simulated eddy current inspection tasks. Experiments under field eddy current inspection conditions did not however show the same decrement. A self-paced, 4-hour task, with inspector led rest breaks, found *no* significant change in hit

or false alarm rate between the first and second halves of the task (Spencer & Schurman, 1995). A similar eddy current experiment examined work of either 30 or 90 minute periods over six days, and found no difference between hit rate and false alarm rate between the 30 and 90 minute duration (Murgatroyd, Worrall & Waites 1994). Contradictory results were obtained with a fluorescent penetrant inspection (FPI) task. This study, performed by Drury, Green, & Lin (2006), directly investigated the effects of period (1 vs. 2 hour task), breaks (none vs. every 20-min) and time on task (defined as 20-minute increments) on performance variables. Utilizing trained FPI inspectors, the study found a significant decrease in detection rate and false alarm rate as a function of time on task. However, performing the same study as a laboratory task with well trained industrial-experienced participants showed no significant change in detection rate across time, although the false alarm rate decreased over time. Results showed that a 3-minute break after each 20-min period positively improved performance, especially for P(FA) and speed, but only during the day and only for the 2-hour task. Other inspection tasks not related to aviation or security have also found a vigilance decrement. A study of the inspection of chicken carcasses on a processing line found an initial warm-up period and subsequent increase in hit rate, followed by a slow decline in hit rate over more than an hour. Eventually the rates reached a level below the initial hit rate (Chapman & Sinclair, 1975). A study of inspection of rubber seals found a significant decrease in performance time from the first- to second- 15 minute period of the task (Fox, 1977). The effects of time of day have also been found to give mixed results. A study measuring an expert's detection of noxious weeds while flying over fields in a helicopter found detection performance to be significantly lower (by almost half) between the mornings and afternoons for half-day sessions. Full-day sessions produced even greater decrements (Hartley et al, 1989). Together these studies of inspection tasks provide a picture of performance decreases over time periods ranging from 15 minutes to half-a-day, with some notable exceptions. Horowitz, Cade, Wolfe, and Cziesler (2003) add another element to the presence of a vigilance decrement. They argue that a decrement may only be found in tasks having no search element. However, most inspection tasks do contain some form of search.

While understanding the presence of any vigilance decrement is important, having evidence of the nature of the decline and the underlying mechanisms should provide information on how to design jobs to reduce performance changes. Within the security inspection tasks, recent research has shown that inspection time is comprised of both search and non-search components (Ghylin, Drury, & Schwaninger; McCarley, Kramer, Wickens, Vidoni, & Boots, 2004). If changes in search behavior, such as a reduced time spent searching, are found, that may help explain the nature of the vigilance decrement, as probabilities of both detection and false alarm decrease with reduced search time. The goal of the current study is to explore changes in vigilance by experienced security inspectors performing a simulated x-ray screening task. Note that the data came from a pre-existing study, made available to us from a national security agency.

METHOD

Participants (n=66) were all current airline security screeners. Participants were asked to identify threats within an x-ray image for 4 hours, or 1000 images, whichever came first. Images were combined into 4 bag sets, each containing an equal number of one of 8 threat types. Each participant saw each image bag set only once; sets were presented in a random order. Presented images were simulated x-ray images of carry-on bags typically found in a security screening setting. The eight threat types consisted of guns, knives, opaque objects, and five distinct types of improvised explosive devices (IEDs). Responses (Threat or No-Threat present), reaction times (measured in seconds), and threat image type (when present) were recorded for each response for each participant.

RESULTS

[NOTE: Due to the security sensitivity of the data, probabilities of detection and false alarm have been scaled by an arbitrary constant for this report.]

Although past research has shown that a significant decrement in performance occurs after 15 to 20 minutes (Fox, 1977), data obtained for this study was limited and only available based on the hour that the task was performed during (1, 2, 3, or 4). Thus, the following analysis will take into account performance metrics based on time-on-task in hour blocks. We start analysis with overall performance and proceed through reaction times to the analysis of search and non-search times.

Overall Performance Analysis

Detection performance, in terms of A', probability of detection (PoD) and probability of false alarm [P(FA)] were subject to ANOVAs with Hours and Bag Set as fixed independent factors and Participants as a random factor. As seen in Table 1, PoD and P(FA) were significant for Hour, but not for A'. Across the four hours of the test, PoD and P(FA) decreased, while A' did not change.

Factors	df	PoD	P(FA)	A'
Hour	3	$p < 0.001$	$p < 0.001$	N.S.
Bag Set	3	$p < 0.001$	N.S.	$p < 0.001$
Participant	65	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table 1: ANOVA results for overall performance

Performance measures of hit rate and false alarm rate were correlated across participants for each hour as seen in Table 2. Uniformly positive correlations resulted, signaling common sensitivity over the four Hours, and (not shown here) confirming the insignificant change in A' (Green & Swets, 1966; Hofer & Schwaninger, 2004). Participants had approximately the same level of overall performance (A') by hour and variation existed only between participants. As PoD and P(FA) both decreased between hours, we conclude that criterion shifts occur between hours for participants.

Hour	r	p-value
1	0.55	< 0.001
2	0.634	< 0.001
3	0.487	< 0.001
4	0.431	= 0.001

Table 2: PoD/P(FA) correlations of reach hour

Reaction Time Analysis

Reaction times for hits, false alarms, correct rejections, and misses were calculated and analyzed separately. The time to correctly identify a threat was detection time, whereas the time taken to end the search without finding a threat, or search time, was called stopping time. Reaction times inter-correlated significantly across participants (all $r \geq 0.50$ at $p < 0.001$). In mixed model ANOVAs, highly significant effects of Hour were found for all response times. [Hits: $F(3,256) = 16.36$, $p < 0.001$; False Alarms: $F(3,256) = 28.18$, $p < 0.001$; Correct Rejections: $F(3,256) = 63.56$, $p < 0.001$; and Misses: $F(3,256) = 20.61$, $p < 0.001$]. All times decreased over Hours.

A factor analysis of the four time measures and the two performance measures [PoD and P(FA)] produced only two factors explaining 76.4% of the variance. Factor 1 contained the two performance measures, while Factor 2 contained the four time measures, justifying the separate treatment of speed and accuracy.

ANOVAs of detection time and stopping time were performed with Hour and Bag Set as fixed factors and participants as a random factor, as shown in Table 3.

Factors	df	Detection Time	Stopping Time
Hour	3	p < 0.001	p < 0.001
Bag Set	3	N.S.	N.S.
Participant	65	p < 0.001	p < 0.001

Table 3: ANOVA results for response times

Tukey tests, with 95% confidence, were performed on Hour for PoD, P(FA), detection time, and stopping time and can be seen in Table 4.

PoD			
Hour	2	3	4
1	N.S.	p = 0.0078	p = 0.0002
2		N.S.	N.S.
3			N.S.
P(FA)			
Hour	2	3	4
1	p < 0.001	p < 0.001	p < 0.001
2		N.S.	p = 0.0392
3			N.S.
Detection Time			
Hour	2	3	4
1	p < 0.001	p < 0.001	p < 0.001
2		N.S.	N.S.
3			N.S.
Stopping Time			
Hour	2	3	4
1	p < 0.001	p < 0.001	p < 0.001
2		p = 0.0026	p < 0.001
3			N.S.

Table 4: Tukey tests for significant results on Hours

The Tukey comparisons showed that the PoD decreased significantly only when comparing hour 1 to hours 3 and 4. However, the P(FA) significantly decreased while comparing hour 1 to hours 2, 3, and 4, and when comparing hour 2 to hour 4. Reaction times produced somewhat similar patterns, with detection time significantly decreasing between hours 1 and hours 2, 3, and 4 only. Stopping time showed significant decreases between hours 1 and hours 2, 3, and 4, also from hour 2 to hours 3 and 4.

Search and Non-search Times Analysis

The basic results indicate a change in performance across hours, with differing patterns for PoD and P(FA). Similarly, detection times and stopping times decline over the four hours. Perhaps the reduction in time spent searching could account for some of the accuracy changes. To further investigate this idea, the raw time data were transformed to cumulative probability distributions and fitted to a search and non-search equation to further determine the time components involved in both search and non-search activities. [For explanation of these equations, see Ghylin, Drury, & Schwaninger, 2006]. To obtain a good fit, only data sets with five or more responses were used, resulting in 264 [264 = n(66) x levels(4)] fits for Hits and 230 entries for False Alarms. The analysis produced estimates of mean search and non-search times for hits and false alarms. ANOVAs were performed with these times as dependent variables. Results are in Table 5. [Please note the change in table direction.]

Factors	Hour	Bag Set	Participant
df	3	3	65
Hit Search Time	p < 0.001	N.S.	p < 0.001
Hit Non-search Time	N.S.	P = 0.068	p < 0.001
FA Search Time	p < 0.001	N.S.	p < 0.001
FA Non-search Time	p < 0.001	P = 0.007	p < 0.001

Table 5: ANOVA results for fitted time measures

Results showed significant changes (decreases) in Hit search time, FA search time and FA non-search time by hours. A Tukey test with 95% confidence was then run. Results are in Table 6.

Hit Search Time			
Hour	2	3	4
1	p = 0.0003	p < 0.001	p < 0.001
2		N.S.	N.S.
3			N.S.
Hit Non-search Time			
Hour	2	3	4
1	N.S.	N.S.	N.S.
2		N.S.	N.S.
3			N.S.
FA Search Time			
Hour	2	3	4
1	p = 0.0337	p < 0.001	p < 0.001
2		p = 0.0605	p = 0.0001
3			N.S.
FA Non-search Time			
Hour	2	3	4
1	p = 0.0029	p < 0.001	p < 0.001
2		N.S.	N.S.
3			N.S.

Table 6: Tukey tests for significant results on Hours

Most measures showed that only the first hour was different from the other three. Search and non-search parameters for each hour, along with detection and stopping times are plotted in Figure 1. This shows that at least some of the decrease in the PoD and P(FA) that occurred throughout the task can be predicted by the changes in the search and non-search parameters.

DISCUSSION

A significant decrease in overall performance across hours indicated the presence of a vigilance decrement, as expected and as found in many previous studies (Mackworth, 1970, Davies & Parasuraman, 1982). Due to the limited details available about the collected data, finer grained analysis than hours was not possible. However, even at the hour level, the historical change in performance across time was seen, with both the performance measures declining as time progressed.

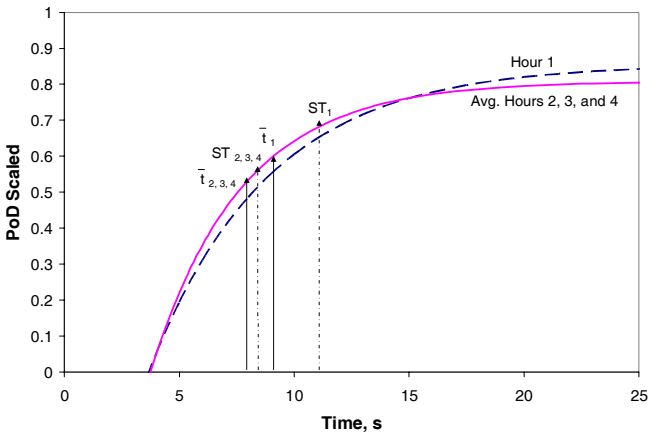

Figure 1: Combined search and non-search times, along with stopping time (ST) and detection time (t) by hour 1 and hours 2, 3, and 4

With regards to Signal Detection Theory (SDT) changes in performance can either be attributed to a change in sensitivity or a shift of participant criterion (Green & Swets, 196). For a change in sensitivity to be present, a decrease in PoD with a relatively constant false alarm detection rate would be expected. This results in a true performance decrement, where the participant has a more difficult time distinguishing between targets and non-target events. Alternately, if both PoD and P(FA) both decrease, then a bias change has occurred and the participant is generally less willing to report anything, signal or not. Positive correlations between hits and false alarms provide further evidence that sensitivity did not change across the runs, but rather there was a shift in operator criterion, or Beta. An insignificant change between hour 1 to hour 2, paired with a significant decline in detection times was found. These changes could be explained by participants performing the search part of the task faster, without in fact changing their reporting criterion.

Significant reductions in reaction times across hours for Hits, Miss, False Alarms, and Correct Rejections provide evidence that operators spent less time searching images as time progressed. These reductions in time were especially significant between the first hour and the remaining hours. This supports past findings of performance changing rapidly at first during the task. However, the PoD was not significantly different between hour 1 and hour 2 but was significantly different between hour 1 and hours 3 and 4. This indicates that after any initial performance reduction occurred in the task, a secondary performance reduction occurred after 2 hours of performing the task. This disagrees with past findings (Teichner, 1974) that suggest no significant performance decrease after only 30 minutes on the task (Murgatroyd, Worrall & Waites, 1994). Limitations however exist, as many studies only reported time durations of up to 90 minutes (Gramopadhye, 1992; Murgatroyd, Worrall & Waites, 1994).

The search and non-search components of the data showed evidence of a significant decrease in search time for both hits and false alarms and a significant decrease in non-search time for false alarms. However, no significant change in hit non-search time was found. If learning was occurring during this task, we would expect to see the hit non-search time improve as well. As images became more familiar to the operator, they would be spending less time discerning the true nature of the image. However, what we find is that participants were spending less time searching an image with time on task for both hits and false alarms. Less time was also spent discerning the potential threats in the false alarms. However, whether due to training or experience, the time spent discerning the potential threat items that were correctly tagged as threats, did not change. This may provide evidence that due to training or experience, this aspect of the process is more automatic and simple for the experienced participant than previously believed. Further research is needed in order to better understand this area.

The current study only focused on a continuous task without breaks. Past research has shown that providing breaks, whether paced (Drury, Green, & Lin, 2006) or self-paced (Spencer & Schurman, 1995) helps negate any significant changes in performance, even for tasks that are 4-hours in duration. Had breaks been given, which is certainly the case for many countries, the performance decrements may have been reduced or eliminated from the current study.

CONCLUSIONS

This study provided further evidence of the presence of a vigilance decrement within security inspection. Significant decreases in PoD occurred between the first and third hour of the task. Significant reductions in P(FA), detection time, and stopping time occurred after the first hour of the task. This indicates the reduction of performance occurring even further along in an inspection task, past the 30-minute mark. Changes in the time spent searching the images were found as time on task increased. Future research needs to consider the effects of breaks on the extended task. Data suggests both a criterion shift to be the cause of the vigilance decrement within this study, as well as a truncation of active search-time as time on task increases.

REFERENCES

Chapman, D.E. & Sinclair, M.S. (1975). Ergonomics in Inspection Tasks in the Food Industry, in C.G. Drury and J.G. Fox (eds), *Human Reliability in Quality Control* (London: Taylore & Francis), pp. 241-251.

Craig, A. (1987). Signal Detection Theory and Probability Matching Apply to Vigilance. *Human Factors*, 29, 645-652.

Davies, D.R. & Parasuraman, R. (1982). *The Psychology of Vigilance*. San Diego, CA: Academic Press.

Drury, C.G., Green, B.D. & Lin, J.F. (2006). Fatigue and FPI Inspection.

Fisk, A.D. & Schneider, W. (1981). Automatic and Control Processing during Tasks Requiring Sustained Attention: A New Approach to Vigilance. *Human Factors*, 25, 391-399.

Finomore, V. S. Warm, J. S., Matthews, G., Riley, M.A., Dember, W. N., Shaw, T. H., Ungar, N. R. and Scerbo, M. W. (2006) Measuring The Workload Of Sustained Attention *Proceedings Of The Human Factors And Ergonomics Society 50th Annual Meeting—2006, 1614-1618*

Ghylin, K.M., Drury, C.G., & Schwaninger, A. (2006). Two-component Model of Security Inspection: Application and Findings. *Proceedings of the 16th World Congress of the International Ergonomics Association*, 2006.

Gramopadhye, A.K. (1992). Training for Visual Inspection. PhD Dissertation, State University of New York at Buffalo, Buffalo, NY.

Green, D. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Wiley: New York, NY.

Hofer, F. & Schwaninger, A. (2004). Reliable and Valid Measures of Threat Detection Performance in X-ray Screening. *IEEE ICCST Proceedings*, 38, 303-308.

Horowitz, Cade, Wolfe, & Cziesler (2003).

Mackie, R.R. (1987). Vigilance Research: Are WE Ready for Countermeasures? *Human Factors*, 29, 707-723.

Mackworth, N.H. (1948). The Breakdown of Vigilance during Prolonged visual Search. *Quarterly Journal of Experimental Psychology, 1*, pp. 6-21.

Mackworth, J. (1970). *Vigilance and Attention*. Penguin Publishers, Baltimore, MD.

McCallum, M., Bittner, A., Rubinstein, J., Brown, J., Richman, J. & Taylor, R. (2005). Factors Contributing to Airport Screener Expertise. *Proceedings of the Human Factors and Ergonomic Society 49th Annual Meeting*, p. 922-926.

McCarley, J.S., Kramer, A.F., Wickens, C.D. Vidoni, E.D., & Boot, W.R. (2004). Visual Skills in Airport Security Inspection. *Psychological Science*, 15, 302-306.

Murgatroyd, R.A. Worrall, G.M. & Waites, C. (1994). A Study of the Human Factors Influencing the Reliability of Aircraft Inspection, AEA/TSD,0173. Risley, AEA Technology.

Spencer, F.W. & Schurman, D.L. (1995). Reliability Assessment at Airline Inspection Facilities, Vol III: Results of an Eddy-Current Inspection Reliability Experiment, DOT/FAA/CT-92/12, III, May 1995.

Teichner, W.H. (1974). The Detection of a Simple Visual Signal as a Function of Time of Watch. *Human Factors*, 16(4), pp. 339-353.

Wickens, C.D. & Hollands, J. (2000). *Engineering Psychology and Human Performance*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Williams, P.S. (1986). Processing Demands, Training, and Vigilance Decrement. *Human Factors*, 28, 567-579.