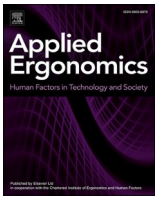
Applied Ergonomics 111 (2023) 103995

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Time on task and task load in visual inspection: A four-month field study with X-ray baggage screeners

D. Buser [a], A. Schwaninger [a], J. Sauer [b], Y. Sterchi [a,*]

[a] University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Riggenbachstrasse 16, CH-4600, Olten, Switzerland
[b] University of Fribourg, Department of Psychology, Rue P.A. de Faucigny 2, CH-1700, Fribourg, Switzerland

ARTICLE INFO

Keywords:
Time on task
Visual search
X-ray image inspection

ABSTRACT

Previous studies suggest that performance in visual inspection and typical vigilance tasks depend on time on task and task load. European regulation mandates that security officers (screeners) take a break or change tasks after 20 min of X-ray baggage screening. However, longer screening durations could reduce staffing challenges. We investigated the effects of time on task and task load on visual inspection performance in a four-month field study with screeners. At an international airport, 22 screeners inspected X-ray images of cabin baggage for up to 60 min, while a control group (N = 19) screened for 20 min. Hit rate remained stable for low and average task loads. However, when the task load was high, the screeners compensated by speeding up X-ray image inspection at the expense of the hit rate over time on task. Our results support the dynamic-allocation resource theory. Moreover, extending the permitted screening duration to 30 or 40 min should be considered.

1. Introduction

The continuous visual inspection of X-ray images of passenger baggage is legally limited to 20 min at European airport security checkpoints (European Commission, 2015). Thereafter, security officers (screeners) take a break of 10 min or rotate positions to perform a different task. While this regulatory limit might prevent a decrease in performance, it restricts options for staffing and can lead to operational challenges. The current time limit does not originate from research in X-ray image inspection, but it is believed to be based on findings from vigilance research (personal communication with an airport security expert, March 2019). There, a decrease in performance was often observed after about 15 min (Davies and Parasuraman, 1982; Mackworth, 1948; See, 2012; Teichner, 1974) or even earlier in difficult tasks (Jerison, 1963; Nuechterlein et al., 1983). The decrease in vigilance, called vigilance decrement, typically manifests as fewer detections and slower response times (Davies and Parasuraman, 1982; See et al., 1995). Additionally, it is frequently accompanied by a decrease in task engagement and an increase in distress, compared to pre-task values (Claypoole et al., 2019; Teo and Szalma, 2011; Tiwari et al., 2009; Warm et al., 2008a).

The underlying causes of the vigilance decrement have been predominantly explained by two different theories (Helton and Warm, 2008; MacLean et al., 2010; Neigel et al., 2020). Resource theory assumes that maintaining attention depletes limited attentional resources, which causes a decline in performance (Helton and Warm, 2008; Matthews et al., 2010). This is supported by the observation that vigilance declines more strongly when the event rate (number of stimuli to be processed per time unit) is higher (Claypoole et al., 2019; Davies and Parasuraman, 1982; See et al., 1995). Underload theory assumes that vigilance tasks' monotony induces under-stimulation that causes lapses in attention, whereby targets go undetected (Robertson et al., 1997).

However, visual inspection differs from typical vigilance tasks (Drury and Watson, 2002). X-ray image inspection involves visual search and decision making (Koller et al., 2009) regarding visually complex stimuli (Schwaninger et al., 2005) and it requires multiple target search (Biggs et al., 2018; Biggs and Mitroff, 2015; Donnelly et al., 2019; Godwin et al., 2010a). In traditional vigilance tasks, simple and single signals must be distinguished from background noise (Davies and Parasuraman, 1982). Moreover, visual inspection tasks, such as in X-ray baggage screening or industrial inspection, elicit a different vigilance decrement pattern compared to traditional vigilance tasks: The decrease in detected targets is often accompanied by a decrease in reaction times and false alarms (Basner et al., 2008; Ghylin et al., 2007). To account for

* Corresponding author.
E-mail addresses: daniela.buser@fhnw.ch (D. Buser), adrian.schwaninger@fhnw.ch (A. Schwaninger), juergen.sauer@unifr.ch (J. Sauer), yanik.sterchi@fhnw.ch (Y. Sterchi).

https://doi.org/10.1016/j.apergo.2023.103995
Received 19 July 2022; Received in revised form 20 October 2022; Accepted 6 February 2023
Available online 17 May 2023

this alternative pattern, Rubinstein (2020) proposed the dynamic-allocation resource theory (DART), which suggests that vigilance decrements in inspection tasks are caused by changes in behavior to preserve resources as opposed to limited resources or under-stimulation. More specifically, searchers try to save resources by speeding up and changing their response tendency. A few studies have investigated how time on task affects performance in X-ray baggage screening and reported results consistent with this pattern. Ghylin et al. (2007) investigated screener performance by comparing the performance of professional screeners between four 1-h blocks in a laboratory study. They observed a decrease in the hit rate, false alarm rate, and reaction times when comparing the first hour to the later hours indicating a shift in response tendency rather than a decrease in sensitivity. Buser et al. (2020) investigated how performance changed during 60 min of baggage screening among professional screeners. One group took a 10-min break after every 20-min of screening, the other group analyzed X-ray images for 60 min continuously. Performance was compared for three consecutive 20-min blocks. At a target prevalence of 12.5%, the false alarm rate decreased from the first to the second 20-min block, whereas processing times decreased continuously across blocks. Consistent with the results of Ghylin et al. (2007), a change in response tendency was found from the first to the second 20-min block. Additionally, screeners who worked continuously reported more distress than those who took breaks. To understand the study from Meuter and Lacherez (2016), threat image projection (TIP) first has to be introduced. During X-ray baggage screening at airports, the frequency of real threat articles (target prevalence) is very low, and a low frequency of targets reduces their detection (Godwin et al., 2010b; Wolfe et al., 2005, 2007). Airports counteract this by projecting prerecorded images of threat items (fictional threat items, FTIs) onto randomly selected X-ray images of passenger baggage using a technology called threat image projection (TIP) (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005). Therefore, screeners are exposed to more threats. Because it is recorded whether a TIP was detected by the screener or not, TIP data can be used to calculate the screeners' hit rates as an indicator of their detection performance (Meuter and Lacherez, 2016; Skorupski and Uchroński, 2016). Meuter and Lacherez (2016) used TIP data from an airport to investigate the effects of time on task and event rate (number of analyzed images per minute; task load) on detection performance for screening durations of up to 30 min. They found no main effect of time on task; however, they observed an interaction effect of time on task and event rate. Screeners showed a stable hit rate over time when the event rate was low. However, when the event rate was high (more than 5.4 analyzed images per min), the hit rate dropped from 94% to 92% after 20 min and to 87% after 30 min of screening. Chavaillaz et al. (2019) investigated how different break regimes affect detection performance for 60 min of X-ray image inspection in novices. The participants took 10-min breaks every 20 min, 5-min breaks every 20 min, or spontaneous breaks during a 1 h simulated baggage screening task. The study found no performance differences among the break regimes. Therefore, the researchers concluded that more flexible breaks could be implemented, granting the screeners autonomy to take spontaneous breaks when necessary.

Although previous studies indicate how screener performance evolves over time, no conclusions can be drawn about how professional screeners' performance changes over 1 h under real working conditions. Furthermore, airports increasingly move the screening of cabin baggage away from the checkpoint to remote screening rooms. This quieter working environment could have a positive impact on performance (Kuhn, 2017) and reduce performance decline over time on task. Therefore, this study investigated how screener performance evolves with time on task under remote screening conditions. We conducted a four-month study at an international airport using TIP data to investigate whether performance changes with time on task of up to 60 min under real working conditions and whether the effect of time on task is moderated by task load. Based on the DART (Rubinstein, 2020), we hypothesized that the hit rate, reject rate, and processing time decrease with increase in time on task and task load. Moreover, we evaluated whether there is an interaction between time on task and task load.

2. Methods

2.1. Participants

The study was conducted at an international airport with a workforce of about 100 screeners and with several checkpoints. About 50 screeners worked regularly at the checkpoint where the study was conducted. We created two groups by random assignment and, after the study, selected screeners who completed a minimum of eight X-ray baggage screening sessions. Consequently, 41 screeners who met the criteria were selected; the study group engaged in screening for up to 60 min (22 screeners, 11 females; mean age: 30.77 years, SD = 8.38; mean tenure: 3.66, SD = 1.41), and the control group (19 screeners, 9 females; mean age: 34.89 years, SD = 10.97; mean tenure 2.80, SD = 1.42) continued screening as before, which suggests that screeners should rotate their position after 20 min of continuous X-ray image inspection. Our study is based on 2′376 baggage screening sessions (study group: 1′170, control group: 1′206), where 436′512 X-ray images were analyzed. The study was conducted during the screeners' regular working hours without affecting their compensation. The study complied with the American Psychological Association's Code of Ethics and was approved by the Institutional Review Board of the School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland. Informed consent was obtained from all screeners prior to the study. The national civil aviation authorities permitted this study under the condition that we monitor detection performance regularly and stop the study immediately if the airport's security is threatened.

2.2. Materials

In addition to analyzing TIP data, subjective stress was measured using the Short Stress State Questionnaire (SSSQ; Helton, 2004). Screeners were asked to fill in the questionnaire every three weeks after completing a screening session. Upon completion of the study, screeners completed a short survey, which included questions regarding the screening durations. The survey also included airport specific questions comparing different technologies, which are not reported here.

2.3. Procedure

The screeners were informed about the study verbally and in writing by their employer and a member of our team. Their supervisors informed them whether they had been assigned to the study or control group. The study group was instructed to screen for up to 60 min; however, they were given the option to stop earlier if they felt tired or unconcentrated. They were asked to note down the reason for ending a screening session before completing 60 min on a sheet next to their workstation. The control group continued to work by following the current EU regulation. Both groups screened X-ray images of cabin baggage from a remote room, which was located close to the checkpoint and typically staffed by one or two screeners. Screeners were limited to 20 s to decide whether a baggage contained a prohibited article. This included marking the identified article and assigning it to a threat category. They received direct feedback on TIP images with the TIP system indicating FTIs in the X-ray image. For security reasons, all bags containing an FTI were rescreened. After X-ray screening, screeners from both groups switched to a different position at the checkpoint or took a break. The study lasted 18 weeks between January and May. Because adapting the daily operation to longer screening sessions required some time, the first two weeks were excluded from the data analysis.

2.4. Dependent measures

We considered the following dependent variables in the mixed models to assess how performance evolved with time on task: Detection performance (hit rate), reject rate, and processing time. Hit rate was the percentage of correctly identified TIP images. Reject rate was the percentage of all bags sent to a manual bag search. Because the available data only indicated whether a bag was rejected but not whether a real prohibited article (e.g., a bottle of water, knife, etc.) was present, the reject rate is not equivalent to the false alarm rate, which is the percentage of bags that are harmless but wrongly classified as containing a prohibited article. Processing time was the number of seconds screeners took to decide whether an image contained a prohibited article (rounded to full seconds by the TIP system). The session duration was the time difference between the screeners' login and logout for each screening session. For comparing the performance between the study and the control group, the hit rate, reject rate, and mean processing time were calculated for each screener.

2.5. Data analysis

All statistical analyses were performed using R (R Core Team, 2020). Because we were interested in the effects of time on task, we excluded short screening sessions under 10 min. Despite the instruction to screen for up to 60 min, 16 screening sessions exceeded 70 min and were also excluded. This resulted in the exclusion of 3.08% of all images and 3.09% of TIP images. The analyzed sessions were conducted between 04:00 and 20:00. Sessions after 20:00 occurred rarely (0.25% of all images and 0.30% of TIP images) and were therefore excluded. Because we calculated task load as the mean number of images analyzed per minute from the beginning of the session, the first screening minute of each session was omitted, which led to the exclusion of 4.09% of all images and 3.80% of TIP images.

We used mixed-effects models to assess the effects of time on task and task load on hit rate, reject rate, and processing time. The three models included time on task, task load, Time on task × Task load, days since study start, and daytime as fixed effects, and the session nested in the screener as random effects (see Equation (1)). The interaction Time on task × Task load was included because it was found in the only previous study on time on task and task load in X-ray baggage screening by Meuter and Lacherez (2016). Time on task was the number of minutes spent logged into a screening session by the screeners when they analyzed an image and was therefore calculated as the difference between the login time and the time when the decision for that image was made. Sessions conducted by the same screener that were less than 2 min apart were combined and treated as one screening session. Task load was the mean number of images a screener analyzed per minute from the start of the screening session. Days since the beginning of the study was included to examine whether habituation or fatigue occurred with increasing study duration and to account for seasonal changes in bag characteristics. It was defined as the number of days elapsed since the first day of the study (excluding the first two weeks that were excluded from analysis). Daytime was included to control for the variation of passenger types and their baggage throughout the day. Time was therefore split into 2-h blocks and included as dummy variables, with the time block from 12:00 to 14:00 as the reference category.

$$performance = time\ on\ task + task\ load + time\ on\ task \times task\ load + days\ since\ study\ start + daytime + (1|screener/session) \quad (1)$$

$$session\ duration = mean\ session\ task\ load + days\ since\ study\ start + daytime + (1|screener) \quad (2)$$

We fitted logistic mixed models (estimated using ML with Laplace Approximation and Nelder-Mead optimizer) using the glmer function from the lme4 package (Bates et al., 2015) to analyze the binary dependent variables hit rate and reject rate. For the processing time and screening duration, a linear mixed model (estimated using REML and nloptwrap optimizer) was fitted using the lmer function of the same package. The processing time was log-transformed to normalize residuals. To assess the effect of task load on session duration, we fitted a linear mixed model that included the mean task load of the session, days since study start, and daytime as fixed effects and screener as the random effect (see Equation (2)). All metric variables (time on task, task load, log processing time, and duration) were z-transformed to ensure better model convergence. Visual inspection of residual plots using the DHARMa package (Hartig, 2022) did not reveal any obvious deviations from homoscedasticity or normality. For the logistic models, confidence intervals (95%) and p-values were computed using the Wald approximation. The aforementioned analyses focus on how performance was affected by time on task and task load considering longer screening sessions. We further compared the performance of the study and control group to test whether the prolonged screening sessions had a direct impact on performance, for example, knowing that the session will likely be longer might have preemptively affected performance at the beginning of the session. Since the data were not normally distributed, average hit rates, reject rates, and processing times were compared using the Mann–Whitney–Wilcoxon test. SSSQ data was aggregated per screener and construct (Distress, Engagement, Worry). The Mann-Whitney-Wilcoxon test was used to compare the central tendencies of each construct between the two groups.

3. Results

3.1. Descriptive data

Table 1 shows the average session duration, average number of screening sessions conducted per screener, and the average number of images and TIP images inspected per screener for both groups.

3.2. Effects on performance

The mixed model analyzing the detection performance (hit rate) of the study group showed no main effect of time on task ($b = -0.068$, $SE = 0.041$, $p = .092$); however, a significant main effect of task load ($b = -0.137$, $SE = 0.046$, $p = .003$) and a significant interaction of Time on task × Task load ($b = 0.140$, $SE = 0.041$, $p < .001$) were observed. The days since the start of the study ($b = 0.153$, $SE = 0.046$, $p < .001$) also had a significant main effect. The odds ratios and confidence intervals of the mixed models are listed in Table 2. Model statistics on random effects and variance decomposition are provided in Table 3. While the fixed effects explained 1.8% of the variance for the hit rate, 13.2% of the variance was explained by random effects; 9.9% by the screener and 3.4% by the session. For the reject rate, there was a significant main effect of time on task ($b = -0.039$, $SE = 0.006$, $p < .001$), a main effect of task load ($b = -0.049$, $SE = 0.007$, $p < .001$), and a significant interaction of Time on task × Task load ($b = -0.015$, $SE = 0.007$, $p = .022$). Furthermore, a main effect of days since study start was found ($b = 0.121$, $SE = 0.008$, $p < .001$). For processing time, there was a significant main effect of time on task ($b = -0.042$, $SE = 0.002$, $p < .001$) and a main effect of task load ($b = -0.123$, $SE = 0.005$, $p < .001$). The

Table 1
Descriptive statistics for the study and control group.

Group	n	Mean session duration per screener in min	Number of sessions per screener	Number of images per screener	Number of TIP images per screener
		M (SD)	M (SD)	M (SD)	M (SD)
SG	22	34.7 (5.68)	53.2 (36.4)	13′073 (9′621)	287 (211)
CG	19	20.8 (3.04)	63.5 (49.0)	7′836 (641)	175 (125)

Note. SG = study group, CG = control group.

Table 2
Fixed effects of the mixed models for the hit rate, reject rate, and processing time of the study group. Confidence intervals and p-values are based on the Wald approximation.

Coefficient	Hit rate			Reject rate			Processing time		
	Odds ratio	95% CI	*p*	Odds ratio	95% CI	*p*	Estimate	95% CI	*p*
Intercept	6.173	[4.454, 8.556]	**<.001**	0.128	[0.119, 0.138]	**<.001**	.000	[-0.125, 0.126]	.994
Time on task	0.934	[0.862, 1.011]	.092	0.962	[0.950, 0.974]	**<.001**	−.042	[-0.046, −0.037]	**<.001**
Task load [images/min]	0.872	[0.796, 0.955]	**.003**	0.952	[0.937, 0.966]	**<.001**	−.123	[-0.132, −0.113]	**<.001**
Days since study start	1.165	[1.065, 1.275]	**<.001**	1.129	[1.112, 1.146]	**<.001**	.108	[0.092, 0.124]	**<.001**
Time on task × Task load	0.869	[0.802, 0.942]	**<.001**	0.985	[0.972, 0.998]	**.022**	.005	[0.001, 0.010]	**.029**
Day time [04:00–06:00]	1.180	[0.910, 1.530]	.212	0.894	[0.855, 0.936]	**<.001**	−.101	[-0.150, −0.052]	**<.001**
Day time [6:00–08:00]	1.261	[0.958, 1.659]	.098	0.936	[0.893, 0.980]	**.005**	−.037	[-0.086, 0.013]	.145
Day time [8:00–10:00]	1.125	[0.790, 1.601]	.513	1.026	[0.968, 1.088]	.382	.001	[-0.055, 0.056]	.977
Day time [10:00–12:00]	1.142	[0.876, 1.491]	.326	0.997	[0.954, 1.043]	.908	.013	[-0.036, 0.062]	.592
Day time [14:00–16:00]	1.156	[0.873, 1.530]	.312	0.908	[0.865, 0.952]	**<.001**	−.085	[-0.136, −0.034]	**.001**
Day time [16:00–18:00]	1.504	[0.872, 2.593]	.142	0.906	[0.836, 0.983]	**.017**	−.106	[-0.171, −0.040]	**.002**
Day time [18:00–20:00]	0.741	[0.375, 1.466]	.390	0.806	[0.720, 0.902]	**<.001**	−.233	[-0.325, −0.142]	**<.001**

Table 3
Random effects and variance explanation of the mixed models for the hit rate, reject rate, and processing time of the study group.

	Hit rate	Reject rate	Processing time
σ^2	3.29	3.29	0.89
τ_{00}	0.13 $_{\text{Screener/Session}}$[a] 0.38 $_{\text{Screener}}$[b]	0.01 $_{\text{Screener/Session}}$[a] 0.03 $_{\text{Screener}}$[b]	0.05 $_{\text{Screener/Session}}$[a] 0.08 $_{\text{Screener}}$[b]
ICC	0.13	0.01	0.13
N	132 $_{\text{Screener/Session}}$[a] 22 $_{\text{Screener}}$	134 $_{\text{Screener/Session}}$[a] 22 $_{\text{Screener}}$	134 $_{\text{Screener/Session}}$[a] 22 $_{\text{Screener}}$
Observations	6′332	281′289	253′770
Marginal R^2/ conditional R^2	0.018/0.150	0.006/0.016	0.026/0.152

Note. σ^2 = residual variance or within-subject variance; τ_{00} = random intercept variance; ICC = intra-class correlation; marginal R^2 = variance explanation through the fixed effects; conditional R^2 = variance explanation through the fixed and random effects.
[a] Random intercept variance between sessions nested in screener.
[b] Random intercept variance between screeners.

interaction term of Time on task × Task load was also significant (b = 0.005, SE = 0.002, $p = .029$). Additionally, days since study start demonstrated a main effect (b = 0.108, SE = 0.008, $p < .001$). For all three dependent variables, likelihood-ratio tests confirmed the presence of the interaction between time on task and task load (Table 4). Fig. 1 demonstrates the marginal effects of time on task for three levels of task load (mean, one standard deviation below and above the mean) to illustrate how performance changed with time on task and task load (for readability, estimates and confidence intervals have been back-transformed to absolute rates and processing times without any bias correction). The hit rate only decreased when the task load was high; however, we found a general decrease of the reject rate and processing time with time on task.

We observed a main effect of days since study start, showing an increase in the hit rate, reject rate, and processing time over the course of the study. We calculated the same mixed models as described in Equation (1) for the control group to investigate whether this was caused by the study group adapting to longer screening durations or because of other factors (seasonal, operational). We found an increase for reject rate (b = 0.137, SE = 0.010, $p < .001$) and processing time (b = 0.086, SE = 0.008, $p < .001$) for the control group during the study. For the hit rate, no increase was found for the control group (b = −0.017, SE = 0.072, $p = .816$). Owing to large confidence intervals as shown in Fig. 2, it is unclear whether there was a substantial difference between the study and control group.

Group comparisons between the study and control groups found no difference in the average hit rate (study group: M = .855, SD = .070; control group: M = .857, SD = .070; $W = 237$, $p = .472$), reject rate (study group: M = .113, SD = .019; control group: M = .120, SD = .027; $W = 250$, $p = .293$), or processing time (study group: M = 3.6s, SD = 0.77; control group M = 3.8s, SD = 0.84; $W = 254$, $p = .248$).

3.3. Effects on session duration

To determine whether screening durations changed over the course of the study or depended on task load, we further examined the screening sessions that screeners of the study group terminated themselves. In total, the study group conducted 1′170 sessions. Among these, the screeners provided a reason for terminating the session prematurely for 436 sessions; only 129 of these were terminated by the screeners themselves for non-external reasons. Thus, the analyzed data set consisted of 129 sessions conducted by 15 different screeners. Fig. 3 depicts the distribution of the mean duration of these sessions per screener.

The mixed model analyzing the duration of self-terminated sessions showed no significant effects for task load (b = 0.174, SE = 0.091, $p = .059$) or days since the start of the study (b = 0.058, SE = 0.098, $p = .553$). Meanwhile, the fixed effects explained 6.4% of the variance for the screening duration, 24.7% of the variance was explained by random effects, and therefore, by the screener ($\sigma^2 = 0.77$, $\tau_{00\ \text{Screener}} = 0.28$, marginal $R^2 = 0.064$, conditional $R^2 = 0.311$).[1]

3.4. Subjective data

A total of 40 participants (study group: 21, control group: 19) filled in the SSSQ up to five times (M = 3.45, SD = 1.48). Fig. 4 shows the means of the constructs in the questionnaire for each group. Group comparisons found no difference for Distress (W = 261.5, $p = .095$) or Worry (W = 262, $p = .093$). However, the study group reported higher values of Engagement (W = 112, $p = .018$).

Fig. 5 depicts the results of the questionnaire on screening duration (completed by 15 participants of the study group). The distribution in Fig. 5A shows that it became difficult for screeners to continue with screening at around 30–40 min (M = 39.29, SD = 9.17). Further, the screeners stated that a screening duration of around 30 min (M = 31.79, SD = 9.92) was optimal (Fig. 5B).

4. Discussion

This study investigated the effects of time on task and task load on performance and subjective stress among X-ray baggage screeners. A

[1] σ^2 = residual variance or within-subject variance; τ_{00} = random intercept variance or between-subject variance; marginal R^2 = variance explanation through the fixed effects; conditional R^2 = variance explanation through the fixed and random effects.

Table 4
Model comparison between the models with and without the interaction Time on task × Task load for the hit rate, reject rate, and processing time of the study group.

Model	Hit rate				Reject rate				Processing time			
	AIC	R^2 c.	df	*p*	AIC	R^2 c.	df	*p*	AIC	R^2 c.	df	*p*
M0 = **Time on task** + **task load** + days since study start + daytime + (1| Screener/Session)	4′605	.146	13		193′200	.016	13		692′200	.153	14	
M1 = **Time on task** × **task load** + days since study start + daytime + (1| Screener/Session)	4′595	.150	14	**.001**	193′200	.016	14	**.023**	692′200	.152	15	**.020**

Note. R^2 c. = R^2 conditional: variance explanation through the fixed and random effects.

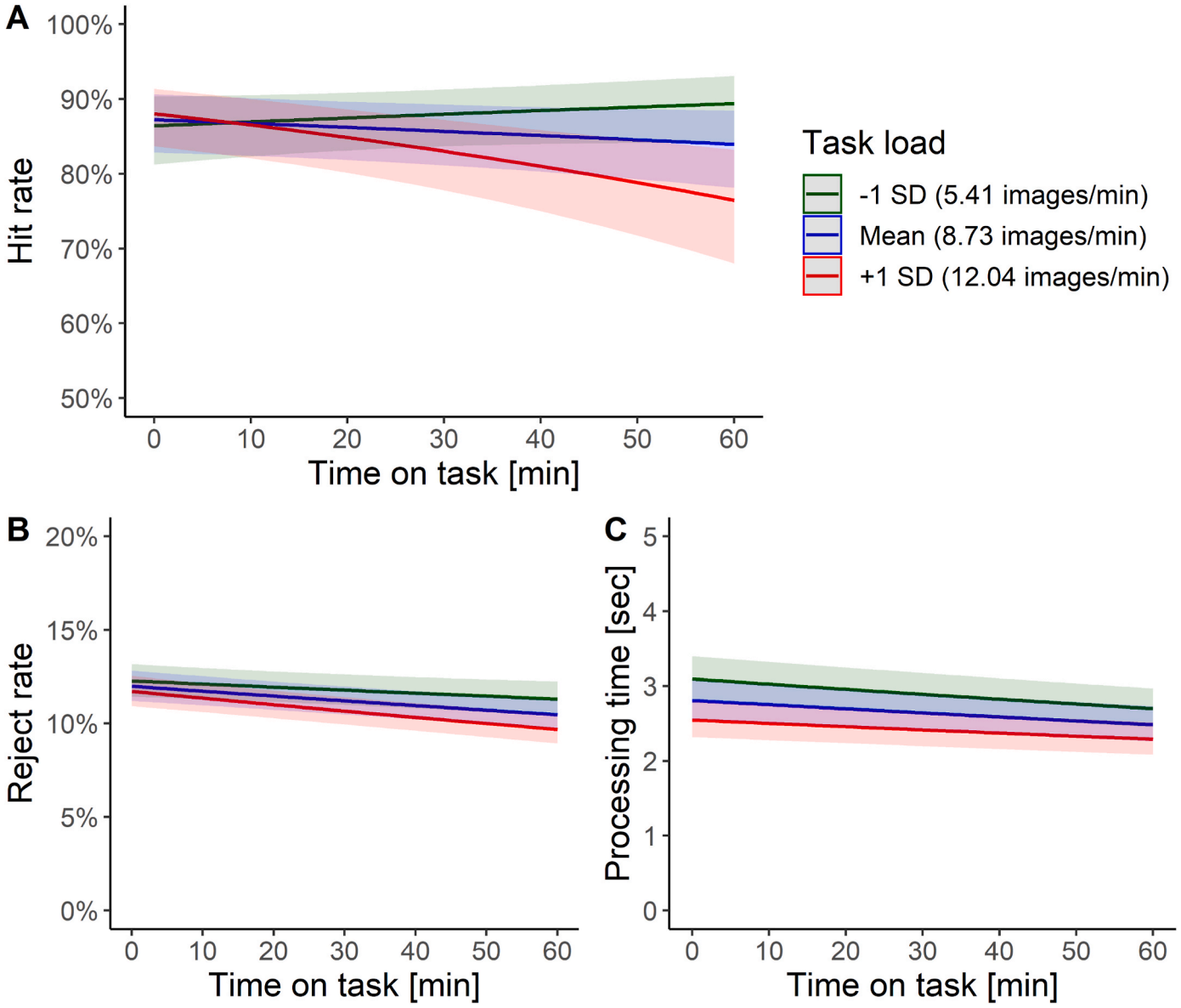


Fig. 1. Effects of time on task on hit rate (A), reject rate (B), and processing time (C) for three levels of task load: M - 1 SD, M, and M + 1 SD.

group of screeners (study group) from an international airport participated in a four-month field study during which they conducted screening sessions for up to 60 min, while a control group engaged in screening for around 20 min. Examining longer screening sessions in the study group revealed an interaction between time on task and task load (number of images inspected per min) for detection performance (hit rate). The hit rate did not decrease with time on task at low or average task load. However, when the task load was high, a decline in the hit rate was observed with time on task. A stable hit rate at low and average task load confirmed the results of X-ray baggage screening studies that found an unchanged hit rate over time (Buser et al., 2020; Wolfe et al., 2007), or did not find performance differences between different break regimes (Chavaillaz et al., 2019). Meuter and Lacherez (2016) also found an interaction between time on task and task load and a decrease in the hit rate at high task load (defined as more than the median of 5.4 images per minute in their case). For the reject rate and processing time, we found small decreases with time on task for all levels of task load; however, slightly stronger decreases were observed for the higher task load. The

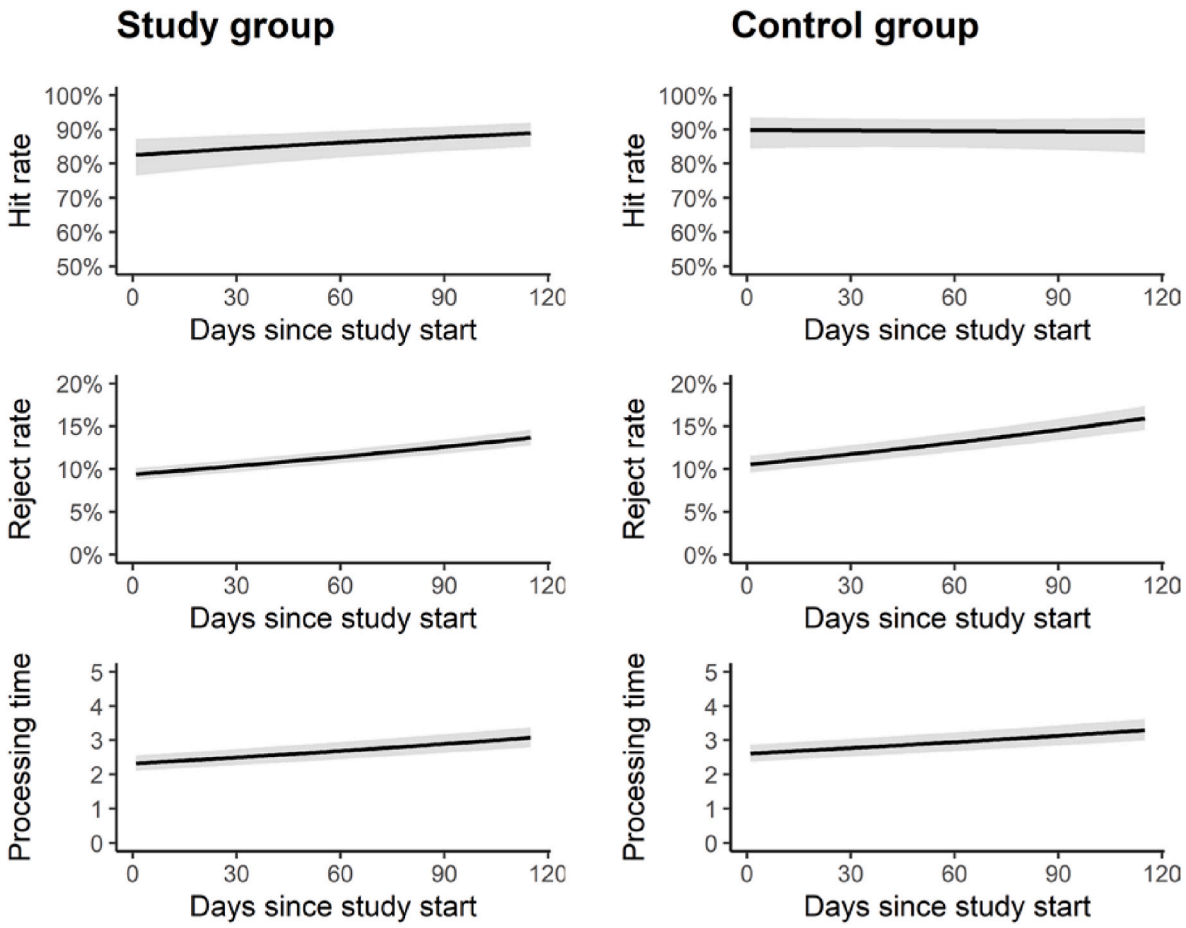


Fig. 2. Change in hit rate, reject rate, and processing time over the course of the study for the study group (left), and the control group (right).

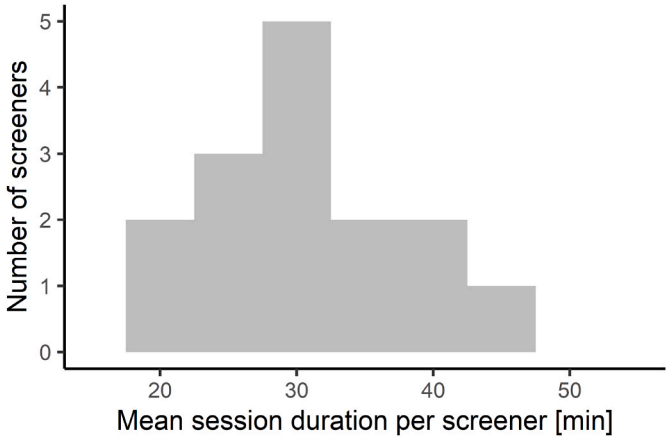


Fig. 3. Distribution of the mean screening duration per screener of the study group for sessions that were ended on the screeners' own terms.

efficiency of X-ray baggage screening therefore increased, with screeners providing faster responses and producing fewer manual bag searches.

Our finding of a declining hit rate at high but not at low or average task loads does not seem compatible with the underload theory (Robertson et al., 1997). If under-stimulation were the cause of the decline in the hit rate, a stronger decline in performance would be expected when screeners only have few images to inspect (low task load). In other words, the interaction between task load and time on task would be expected in the opposite direction. The resource theory, however, assumes that performance decreases due to the depletion of cognitive resources (Helton and Warm, 2008; Matthews et al., 2010) and one would therefore expect a stronger decline when task load is high. While we did observe a decline in the hit rate at high task load, we also found a decline in processing time and reject rate with increasing time on task and task load. The fact that screeners become faster under these conditions cannot be justified with the resource theory. Similarly, a decrease in the reject rate suggests that the false alarm rate does not increase, which cannot be explained by the resource theory. Conversely, our results are in line with Rubinstein's (2020) observation that performance decrements in visual search tasks, such as X-ray baggage screening or industrial inspection, often manifest themselves in a decrease of hit rate, false alarm rate, and response time. His proposed DART theory assumes that this change in performance is due to implicit strategic changes in the behavior to protect cognitive resources. Screeners therefore increase their speed of performing the task when spending long periods of time on the task to save resources, which leads to fewer "target present" responses (Rubinstein, 2020). In this context, one might expect the behavior change to occur more strongly when the task load is high, as we observed in our study: saving resources becomes more important as the number of images to be analyzed increases. Additionally, other studies have also found resource-saving behavior at high task loads. People tend to rely more on automation when faced with high task load (Dixon and Wickens, 2006; Wickens and Dixon, 2007), or choose heuristic search strategies when forced to prioritize speed over accuracy (McCarley, 2009).

The subjective stress measures of the study group were compared to the control group, who undertook screening for around 20 min according to the European Regulation. Unlike in other typical vigilance tasks (Warm et al., 2008b), we did not find increased levels of distress or

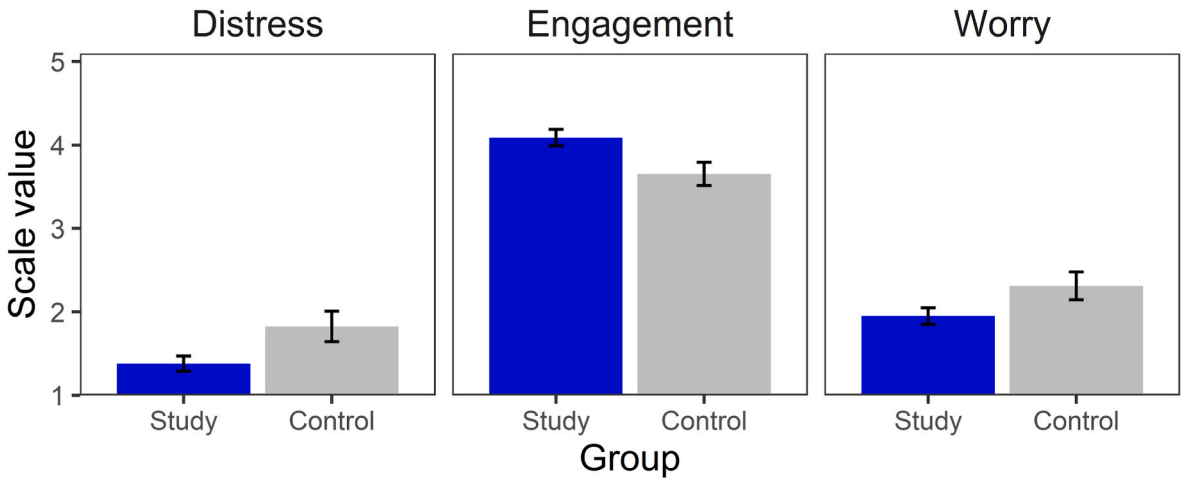


Fig. 4. Mean of Distress, Engagement, and Worry for the study and control group. Error bars represent standard errors.

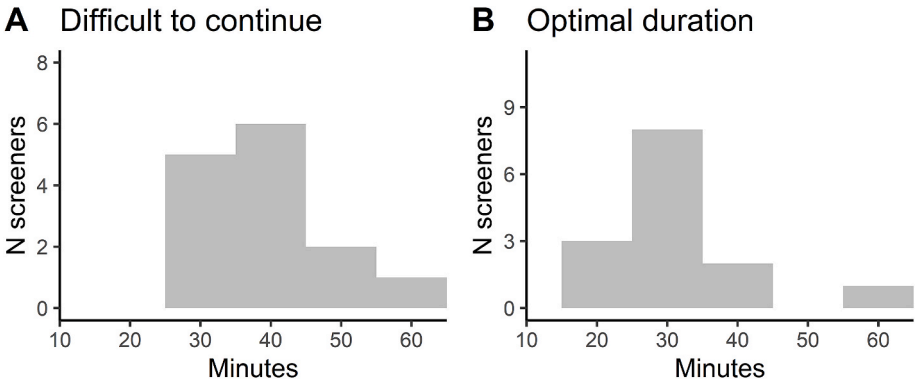


Fig. 5. Distribution of the study group's responses to the survey questions (A) "After what time did it get difficult to continue screening?", and (B) "What do you consider to be an optimal screening duration?"

decreased levels of engagement due to longer screening. The study group, who undertook screening for up to 60 min, did not report more distress or worry compared to the control group. The study group even reported higher values in engagement; this may be because the screening position allows screeners to sit separated from the checkpoint, thus contributing to recovery. Furthermore, this group was able to decide for themselves when to end a screening session. This additional autonomy could be another reason for the higher engagement as defined in the established work design theories (Bakker and Demerouti, 2007; Hackman and Oldham, 1980). Another explanation could be that the study group showed more engagement because participants were more aware about contributing to research than participants in the control group. Regarding the small changes in the analyzed performance measures that were observed over the four months of the study, it is reasonable to infer that they were due to seasonal changes in passengers and their baggage as they were also found for the control group. The session duration did not change across the study. Therefore, the results do not indicate that screeners either became accustomed to screening for longer or showed any negative impact of long-term stress from engaging in screening for longer.

Overall, the effects of time on task and task load on hit rate were relatively small compared to differences between study participants. The estimated random effects suggest that screeners contributed 5.5 times more to the variance in hit rate explained by the model than all fixed effects combined, i.e., time on task, task load, days since study start, and daytime. Previous studies showed that people differ significantly in visual cognitive abilities that are relevant for recognizing objects in X-ray images and that vigilance or working memory capacity of individuals predict visual search performance (Hardmeier and Schwaninger, 2008; Hättenschwiler et al., 2019; Mitroff et al., 2018; Peltier and Becker, 2020; Rusconi et al., 2015; Schwaninger et al., 2005). Along with performance, performed and preferred screening durations also varied considerably between the screeners. Based on the participants' average session duration and their reported preferred duration, 30–40 min of screening would be feasible for most screeners. Whereas we observed a decrease in the hit rate with time on task at high task load, preventing high task load would only have a minor impact on the overall hit rate at the studied airport, as task load was only high for a minority of the inspected images (only for 15% of the images task load was at or above the threshold defined as high in our results). This indicates that focusing on interindividual differences might be more effective than controlling the task load.

A limitation of our study is that we only investigated remote screening. Further research is needed to examine whether different results are obtained when screeners work at the more busy and noisy checkpoint (Kuhn, 2017). Moreover, it remains to be investigated whether the same or similar results are found at other airports, as they can vary regarding their size, implemented technology, task load, and other variables. Because we were only able to assess the reject rate and not the false alarm rate, we could not fully conclude whether the observed decreases in hit rate and reject rate are due to a sensitivity decrement or a change in response bias. Further, it is important to consider that screeners in our study could decide to end screening sessions. Therefore, the generalizability of the conclusions for airports with fixed screening sessions might be limited. Another limitation is that we did not address eye strain, which has been associated with prolonged

and continuous daily use of digital screens (for recent reviews, see Kaur et al., 2022; Mehra and Galor, 2020). When allowing 30–40 min of continuous screening, the recommendations of the American Optometric Association may be considered, that is, taking a 15-min break after 2 h of computer use or focusing on an object 20 feet away for 20 s after 20 min of screen use (American Optometric Association, n.d.).

5. Conclusion

This study investigated how longer screening durations affect screener performance at an international airport. For the detection of prohibited articles (hit rate), there was an interaction between time on task and task load: while detection (hit rate) decreased with an increase in the time on task when task load was high, we found no significant decrement of the hit rate when task load was low or average. Furthermore, time on task and a higher task load resulted in a lower reject rate and faster processing times. While screeners conducting longer screening durations did not report more stress, we observed individual differences in performance and in performed and preferred screening duration. Our results are in line with the DART proposed by Rubinstein (2020), which can explain decreases in the hit rate, reject rate, and processing times as a coping strategy. Accordingly, screeners switch to a resource-efficient response pattern when the task load is high, with negative consequences for hit rates. If the results of our study can be replicated in remote screening conditions with different airports, trials can be extended to the checkpoint. With similar outcomes at checkpoints, screening durations of 30–40 min could be implemented, which can provide operational benefits without, or only, small decreases in the hit rate during periods of high task load.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Swiss National Science Foundation [grant number 100019_188808] and the Swiss Federal Office of Civil Aviation. We thank Robin Riz à Porta for his assistance in data preparation.

References

American Optometric Association. Computer vision syndrome. n.d. https://www.aoa.org/healthy-eyes/eye-and-vision-conditions/computer-vision-syndrome?sso=y. (Accessed 13 February 2023).

Bakker, A.B., Demerouti, E., 2007. The job demands-resources model: state of the art. J. Manag. Psychol. 22 (3), 309–328. https://doi.org/10.1108/02683940710733115.

Basner, M., Rubinstein, J., Fomberstein, K.M., Coble, M.C., Ecker, A., Avinash, D., Dinges, D.F., 2008. Effects of night work, sleep loss and time on task on simulated threat detection performance. Sleep 31 (9), 1251–1259. https://doi.org/10.5665/sleep/31.9.1251.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Biggs, A.T., Kramer, M.R., Mitroff, S.R., 2018. Using cognitive psychology research to inform professional visual search operations. J. Appl. Res. Mem. Cogn. 7 (2), 189–198. https://doi.org/10.1016/j.jarmac.2018.04.001.

Biggs, A.T., Mitroff, S.R., 2015. Improving the efficacy of security screening tasks: a review of visual search challenges and ways to mitigate their adverse effects. Appl. Cognit. Psychol. 29 (1), 142–148. https://doi.org/10.1002/acp.3083.

Buser, D., Sterchi, Y., Schwaninger, A., 2020. Why stop after 20 minutes? Breaks and target prevalence in a 60-minute X-ray baggage screening task. Int. J. Ind. Ergon. 76, 102897 https://doi.org/10.1016/j.ergon.2019.102897.

Chavaillaz, A., Schwaninger, A., Michel, S., Sauer, J., 2019. Work design for airport security officers: effects of rest break schedules and adaptable automation. Appl. Ergon. 79, 66–75. https://doi.org/10.1016/j.apergo.2019.04.004.

Claypoole, V.L., Dever, D.A., Denues, K.L., Szalma, J.L., 2019. The effects of event rate on a cognitive vigilance task. Hum. Factors 61 (3), 440–450. https://doi.org/10.1177/0018720818790840.

Cutler, V., Paddock, S., 2009. Use of threat image projection (TIP) to enhance security performance. In: Proceedings of the 43rd IEEE International Carnahan Conference on Secur. Technol., Zurich, 5-8 October, pp. 46–51. https://doi.org/10.1109/CCST.2009.5335565.

Davies, D., Parasuraman, R., 1982. The Psychology of Vigilance. Academic Press, London.

Dixon, S.R., Wickens, C.D., 2006. Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. Hum. Factors 48 (3), 474–486. https://doi.org/10.1518/001872006778606822.

Donnelly, N., Muhl-Richardson, A., Godwin, H.J., Cave, K.R., 2019. Using eye movements to understand how security screeners search for threats in X-ray baggage. Vision 3 (2). https://doi.org/10.3390/vision3020024, 24.

Drury, C.G., Watson, J., 2002. Good practices in visual inspection. Human factors Aviat. maintenance-phase nine, progress report. FAA/Human Factors Aviation Maintenance 1–90. http://www.dviaviation.com/files/45146949.pdf. (Accessed 13 February 2023).

European Commission, 2015. Commission implementing regulation (EU) 2015/1998 of 5 November 2015 laying down detailed measures for the implementation of the common basic standards on aviation security (Text with EEA relevance). Off. J. Eur. Union. http://data.europa.eu/eli/reg_impl/2015/1998/oj. (Accessed 13 February 2023).

Ghylin, K.M., Drury, C.G., Batta, R., Lin, L., 2007. Temporal effects in a security inspection task: breakdown of performance components. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 51 (2), 93–97. https://doi.org/10.1177/154193120705100209.

Godwin, H.J., Menneer, T., Cave, K.R., Donnelly, N., 2010a. Dual-target search for high and low prevalence X-ray threat targets. Vis. cogn. 18 (10), 1439–1463. https://doi.org/10.1080/13506285.2010.500605.

Godwin, H.J., Menneer, T., Cave, K.R., Helman, S., Way, R.L., Donnelly, N., 2010b. The impact of relative prevalence on dual-target search for threat items from airport X-ray screening. Acta Psychol. 134 (1), 79–84. https://doi.org/10.1016/j.actpsy.2009.12.009.

Hackman, J.R., Oldham, R.G., 1980. Work Redesign. Addison-Wesley, MA.

Hardmeier, D., Schwaninger, A., 2008. Visual cognition abilities in X-Ray screening. In: Proceedings of the 3rd International Conference on Research in Air Transportation. ICRAT, pp. 311–316. https://doi.org/10.13140/RG.2.1.4335.7924.

Hartig, F., 2022. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R packag (version 0.4.5). http://florianhartig.github.io/DHARMa/.

Hättenschwiler, N., Merks, S., Sterchi, Y., Schwaninger, A., 2019. Traditional visual search vs. X-ray image inspection in students and professionals: are the same visual-cognitive abilities needed? Front. Psychol. 10, 1–17. https://doi.org/10.3389/fpsyg.2019.00525.

Helton, W.S., 2004. Validation of a short stress state questionnaire. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 48 (11) https://doi.org/10.1177/154193120404801107.

Helton, W.S., Warm, J.S., 2008. Signal salience and the mindlessness theory of vigilance. Acta Psychol. 129 (1), 18–25. https://doi.org/10.1016/j.actpsy.2008.04.002.

Hofer, F., Schwaninger, A., 2005. Using threat image projection data for assessing individual screener performance. WIT Trans. Built Environ. 82, 417–426. https://doi.org/10.2495/SAFE050411.

Jerison, H.J., 1963. On the decrement function in human vigilance. In: Buckner, D.N., McGrath, J.J. (Eds.), Vigilance: A Symposium. McGraw-Hill, New York, pp. 199–212.

Kaur, K., Gurnani, B., Nayak, S., Deori, N., Kaur, S., Jethani, J., Singh, D., Agarkar, S., Hussaindeen, J.R., Sukhija, J., Mishra, D., 2022. Digital eye strain- a comprehensive review. Ophthalmol. Ther. 11, 1655–1680. https://doi.org/10.1007/s40123-022-00540-9.

Koller, S.M., Drury, C.G., Schwaninger, A., 2009. Change of search time and non-search time in X-ray baggage screening due to training. Ergonomics 52 (6), 644–656. https://doi.org/10.1080/00140130802526935.

Kuhn, M., 2017. Centralised image processing: the impact on security checkpoints. Aviat Secur. Int. 23 (5), 28–30.

Mackworth, N.H., 1948. The breakdown of vigilance durning prolonged visual search. Q. J. Exp. Psychol. 1 (1), 6–21. https://doi.org/10.1080/17470214808416738.

MacLean, K.A., Ferrer, E., Aichele, S.R., Bridwell, D.A., Zanesco, A.P., Jacobs, T.L., King, B.G., Rosenberg, E.L., Sahdra, B.K., Shaver, P.R., Wallace, B.A., Mangun, G.R., Saron, C.D., 2010. Intensive meditation training improves perceptual discrimination and sustained attention. Psychol. Sci. 21 (6), 829–839. https://doi.org/10.1177/0956797610371339.

Matthews, G., Warm, J.S., Reinerman-Jones, L.E., Langheim, L.K., Washburn, D.A., Tripp, L., 2010. Task engagement, cerebral blood flow velocity, and aiagnostic monitoring for sustained attention. J. Exp. Psychol. Appl. 16 (2), 187–203. https://doi.org/10.1037/a0019572.

McCarley, J.S., 2009. Effects of speed - accuracy instructions on oculomotor scanning and target recognition in a simulated baggage X-ray screening task. Ergonomics 52 (3), 325–333. https://doi.org/10.1080/00140130802376059.

Mehra, D., Galor, A., 2020. Digital screen use and dry eye: a review. Asia-Pacific J. Ophthalmol. 9 (6), 491–497. https://doi.org/10.1097/APO.0000000000000328.

Meuter, R.F.I., Lacherez, P.F., 2016. When and why threats go undetected: impacts of event rate and shift length on threat detection accuracy during airport baggage screening. Hum. Factors 58, 218–228. https://doi.org/10.1177/0018720815616306.

Mitroff, S.R., Ericson, J.M., Sharpe, B., 2018. Predicting airport screening officers ' visual search competency with a rapid assessment. Hum. Factors 60 (2), 201–211. https://doi.org/10.1177/0018720817743886.

Neigel, A.R., Claypoole, V.L., Smith, S.L., Waldfogle, G.E., Fraulini, N.W., Hancock, G.M., Helton, W.S., Szalma, J.L., 2020. Engaging the human operator: a review of the

theoretical support for the vigilance decrement and a discussion of practical applications. Theor. Issues Ergon. Sci. 21 (2), 239–258. https://doi.org/10.1080/1463922X.2019.1682712.

Nuechterlein, K.H., Parasuraman, R., Jiang, Q., 1983. Visual sustained attention: image degradation produces rapid sensitivity decrement overtime. Science 220, 327–329. https://doi.org/10.1126/science.6836276.

Peltier, C., Becker, M.W., 2020. Individual differences predict low prevalence visual search performance and sources of errors: an eye-tracking study. J. Exp. Psychol. Appl. 26, 646–658. https://doi.org/10.1037/xap0000273.

R Core Team, 2020. R: a language and environment for statistical computing. https://www.r-project.org/.

Robertson, I.H., Manly, T., Andrade, J., Baddeley, B.T., Yiend, J., 1997. 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia 35 (6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8.

Rubinstein, J.S., 2020. Divergent response-time patterns in vigilance decrement tasks. J. Exp. Psychol. Hum. Percept. Perform. 46 (10), 1058–1076. https://doi.org/10.1037/xhp0000813.

Rusconi, E., Ferri, F., Viding, E., Mitchener-Nissen, T., 2015. XRIndex: a brief screening tool for individual differences in security threat detection in X-ray images. Front. Hum. Neurosci. 9 https://doi.org/10.3389/fnhum.2015.00439.

Schwaninger, A., Hardmeier, D., Hofer, F., 2005. Aviation security screeners visual abilities & visual knowledge measurement. IEEE Aerosp. Electron. Syst. 20, 29–35.

See, J.E., 2012. Visual inspection : a review of the literature. Albuquerque, NM, and Livermore, CA (United States). https://doi.org/10.2172/1055636.

See, J.E., Howe, S.R., Warm, J.S., Dember, W.N., 1995. Meta-analysis of the sensitivity decrement in vigilance. Psychol. Bull. 117 (2), 230–249. https://doi.org/10.1037/0033-2909.117.2.230.

Skorupski, J., Uchroński, P., 2016. A human being as a part of the security control system at the airport. Procedia Eng. 134, 291–300. https://doi.org/10.1016/j.proeng.2016.01.010.

Teichner, W.H., 1974. The detection of a simple visual signal as a function of time of watch. Hum. Factors 16 (4), 339–352. https://doi.org/10.1177/001872087401600402.

Teo, G., Szalma, J.L., 2011. The effects of task type and source complexity on vigilance performance, workload, and stress. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 55 (1), 1180–1184. https://doi.org/10.1177/1071181311551246.

Tiwari, T., Singh, A.L., Singh, I.L., 2009. Task demand and workload: effects on vigilance performance and stress. J. Indian Acad. Appl. Psychol. 35 (2), 265–275.

Warm, J.S., Matthews, G., Finomore, V.S., 2008a. Workload and stress in sustained attention. In: Hancock, P.A., Szalma, J.L. (Eds.), Performance under Stress. CRC Press, London, pp. 115–141. https://doi.org/10.1201/9781315599946.

Warm, J.S., Parasuraman, R., Matthews, G., 2008b. Vigilance requires hard mental work and is stressful. Hum. Factors 50 (3), 433–441. https://doi.org/10.1518/001872008X312152.

Wickens, C.D., Dixon, S.R., 2007. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. 8 (3), 201–212. https://doi.org/10.1080/14639220500370105.

Wolfe, J.M., Horowitz, T.S., Kenner, N.M., 2005. Rare items often missed in visual searches. Nature 435, 439–440. https://doi.org/10.1038/435439a.

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. J. Exp. Psychol. Gen. 136 (4), 623–638. https://doi.org/10.1037/0096-3445.136.4.623.