Universität Basel

Fakultät für Psychologie

Measuring and Maintaining Performance in X-ray Baggage Inspection at Security Checkpoints: Methodological and Practical Considerations

Inauguraldissertation zur Erlangung der Würde einer Doktorin der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Daniela Buser

aus Riehen

Basel, 2023

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Klaus Opwis (Erstgutachter)

Prof. Dr. Adrian Schwaninger (Zweitgutachter)

Datum des Doktoratsexamen:

Dekan:in der Fakultät für Psychologie

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

1. Buser, D., Sterchi, Y., & Schwaninger, A. (2020). Why stop after 20 min? Breaks and target prevalence in a 60-min X-ray baggage screening task. *International Journal of Industrial Ergonomics*, *76*, 102897. https://doi.org/10.1016/j.ergon.2019.102897

2. Buser, D., Schwaninger, A., Sauer, J., & Sterchi, Y. (2023). Time on task and task load in visual inspection: a four-month field study with X-ray baggage screeners. *Applied Ergonomics*, *111*, 103995. https://doi.org/10.1016/j.apergo.2023.103995

3. Buser, D., Schwaninger, A., Rehor, V., & Sterchi, Y. (under review). Reliability and validity of threat image projection data on X-ray baggage screening.

Spezifizierung des eigenen Forschungsbeitrags zu den Manuskripten:

1. Eigener Beitrag nach CRediT[1]:

☐ Conceptualization	☒ Data curation	☒ Formal Analysis
☐ Funding acquisition	☒ Investigation	☐ Methodology
☒ Project administration	☒ Resources	☐ Software
☐ Supervision	☐ Validation	☒ Visualization
☒ Writing – original draft		
☐ Writing – review & editing		

Das Manuskript wurde bereits für eine andere Qualifikationsarbeit von Yanik Sterchi für seine Dissertation an der Universität Basel eingereicht.

2. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization	☒ Data curation	☒ Formal Analysis
☐ Funding acquisition	☒ Investigation	☒ Methodology
☒ Project administration	☒ Resources	☐ Software
☒ Supervision	☐ Validation	☒ Visualization
☒ Writing – original draft		
☐ Writing – review & editing		

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

3. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization	☒ Data curation	☐ Formal Analysis
☒ Funding acquisition	☐ Investigation	☒ Methodology
☒ Project administration	☒ Resources	☐ Software
☐ Supervision	☒ Validation	☒ Visualization
☒ Writing – original draft		
☐ Writing – review & editing		

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

[1] https://casrai.org/credit/

Universität Basel

Fakultät für Psychologie

Open-Science Aspekte der Manuskripte:

1. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☐ ja ☒ nein
   Ort/URL der Daten und Analysen: ……..

2. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☐ ja ☒ nein
   Ort/URL der Daten und Analysen: ……..

3. Preregistration: ☐ ja ☒ nein
   Open-Access-Publikation: ☒ ja ☐ nein
   Open-Access-Data/Analyse: ☐ ja ☒ nein
   Ort/URL der Daten und Analysen: ……..

Ort, Datum Basel, 01.08.2023

Signatur

Vorname Nachname Daniela Buser

Contents

Abstract

Inspecting X-ray images of passenger baggage for prohibited items at security checkpoints is crucial to ensuring aviation security. To prevent performance declines during inspection, the EU allows screeners to perform this task for only 20 min, although little is known about how this performance actually evolves over time. For many airports, longer screening durations would be practical, and this raises the question of the ideal screening duration in terms of both performance and screener well-being. To measure screeners' performance, airports typically implement threat image projection (TIP). TIP projects fictional threat items (FTIs) onto the X-ray images of passenger baggage; and by recording the screeners' decisions, it allows measurement of their detection rate. To draw meaningful conclusions from these data, it is essential for them to be reliable and valid. However, their reliability and validity are still poorly researched and not confirmed. This thesis addresses the question of how time on task affects performance in X-ray image inspection of cabin baggage, and it asks whether TIP performance data collected at airports provide a reliable and valid measure of operational threat detection.

Manuscript 1 investigated how performance evolves with time on task in two groups of screeners who performed a 1-hr X-ray image inspection task in the laboratory. One group took 10-min breaks every 20 min; the other group screened continuously without breaks. To assess the validity of measurements of detection performance, we varied target prevalence. Results confirmed the typical target prevalence effect and showed that d_a is a valid measure of detection performance for X-ray images inspection. Manuscript 1 provides evidence that screeners were able to maintain performance for a full hour, and that breaks had no effect on performance. However, time on task caused a shift in response tendency and might cause more distress. In Manuscript 2, we investigated the effects of time on task on performance under real working conditions by analyzing performance data from a 4-month field study. A group of screeners at a European airport were asked to analyze X-ray images from a remote screening room for up to 60 min. Only when task load was high (number of images analyzed per min), did the screeners' hit rate decrease with time on task. The efficiency, in terms of the reject rate and processing time, increased with time on task. Screeners who conducted longer screening durations did not report more distress. Yet, there were marked individual differences in performance, in performed screening durations, and in preferred screening durations. In Manuscript 3, we examined the reliability and validity of TIP performance by analyzing a large data set from a European airport. We showed that TIP data can be a reliable and valid measure of operational threat detection, and that around 100 TIP events per screener should be considered to attain minimum reliability values of 0.7. The manuscript further provides recommendations on how to increase the reliability of TIP data.

Taken together, these findings show that TIP data, which are in frequent use, can provide a reliable and valid measure of operational threat detection and that screeners can maintain performance for more than 20 min. Manuscripts 1 and 2 provide evidence that time on task in X-ray image inspection leads to a shift in response tendency rather than a decline in sensitivity. Based on performance and survey results, screening sessions could be designed more flexibly and an extension to 30–40 min could be considered. The manuscripts provide meaningful theoretical insights into performance in X-ray image inspection, especially with regard to the effect of time on task. They further provide methodological and practical contributions on appropriate detection performance measures, on how to measure performance reliably and validly, and on the design of screening durations.

Introduction

Airport security checkpoints are an essential component of ensuring aviation security. Security personnel (screeners) check passengers and their belongings to prevent them from bringing potential security threats onto the plane. This includes the inspection of X-ray images of cabin baggage in which screeners search for prohibited items among harmless daily objects. For each image, screeners decide whether it contains a prohibited item and whether it needs to be further examined at a second search station. This task entails visual search and decision making (Koller et al., 2009; McCarley et al., 2004; Wales et al., 2009) consistent with Spitz and Drury's (1978) two-component model. However, compared to traditional search or inspection tasks (Treisman & Gelade, 1980), X-ray image inspection is more complex and cognitively demanding (see, for reviews, Biggs et al., 2014, 2018) and requires different visual cognitive abilities (Hättenschwiler et al., 2019). One difference lies in the complexity of the targets and distractors (Wolfe et al., 2013). The list of prohibited items screeners must search for is very long and can change as new threats emerge. Moreover, whereas the main categories of prohibited items are clear (guns, knives, improvised explosive devices [IEDs]), the shape, size, or even material of these threats can vary widely. Moreover, targets have to be found among a large variety of distractors that can superimpose the target (Schwaninger et al., 2005). Also, a baggage can contain multiple prohibited items that need to be identified (Menneer et al., 2009).

Events such as the 9/11 attacks show how important the task is and what fatal consequences a mistake can have. Subsequently, there has been increasing research on X-ray image inspection and the factors that contribute to good detection performance (see Biggs et al. 2018 for a recent review). In addition, to minimize errors, the work of security officers is subject to strict regulations (Bassetti, 2021; Walter et al., 2021). For instance, to prevent performance declines during inspection, the continuous screening of X-ray images is legally limited to 20 min at European checkpoints (European Commission, 2015). After this, screeners must take a 10-min break or rotate to another position at the checkpoint where they perform other tasks such as instructing passengers how to prepare their baggage, checking passengers at the walk-through metal detector or person scanner, or manually checking baggage at secondary search (Michel et al., 2014). However, the validity of this regulation needs to be examined for two main reasons: First, this restriction is probably based on evidence from traditional vigilance studies conducted in the laboratory (personal communication with airport security expert, 2019), and not on results from X-ray cabin baggage screening (CBS). Second, airports are interested in extending screening durations due to an emerging technology known as Remote Cabin Baggage Screening (RCBS; Buser & Merks, 2020; Kuhn, 2017; Wetter, 2013). With RCBS, screeners analyze X-ray images in remote rooms that are separate from the checkpoint. Although this

allows for a quieter work environment, time is lost through the rotation to and from the checkpoint. Extending screening durations could alleviate this issue.

Fortunately, passengers rarely carry real threats such as bombs, guns, or knives with them. Yet, this makes it even more difficult for screeners to detect such threats. Research shows that the relative frequency with which targets appear, referred to as target prevalence, affects detection performance (Wolfe et al., 2005, 2007). The rarer targets are, the less likely people are to detect them and vice versa. Airports counteract this by implementing a technology called threat image projection (TIP) with which they project prerecorded images of fictional threat items (FTIs) onto about 1–4% of the X-ray images of real passenger baggage (Cutler & Paddock, 2009; Hofer & Schwaninger, 2005; Meuter & Lacherez, 2016; Skorupski & Uchroński, 2018). By increasing the number of threats to be detected (target prevalence), they raise the chances of their detection, while also increasing screener motivation and attention (Cutler & Paddock, 2009; Schwaninger, 2006). Moreover, TIP systems record screener responses to each TIP event, allowing calculation of on-the-job performance. TIP data are frequently used by airports and security companies for quality control purposes, but they are also used by researchers (Buser et al., 2023; Meuter & Lacherez, 2016; Skorupski & Uchroński, 2016). Despite their frequent use, the reliability and validity of TIP data have hardly been investigated, and their predictive power correspondingly remains unclear.

In line with the above, this thesis addresses two main topics: First, it evaluates the effects of time on task on performance in X-ray image inspection of cabin baggage. Second, it investigates the reliability and validity of TIP data. The following research questions are addressed: (a) How does time on task affect performance in X-ray image inspection in the laboratory? (b) Can the findings from the laboratory be found in the field? (c) Is TIP data a reliable and valid measure of operational threat detection? The thesis is structured as follows: It starts by providing a theoretical background addressing those concepts that are relevant to the three manuscripts. The theoretical background is followed by detailed summaries of the manuscripts. The framework closes with a general discussion of the publications and the conclusions that can be drawn from them by highlighting their relevance for X-ray image inspection. The three manuscripts are attached in the appendix.

Theoretical Background

Measuring performance in X-ray image inspection

In X-ray image inspection, commonly used measures are the hit rate, which refers to the percentage of detected prohibited items, and the false alarm rate, which refers to the percentage of harmless baggage falsely declared as containing a prohibited item. However, these measures depend on a person's response tendency, and this can be affected by factors such as the relative target frequency, costs associated with responses, or confidence in decision making (Macmillan & Creelman, 2005). Thus, when someone's response tendency changes, they may, for example, become more inclined to respond that a target is present, so that both their hit and false alarm rate will increase. Therefore, people with the same detection abilities may have different hit and false alarm rates due to individual response tendencies. To assess detection performance, it is therefore recommended to use measures that are independent of response tendency (Macmillan & Creelman, 2005). Signal detection theory (SDT; Green & Swets, 1966) provides a framework that distinguishes between a person's ability to detect a signal, called sensitivity, and their response tendency also referred to as criterion *c*. It claims that this sensitivity is unaffected by response tendency. Commonly used measures of sensitivity are *d'* and *A'*, which are derived from the hit rate and false alarm rate (Green & Swets, 1966; Pollack & Norman, 1964). However, for X-ray image inspection, several studies have cast doubt on the validity of these measures (Godwin, Menneer, Cave, & Donnelly, 2010; Hofer & Schwaninger, 2004; Sterchi et al., 2019; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010). To understand why this is the case, it is necessary to look at the fundamental assumptions of SDT (see, for more comprehensive discussions, Green & Swets, 1966; Macmillan & Creelman, 2005; T. D. Wickens, 2001).

SDT originated in psychophysics and decision psychology in which it was used to analyze decision making in uncertain situations, and it can be applied whenever two types of stimuli need to be distinguished (Stanislaw, 1999). In X-ray image inspection, screeners have to discriminate prohibited items (signal) from daily, harmless objects (noise). SDT assumes that our cognitive processes generate subjective evidence for or against a target's presence, and this is called the decision variable (see x-axis in Figures 1A and 1C). SDT also posits that a decision is made by establishing a threshold to the decision variable, called the criterion. Depending on where the criterion is positioned on the decision variable, a person is more or less likely to indicate the presence of a target. A liberal criterion indicates that one is more likely to respond that a target is present, whereas a conservative criterion indicates that one is less likely to declare that a target is present. As shown in Figure 1, both target-absent (noise) and target-present (signal-plus-noise) trials result in distributions of the decision variable due to the inherent noise of the process. The commonly used SDT model (Pastore et al.,

2003) assumes that these distributions follow a normal distribution with equal variance (Figure 1A), and this forms the basis for calculating detection performance measure *d'*. However, drawing on a series of studies, Wolfe et al. (2007) argue that visual search in X-ray images does not meet the assumptions underlying *d'*—namely, the requirement of equal-variance distributions for signal and signal-plus-noise (Macmillan & Creelman, 2005). For X-ray image inspection, d_a, as proposed by Simpson and Fitter (1973), seems to be a more valid measure of sensitivity, because it allows us to model unequal variance distribution of signal and signal-plus-noise (Figure 1B). For X-ray image inspection, recent studies have found a slope of 0.6 to be applicable for d_a (Godwin, Menneer, Cave, & Donnelly, 2010; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010). Figure 1B shows the slope of 1 for *d'*; Figure 1D, the slope for d_a. The slope value of 0.6 for d_a indicates that the noise distribution has a smaller standard deviation than the signal-plus-noise distribution (Sterchi et al., 2019). Manuscript 1 examines the valid measure of detection performance in X-ray baggage inspection by varying the target prevalence and computing both *d'* and d_a.

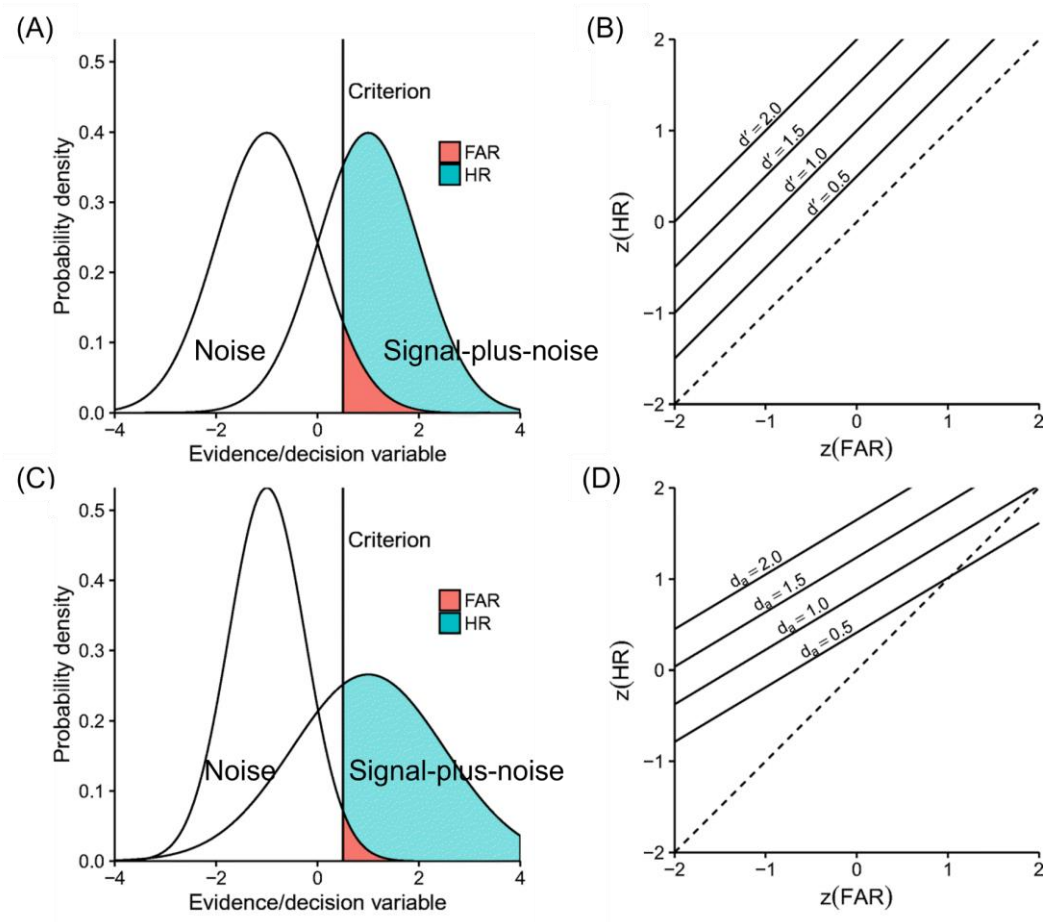


Figure 1: Illustration of signal-plus-noise and noise distribution of d' (A) and d_a (C) and corresponding slopes (B & D). FAR = False alarm rate, HR = hit rate. Graph adapted from Sterchi et al. (2019).

At airports, performance is often measured with TIP. TIP systems record screener responses to each TIP event so that the TIP hit rate can be calculated for each individual screener in a given time period (Hofer & Schwaninger, 2005). Hence, screener performance can be monitored, and action can be taken if a screener shows a low detection performance (Bassetti, 2021; Riz à Porta et al., 2022). To draw meaningful conclusions from TIP data, it is essential for them to be reliable and valid. To date, however, there are no conclusive results regarding the data's reliability. Hofer and Schwaninger (2005) analyzed up to 7 months of TIP data and found that the reliability values for CBS data were insufficient (below .58) and depended on how the data were aggregated. Moreover, the validity of TIP data has not been investigated. Thus, it is not clear whether TIP truly measures threat detection or possibly other abilities. Riz à Porta et al. (2022) found that one-third of all TIP images look unrealistic, making it possible for screeners to recognize TIP images based on artifact detection. Because TIP aims to measure how well screeners detect prohibited items in passengers' baggage, its validity can be examined by assessing how well screeners detect real, physical threat items at the checkpoint. To test and train screeners' ability to detect real threats, many airports conduct covert tests in which instructed individuals attempt to smuggle real threats past the checkpoint (Walter et al., 2021; Wetter et al., 2008). The result of each covert test is recorded, and by relating these data to TIP performance, the validity of TIP can be evaluated. Analyzing a large TIP dataset from a European airport, Manuscript 3 investigates the reliability of TIP by calculating the split-half reliability and covert test data by testing whether TIP performance can predict covert test performance.

Effects of time on task on detection performance

As a precautionary measure to prevent performance declines, the EU limits the continuous inspection of X-ray images to 20 min (European Commission, 2015). Because evidence on the effects of time on task in X-ray image inspection is scarce, this 20-min limitation is probably based on findings from vigilance research (personal communication with airport security expert, 2019) in which the effect of time on task on performance has been investigated extensively. Vigilance refers to a state of heightened attention and alertness characterized by sustained observation of stimuli in order to identify specific targets (Davies & Parasuraman, 1982; Mackworth, 1948; Warm, 1984). It requires the ability to maintain a high level of concentration over an extended period of time in what are often monotonous situations while being prepared to respond promptly and accurately to infrequent but relevant cues. In vigilance tasks, a decline in performance is often found after 15 to 20 min (Mackworth, 1948; Teichner, 1974; Warm, 1984), and typically manifests as a decrease in hits, an increase in false alarms, and increasing reaction times. This pattern of performance decline can therefore be attributed to a decline in the searchers' perceptual ability—that is, the ability to

differentiate between targets and nontargets (e.g., sensitivity in SDT; See et al., 1995). In addition to performance declines, participants often report a decrease in engagement and an increase in distress and subjective workload after such a task (Claypoole et al., 2019; Teo & Szalma, 2011; Tiwari et al., 2009; Warm, Parasuraman & Matthews, 2008). There are two main theories explaining why performance in vigilance tasks decreases with time on task (Helton & Warm, 2008; MacLean et al., 2010; Neigel et al., 2020). Underload theory suggests that the monotony of vigilance tasks (searching for rare targets over an extended period of time) causes attention to fade, resulting in targets going undetected (Robertson et al., 1997). There is more support, in contrast, for resource theory in which it is assumed that our cognitive resources are limited, and that these are depleted with increasing time on task or task difficulty, with the result that performance declines (Helton & Warm, 2008; Matthews et al., 2010).

As a vigilance task, X-ray image inspection of cabin baggage is characterized by long search periods in which the searcher is required to remain attentive although only few targets appear. Whereas there are similarities between the tasks, there are also significant differences (Drury & Watson, 2002; Wolfe et al., 2007). In traditional vigilance tasks, participants typically monitor a screen and must detect simple and single signals (Davies & Parasuraman, 1982). Hence, a short distraction can lead to missing a signal (Wolfe et al. 2007). X-ray image inspection involves searching for multiple, visually complex targets among many distractors (Schwaninger et al., 2005); and for each image, screeners have to actively declare whether or not a target is present (Koller et al., 2009). Additionally, whereas certain types of targets are very rare (e.g., bombs) in X-ray image inspection, other targets occur rather frequently in cabin baggage (e.g., liquids and gels). Because of these differences, it is unclear how well the results of time on task in vigilance tasks translate to X-ray image inspection of cabin baggage. Furthermore, most vigilance studies have been conducted in the laboratory, and it is unclear how well these results transfer to the real world in which tasks and environments tend to be more complex (Drury & Watson, 2002).

Only a few studies have investigated how professional screeners' performance evolves with time on task in X-ray image inspection. A study by Ghylin et al. (2007) examined different performance measures across four 1-hr time blocks in a laboratory study. The researchers observed a decrease in the hit rate, false alarm rate, and reaction times between the first and fourth hour. However, they did not observe a decline in the sensitivity measure A', suggesting that screeners shifted their response tendency. Unfortunately, this study does not indicate how performance changed within the 1-hr blocks. To date, only one study has investigated performance in professional screeners in the field. Meuter and Lacherez (2016) analyzed TIP data from an international airport taken from screening durations of up to 30 min. The researchers found a decrease in the TIP hit rate of 2

percentage points with time on task, but only when workload was high (operationalized as more than 5.4 X-ray images analyzed per min). In their study, workload was calculated using a median split across all screening sessions that categorized workload as low or high. The TIP hit rate refers to the proportion of projected fictional threat items detected by screeners. Due to technical limitations, the researchers were unable to measure the false alarm rate. Therefore, they could not distinguish whether the decline in hit rate was due to decreasing sensitivity or to a shift in response tendency.

Time on task appears to elicit a somewhat different vigilance pattern in X-ray image inspection compared to traditional vigilance tasks (Ghylin et al., 2007; Rubinstein, 2020). This cannot be explained by resource (Helton & Warm, 2008; Matthews et al., 2010) or underload theory (Robertson et al., 1997). In X-ray image inspection, the hit rate and false alarm rate both decline with time on task, while people also become faster at responding. This pattern is better explained by a change in the response tendency rather than a decline in sensitivity. To account for this alternative vigilance pattern in X-ray image inspection tasks, Rubinstein (2020) proposed dynamic-allocation resource theory (DART). This posits that vigilance decrements in X-ray image inspection are caused by active changes in response tendency rather than by a decline in sensitivity due to limited resources or under-stimulation. To better understand how performance evolves with time on task in X-ray baggage inspection, the studies in Manuscript 1 and 2 investigated different performance measures with professional screeners in screening durations up to one hour. Because vigilance studies have associated time on task with increased distress and less engagement (Claypoole et al., 2019; Teo & Szalma, 2011; Tiwari et al., 2009; Warm, Parasuraman & Matthews, 2008), we also assessed these constructs with the Short Stress State Questionnaire (SSSQ; Helton, 2004).

Summary of the Manuscripts

Manuscript 1: Why stop after 20 minutes? Breaks and target prevalence in a 60-minute X-ray baggage screening task

Motivation and aim of the study. Because there is an interest in extending the duration of X-ray image inspection, which is currently limited to 20 min at security checkpoints, more evidence is needed on how performance on this task evolves over time. The few studies to date suggest that with time on task, a shift in response tendency occurs rather than a decrease in sensitivity (Basner et al., 2008; Ghylin et al., 2007; Rubinstein, 2020). Analyzing TIP data, Meuter and Lacherez (2016) found a decrease in the hit rate over time for 30-min session durations, but only when workload (number of images analyzed per min) was high. It is, however, still unclear how long screeners can maintain performance in X-ray image inspection and how performance evolves over longer screening durations up to 1 hr. Several studies have reported positive effects of breaks on performance in a variety of detection tasks (Arrabito et al., 2015; Kopardekar & Mital, 1994). However, in an X-ray image inspection task with student participants, Chavaillaz et al. (2019) found no performance differences for different break regimens. To examine the effects of time on task and breaks on performance, we compared the performance of two groups of screeners during a 60-min X-ray image inspection task in Manuscript 1. Whereas one group screened continuously for 60 min, the other took 10-min breaks between 20-min screening blocks in line with the current EU regulation (European Commission, 2015). Participants in vigilance tasks typically report increased distress and decreased engagement after task completion (Helton, 2004; Matthews et al., 2002). When it comes to the effects of time on task in X-ray image inspection, screener well-being has not yet been considered. We therefore asked screeners to complete the SSSQ (Helton, 2004) at the end of the screening task. Because there is no evidence yet on how screener performance changes after 30 min of screening, we first investigated this in the laboratory in a situation similar to remote screening. This enabled us to prevent negative consequences if performance were to decline, because this could result in missed threats. Only if we find that screeners can maintain performance in the laboratory for more than 30 min, can we then conduct similar studies in the field.

Because time on task in X-ray image inspection is likely to influence the response tendency (Ghylin et al., 2007; Rubinstein, 2020), it is important to consider a valid performance measure that is independent of this. Previous research (Godwin, Menneer, Cave, & Donnelly, 2010; Hofer & Schwaninger, 2004; Sterchi et al., 2019; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010), suggests that for X-ray image inspection, d_a, with a slope parameter of around 0.6 is a more valid measure of sensitivity compared to d' that might be affected by target prevalence. Accordingly, the same slope should be used to calculate the criterion c_a. In Manuscript 1, we varied

target prevalence to validate the appropriate detection measure and to determine whether time on task causes a change in sensitivity and/or the response tendency.

Method. A total of 71 professional screeners (33 female; age: M = 32.01 years, SD = 12.82; length of employment: M = 2.08 years, SD = 2.23) completed a 1-hr X-ray image inspection test twice. A 2 (breaks: with vs. without) × 2 (prevalence: high vs. low) × 3 (time on task: 0–20 min, 20–40 min, 40–60 min) mixed factorial design was employed. As a between-subject variable, the break conditions with and without breaks were used. Both groups conducted two test sessions using different target prevalences that were separated by 3–5 weeks. The hit rate, false alarm rate, sensitivity (d', d_a), criterion (c, c_a), and processing time served as dependent variables. The influence of the break and prevalence condition on the three SSSQ factors distress, worry, and engagement (Helton, 2004) was also analyzed. The test consisted of 864 X-ray images of passenger cabin baggage. In the low prevalence condition, one in eight images contained a threat (12.5%); in the high prevalence condition, one in two (50%). Guns, knives, and IEDs accounted for an equal number of threat items. To ensure that screener's performance was based on the same images, the test was divided into 12 blocks of 72 images. After 5 min, the test switched to the next block. Each participant had enough time to analyze the first 24 images of each block that were then used to measure detection performance. The order of blocks was counterbalanced.

The group with breaks took 10-min breaks every 20 min, whereas the group without breaks analyzed X-ray images for 60 min continuously and took a break of 20 min at the end. The order of the prevalence conditions was counterbalanced across participants. The slope parameter was estimated by comparing individual differences in hit rate and false alarm rate between the two prevalence conditions. After the screening task, screeners filled in the SSSQ survey (Helton, 2004) to assess perceived stress and provided demographic information.

Results. Both the hit rate and false alarm rate were higher in the high prevalence condition, $F(1, 69) = 37.99$, $p < .001$, $\eta_p^2 = .36$ and $F(1,69) = 118.53$, $p < .001$, $\eta_p^2 = .63$ respectively (see Figure 2). The estimated slope resulted in a value of 0.65 (95% BCa-CI [0.41, 0.89]), which was lower than the slope of 1.0 assumed by d'. In line with Wolfe et al. (2007), this suggests using d_a as a sensitivity measure and c_a as criterion (with the estimated slope of 0.65). The effect of time on task depended on the prevalence condition for the false alarm rate, $F(1.97, 136.18) = 17.9$, $p < .001$, $\eta_p^2 = .21$, criterion c_a , $F(1.95, 134.28) = 11.82$, $p < .001$, $\eta_p^2 = .15$), but not for the hit rate, $F(1.96, 134.94) = 3.06$, $p = .051$, $\eta_p^2 = .04$, or d_a , $F(1.95, 134.72) = 0.11$, $p = .895$, $\eta_p^2 = .00$ (see Figure 3). To be more precise, the criterion decreased for high prevalence from the first 20-min block (0–20 min) to the second 20-min block (20–40 min), whereas the criterion increased for low prevalence. There was a

small main effect of time on task for d_a, $F(1.97, 135.91) = 3.43$, $p = .036$, $\eta_p^2 = .05$. Post hoc tests revealed an increase from the first 20-min block to the second 20-min block ($p = .034$). The hit rate and false alarm rate, d_a and c_a were not affected by breaks $F(1, 69) = 1.84$, $p = .180$, $\eta_p^2 = .03$ and $F(1, 69) = 0.00$, $p = .957$, $\eta_p^2 = .00$ respectively.

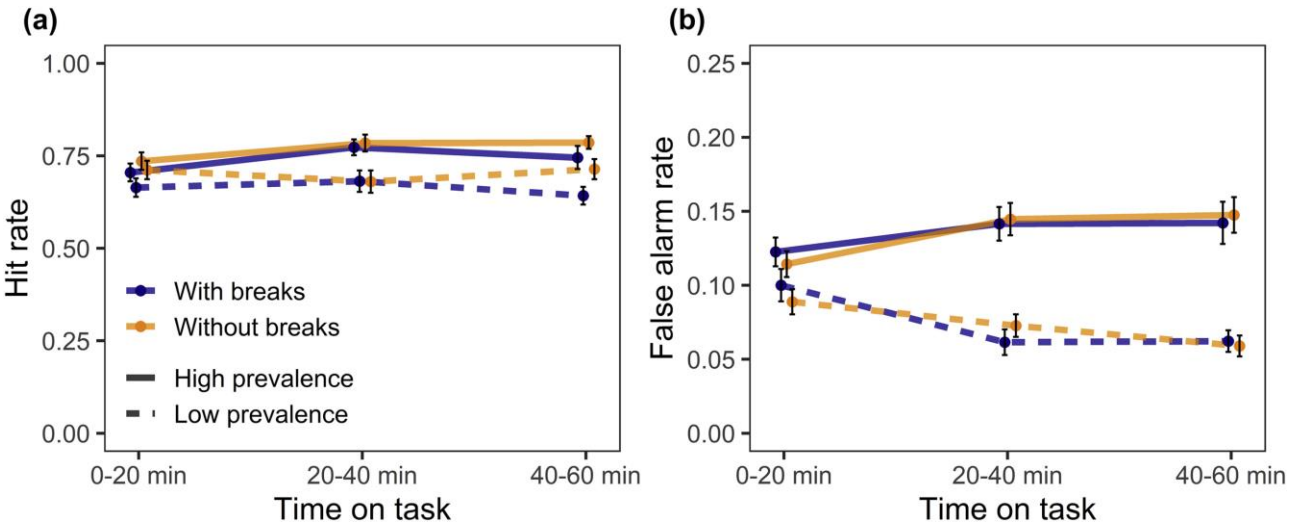


Figure 2. Hit rate (a) and false alarm rate (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task. Error bars represent standard errors.

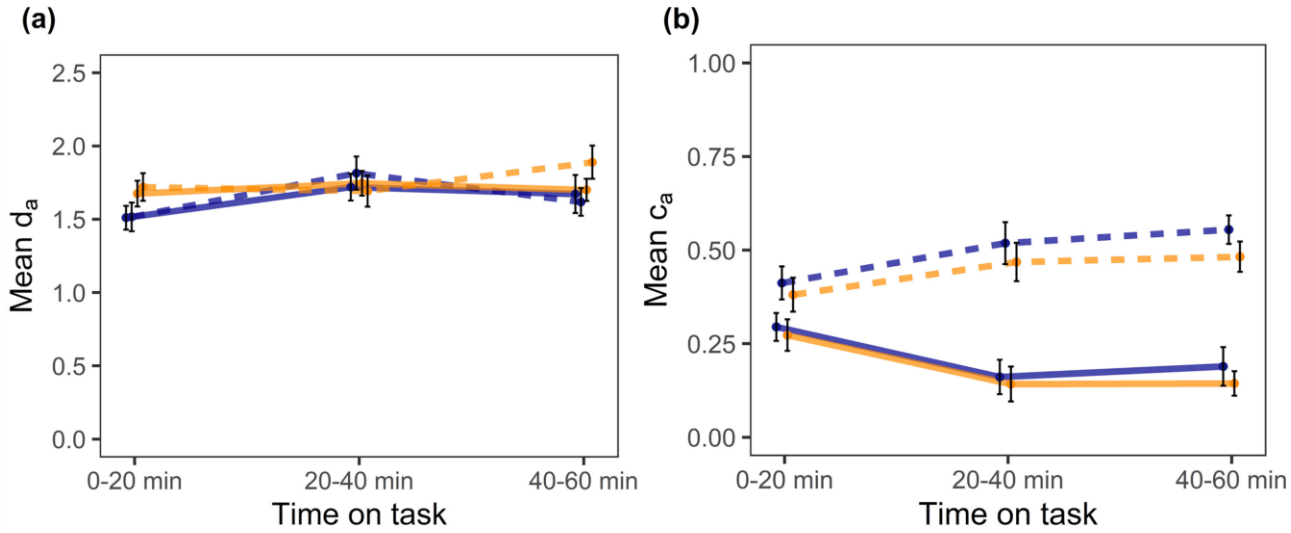


Figure 3. Sensitivity measure d_a (a) and criterion c_a (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task. Error bars represent standard errors.

The three constructs distress, worry, and engagement were used as dependent variables in 2 (with vs. without breaks) x 2 (high vs. low prevalence) ANOVA calculations for the subjective stress levels. A main effect of break, $F(1, 66) = 9.17$, $p = .004$, $\eta_p^2 = .12$, was found for distress. Because the data did not meet the assumptions of normality or homoscedasticity, a Wilcoxon rank-sum test was calculated. This revealed a significant difference between the two break conditions ($W = 1616$, $p = .003$). No effects were found for worry or engagement.

Discussion and conclusion. To examine screeners' ability to maintain performance over an hour and the effects of breaks, two groups of screeners performed a 1-hr X-ray image inspection task. One group took breaks every 20 min in accordance with EU regulations, the other group screened for 1 hr without breaks. To determine the valid detection measure target, prevalence was varied. Performance did not decrease over the course of 60 min of X-ray baggage inspection, but a shift in response tendency was evident. Moreover, breaks had no effect on performance. However, screeners without breaks reported more distress.

Consistent with previous studies, we found the target prevalence to cause a shift in the response tendency resulting in a lower hit rate and a lower false alarm rate with screeners needing less time to inspect images in the low target prevalence condition (Godwin, Menneer, Cave, Helman, et al., 2010; Ishibashi et al., 2012; Ishibashi & Kita, 2014; Lau & Huang, 2010; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010). In line with previous research on X-ray image inspection (Godwin, Menneer, Cave, & Donnelly, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010), we found higher *d'* values for the low target prevalence condition compared to the high target prevalence condition. In agreement with Kundel (2000) and Wolfe et al. (2007), we argue that it is implausible for screeners to become better and faster at detection when fewer threats occur. It is more likely that the equal variance assumption of *d'* (Green & Swets, 1966) is not met, and that the change in the hit rate and the false alarm rate reflect a change in response tendency (c_a) as assumed in SDT (Macmillan & Creelman, 2005). We found an average slope parameter of 0.65, which is close to the slope found in previous studies for the task of X-ray image inspection (Godwin, Menneer, Cave, & Donnelly, 2010; Sterchi et al., 2019; Wolfe et al., 2007; Wolfe & Van Wert, 2010).

For the false alarm rate, we found an interaction between target prevalence and time on task. The false alarm rate increased from the first (0–20 min) to the second (20–40 min) screening block at high prevalence and decreased at low prevalence. Previous research suggests that the target prevalence effect depends on implicit learning rather than explicit instruction, and that it takes some time for searchers to adapt to the current target prevalence by shifting their criterion accordingly (Ishibashi et al., 2012; Lau & Huang, 2010). According to Lau and Huang (2010), the instructed target

prevalence is insufficient to produce the target prevalence effect. Although our participants were instructed about the target prevalence, the target prevalence effect evolved over time, showing that participants first had to experience the target prevalence for the effect to fully develop (Ishibashi et al., 2012; Lau & Huang, 2010). Instructions alone were not sufficient to evoke this effect. In the high target prevalence condition, screeners shifted their response tendency to a more liberal location, increasing their likelihood of declaring the presence of a prohibited item. In the low prevalence condition, they shifted their response tendency to a more conservative location, therefore giving more no-target-present answers. After the first 20 min, the criterion remained stable in both conditions. The sensitivity d_a increased from the first 20-min block to the second 20-min block. As with other recognition tasks, there may be a warm-up phase in X-ray image inspection during which the cognitive processes required for this task are fully activated (Allport & Wylie, 1999; Monsell, 2003). However, the observed performance increase could also be due to adaptation to the task specifics in our experiment. Breaks have been shown to improve performance in previous studies (Arrabito et al., 2015; Balci & Aghazadeh, 2003; Kopardekar & Mital, 1994; Steinborn & Huestegge, 2016), but they are primarily thought to provide rest, recuperation, and fatigue prevention (Tucker, 2003). Because participants who screened for 60 min continuously showed no performance decrease, there was no need for recovery during breaks. Even though breaks had no effect on detection performance, they did appear to influence perceived distress. In the SSSQ, screeners in the condition without breaks reported more distress. This may influence performance in the long run. It should be noted, however, that there were significant differences in distress perception between screeners in the condition without breaks. This study serves as an initial indication that longer screening sessions are feasible without compromising performance, and it provides a solid foundation for future research in the field.

Manuscript 2: Time on task and task load in X-ray image inspection: a four-month field study with X-ray baggage screeners

Motivation and aim of the study. It is believed that the current 20-min limit in X-ray baggage screening is based on vigilance research (personal communication with airport security expert, March 2019). However, differences can be found between typical vigilance tasks and X-ray image inspection and their vigilance decrement patterns (Drury & Watson, 2002; Rubinstein, 2020; Wolfe et al., 2007). In vigilance tasks, a performance decline is often observed in the form of an increase in misses, false alarms, and reaction times (Davies & Parasuraman, 1982; See et al., 1995). This indicates a decline in sensitivity, and is often accompanied by a decrease in task engagement and an increase in distress compared to pretask values (Claypoole et al., 2019; Teo & Szalma, 2011; Tiwari et al., 2009; Warm, Parasuraman,& Matthews, 2008). In X-ray image inspection, the vigilance pattern with time on task has been found to manifest in an increase in misses and a decrease in both false alarms and reaction times (Ghylin et al., 2007; Rubinstein, 2020). This pattern is more consistent with a shift in response tendency and was also found for the low target prevalence condition in Manuscript 1. The classical vigilance decrement is explained mostly through limited attentional resources (Helton & Warm, 2008; Matthews et al., 2010) or underload theories (Robertson et al., 1997). To explain the deviant pattern in inspection tasks, Rubinstein (2020) proposed a new theory: DART. To gather further evidence on how time on task affects performance in X-ray baggage inspection and whether the vigilance pattern corresponds to the one suggested by Rubinstein, we conducted a 4-month field study with professional screeners.

Following Manuscript 1, Manuscript 2 investigated how screener performance evolves with time on task under real working conditions with remote screening. This study builds on Manuscript 1 and can address some of its limitations. Whereas the consequences of missing a prohibited item are minor in an experiment (Manuscript 1), they can be catastrophic in real life. Consequently, working at the checkpoint entails a greater degree of responsibility. Further, target prevalence at the checkpoint is significantly lower, and it is unclear whether screeners can maintain performance under this condition. Also, in the field study, screeners repeatedly perform longer screening durations over a longer period of time. Airports are increasingly moving the screening of cabin baggage away from the checkpoint to remote screening rooms (Kuhn, 2017). This quieter working environment could have a positive impact on performance. Because longer screening durations are desired especially by airports implementing remote screening, the field study was conducted at such an airport. At an international airport, a group of professional screeners inspected X-ray images of cabin baggage for up to 60 min, while a control group screened for 20 min in line with the current regulation (European

Commission, 2015). We used TIP data to investigate changes in detection and the SSSQ (Helton, 2004) to investigate distress and engagement after screening.

Method. The study was conducted at an international airport with 50 professional screeners who worked regularly at the investigated checkpoint. Screeners were randomly divided into two groups. The study group (22 screeners, 11 females; age M = 30.77 years, SD = 8.38; length of employment: M = 3.66 years, SD = 1.41) was instructed to screen for up to 60 min. However, they were given the option to stop earlier if they felt tired or unconcentrated. If they ended a screening session before 60 min, they were asked to note down the reason. A control group (19 screeners, 9 females; age M = 34.89 years, SD = 10.97; length of employment: M = 2.80 years, SD = 1.42) screened according to the EU rule, thereby rotating position after 20 min of screening. Both groups screened from a remote room located next to the checkpoint. The study was conducted during the screeners' regular working hours without affecting their compensation. To measure subjective stress, screeners were asked to fill in the SSSQ (Helton, 2004) after completing a screening session every 3 weeks. Upon completion of the study, screeners completed a short survey that included questions on the screening durations. For the analysis, we selected screeners who conducted a minimum of eight X-ray baggage screening sessions during the study. Consequently, 41 screeners were selected.

We used linear mixed models to assess the effects of time on task and task load on the dependent variables hit rate, reject rate, and processing time. Because Meuter and Lacherez (2016) found an interaction between the number of images screened per min and time on task, we also investigated this interaction. The hit rate was the percentage of correctly identified TIP images. The reject rate was the percentage of all bags sent to a manual bag search. Processing time was defined as the number of seconds screeners took to decide whether an image contained a prohibited item (rounded to full seconds by the TIP system). For these analyses, only sessions by the study group that lasted from 10 to 70 min were included, resulting in 1,170 screening sessions and approximately 250,000 analyzed X-ray images and over 6,000 TIP images. The models included time on task, task load, Time on task x Task load, days since study start, and daytime as fixed effects; and the session nested in the screener as random effects. Time on task was calculated as the difference between the time the screeners logged into a screening session and the time the decision for that image was made. Task load was the mean number of images a screener analyzed per min from the start of the screening session. “Days since study start” was included to examine whether habituation or fatigue occurred with increasing study duration and to account for seasonal changes. Daytime was included to control for the variation of passenger types and their bags throughout the day. To assess the effect of task load on session duration we fitted a linear mixed model that included the mean session task load, days since study start, and daytime as fixed effects and screener as the random effect. All metric variables

(time on task, task load, log processing time, and duration) were *z*-transformed to ensure better model convergence. Data from the SSSQ were aggregated and averaged per screener and construct. Group means were computed by averaging these individual means for each group. A Wilcoxon–Mann–Whitney test was used to compare the two groups' means for each construct.

Results. An average screener in the study group conducted 53.2 screening sessions (*SD* = 36.4) of 34.7 min (*SD* = 5.68) in duration and analyzed 287 TIP images (*SD* = 211). By examining longer screening sessions in the study group with a mixed model, we found no main effect of time on task for the hit rate (b = -0.068, SE = 0.041, p = .092), but a main effect for the reject rate and processing time (reject rate: b = -0.039, SE = 0.006, $p < .001$; processing time: b = -0.042, SE = 0.002, $p < .001$). For the three dependent variables hit rate, reject rate, and processing time, we found a significant main effect of the task load, a significant interaction of Time on task x Task load, and a main effect of days since study start (task load: hit rate, b = -0.137, SE = 0.046, p = .003; reject rate, b = -0.049, SE = 0.007, $p < .001$; processing time, b = - 0.123, SE = 0.005, $p < .001$; Time on task x Task load: hit rate, b = 0.140, SE = 0.041, $p < .001$; reject rate, b = -0.015, SE = 0.007, p = .022; processing time, b = 0.005, SE = 0.002, p = .029; days since study start: hit rate, b = 0.153, SE = 0.046, $p < .001$; reject rate, b = 0.121, SE = 0.008, $p < .001$; processing time, b = 0.108, SE = 0.008, $p < .001$). Figure 4 shows the effects of time on task for the three levels of task load to illustrate how performance changed with time on task and task load. For the mixed-effect models of the hit rate as well as the screening duration, a substantial amount of variance was explained by the random effects and therefore by variance between screeners. For the hit rate, the fixed effects (time on task, task load, Time on task x Task load, days since study start, daytime) explained 1.8% of the variance, whereas random effects (session and screener) explained 13.2% of the variance: 9.9% by the screener and 3.4% by the session. For the screening duration, the fixed effects (mean session task load, days since study start, daytime) explained 6.4% of the variance, whereas 24.7% of the variance was explained by the random effect, and therefore, by the screener.

The SSSQ was filled in up to five times (M = 3.45, SD = 1.48) by 21 participants in the study group and 19 in the control group. Figure 5 shows the means of individual means of the constructs in the questionnaire for each group. Group comparisons found no difference for distress (W = 261.5, p = 0.095) or worry (W = 262, p = 0.093), but higher values in engagement for the study group (W = 112, p = 0.018). Fifteen screeners in the study group completed the questionnaire on screening durations. Screeners reported that it became difficult for them to continue screening at around 30 to 40 min (M = 39.29, SD = 9.17) and that the optimal screening duration was around 30 min (M = 31.79, SD = 9.92).

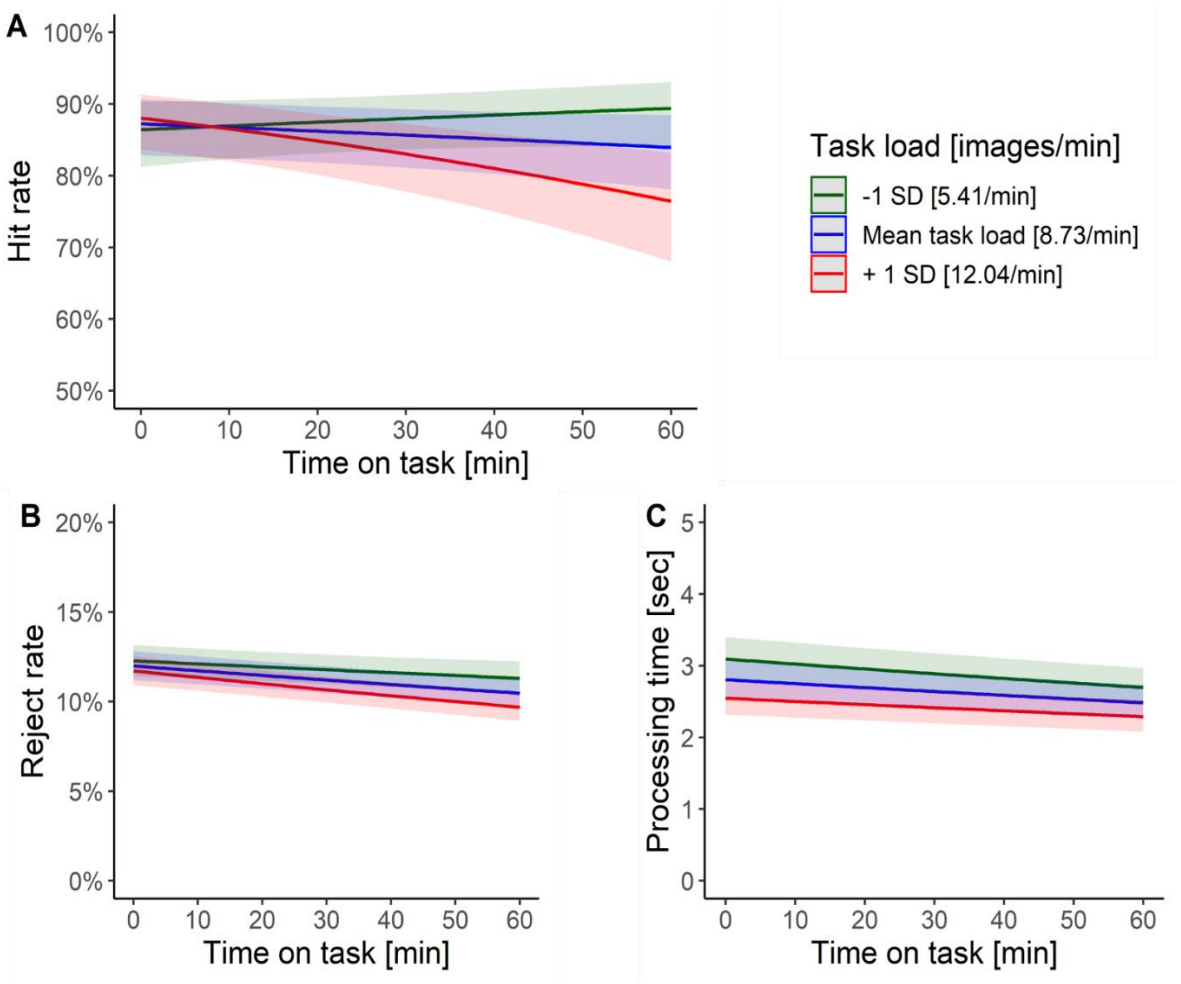


Figure 4. Effects of time on task on hit rate (A), reject rate (B), and processing time (C) depending on task load (SD = standard deviation).

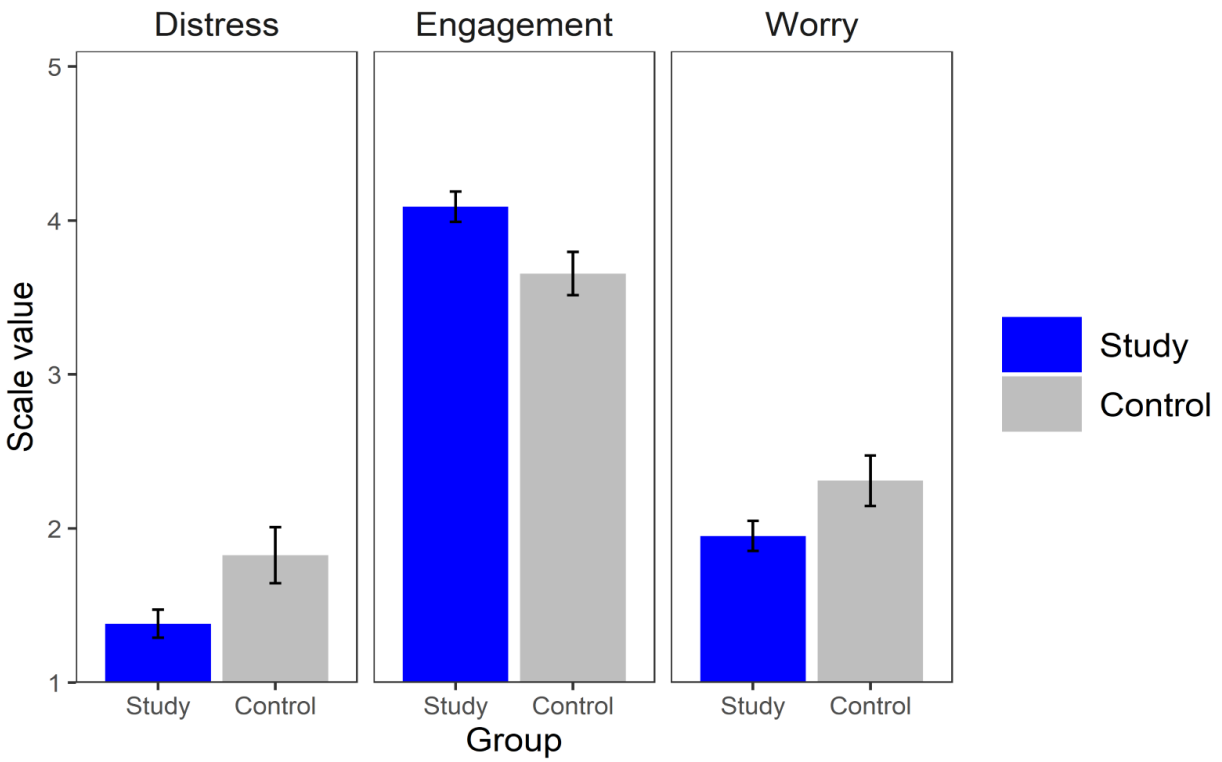


Figure 5. Mean of reported levels of distress, engagement, and worry for the study and control group. Error bars represent standard errors.

Discussion and conclusion. This study examined the effects of time on task and task load on the performance and subjective stress levels of X-ray baggage screeners. A group of screeners (study group) from an international airport conducted screening sessions lasting up to 60 min, whereas a control group screened as usual for around 20 min. For the longer screening durations of the study group, we found an interaction between time on task and task load for the performance measures. When task load was high (number of images analyzed per min), the hit rate decreased with time on task, but it stayed stable when task load was low or average. 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. Furthermore, increased time on task and task load resulted in a lower reject rate and faster processing times. Our findings are not in line with the assumptions of underload theory (Robertson et al., 1997), and cannot be explained fully by resource theory (Helton & Warm, 2008; Matthews et al., 2010), although both theories are widely used to account for the vigilance decrement in vigilance tasks (Helton & Warm, 2008; MacLean et al., 2010; Neigel et al., 2020). Conversely, our results are in line with DART as proposed by Rubinstein (2020). This can explain decreases in the hit rate, reject rate, and processing times as a coping strategy. Rather than getting worse at discriminating between targets and nontargets with time on task, DART proposes that screeners actively shift their response tendency to save resources. In other words, with

increasing task demands (time on task, or higher task load), screeners switch to a resource-efficient response pattern that has negative consequences for hit rates. This suggests that the effects of time on task manifest differently for X-ray baggage screening compared to other vigilance tasks. This observation is consistent with other studies that observe resource-conserving behavior such as increased reliance on automation when workloads are high (Dixon & Wickens, 2006; C. D. Wickens & Dixon, 2007). Our findings confirm that a different vigilance decrement pattern occurs for X-ray image inspection compared to the typical vigilance decrement for which a decrease in the hit rate, increase in false alarms, and increase in reaction times is found.

The study group who screened 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 and thus contributes to recovery. Another explanation could be that the study group showed more engagement because participants were more aware of contributing to research than participants in the control group. In addition, this group was able to determine when to end a screening session on their own, giving them additional autonomy that could increase engagement (Bakker & Demerouti, 2007; Hackman & Oldham, 1975). However, we did observe high individual differences in the study group not only in preferred and conducted screening durations but also in performance. The results from the questionnaire on screening duration suggest that screening durations around 30 to a maximum of 40 min would be feasible, which is consistent with the actual performed screening duration of about 35 min by the study group. If the results of our study can be replicated in remote screening conditions with different airports, trials can be extended to settings where screeners analyze X-ray images at the checkpoint. Hence, extended screening durations can provide operational benefits with no or only small decreases in the hit rate during periods of high task load.

Manuscript 3: Reliability and validity of threat image projection data on X-ray baggage screening

Motivation and aim of the study. To measure how well screeners detect prohibited items, most airports use TIP. During baggage inspection, TIP projects prerecorded X-ray images of prohibited items (bombs, guns, knives, etc.) onto the X-ray images of 1–4% of all passenger baggage (Cutler & Paddock, 2009; Hofer & Schwaninger, 2005; Meuter & Lacherez, 2016; Skorupski & Uchroński, 2018). The responses to these TIP images are recorded and used for quality control by airports, governments, and security companies that typically determine the average TIP hit rate per screener on a half-yearly basis (Cutler & Paddock, 2009; Hofer & Schwaninger, 2005; Riz à Porta et al., 2022; Skorupski & Uchroński, 2016). TIP data are also used to answer research questions (Buser et al., 2023; Meuter & Lacherez, 2016; Skorupski & Uchroński, 2018). Although widely used, there are, however, no conclusive results on the reliability of TIP data (Hofer & Schwaninger, 2005) and their validity remains uninvestigated. Manuscript 3 therefore examines the reliability and validity of TIP by analyzing a large set of data from an international airport. The reliability of TIP performance measurement can be quantified by adopting a statistical model such as classical test theory (CTT). Due to differences between TIP and standardized tests, it is, however, unclear whether assumptions of CTT apply to TIP data. Therefore, Manuscript 3 also examines how well TIP meets the CTT assumptions, and whether estimates based on CTT such as the Spearman–Brown prediction (Brown, 1910; Spearman, 1910) can be applied to TIP. To assess whether TIP is a valid measure of real prohibited items, we analyzed whether it can predict how well screeners detect prohibited items in covert tests in which instructed people attempt to smuggle prohibited items (e.g., knives, bombs, or guns) past the checkpoint in their baggage (Walter et al., 2021; Wetter et al., 2008). If TIP performance is a valid measure, it should be able to predict how probably a screener will detect real prohibited items in a baggage X-ray image.

Method. We analyzed 4 years of CBS TIP and covert test data from an international airport. This was composed of 1,206,076 TIP events from 728 screeners and 1,194 from 474 screeners. On average, the TIP systems at this airport projected fictional threats (TIP events) onto 2.9% of all X-ray images of passengers' baggage. Reliability was assessed by computing the split-half reliability. Therefore, for each screener, TIP events were sorted by date and time of occurrence, and every two consecutive TIP events were paired. To estimate the reliability for *n* number of TIP events, *n* pairs of TIP events were randomly selected (without replacement) and the TIP events of each pair were randomly split into two groups. For each screener and each of the two groups of TIP events, the percentage of detected TIP events (hit rate) was calculated, and the Pearson correlation between the hit rates of the two groups was computed. To estimate reliability, all random steps (i.e., sampling and

splitting of pairs) were repeated 10,000 times and the resulting correlation coefficients were averaged. To assess whether TIP meets the CTT assumptions and whether estimates based on CTT can be applied, we determined whether the Spearman–Brown prediction accurately described how the reliability varied as a function of the number of TIP events considered for the calculation of TIP performance. To estimate the reliability for the various numbers of TIP events using the same sample of screeners, we included only screeners with at least 100 TIP events per 6 months. We calculated the split-half reliabilities (as described above) considering 5–50 TIP events for performance evaluation in increments of five. The reliability of the corresponding number of TIP events was then estimated using the Spearman–Brown prediction. For this purpose, the reliability of 25 TIP events (per split) was used as the baseline. We then compared the reliability values of the two calculations. To determine how reliable TIP performance measurement is, because it is often calculated by airports and authorities, we calculated the reliability per screener for each half year. To retain as many screeners as possible for this analysis, 10 TIP events per screener and per split and half-year were used to calculate the split-half reliability. Consequently, screeners with less than 20 TIP events within the respective half-year period were excluded. To analyze the validity of the TIP data, we performed correlational analyses and thereby examined the effect of the TIP hit rate on covert test performance using binomial generalized estimation equations (GEE; Ballinger, 2004; Liang & Zeger, 1986) and the R-package GEE (R Core Team, 2020). The TIP hit rate was included by aggregating TIP events that occurred within half a year before or after the covert test. The model controlled for the prohibited item category (gun, knife, IED, and other), different checkpoints within the airport, X-ray machine type, and complexity of the covert test as factors, and the screener as a random variable. Per screener, an average of 2.52 covert tests were analyzed (*SD* = 1.79).

Results. The Spearman–Brown prediction (Brown, 1910; Spearman, 1910) corresponded well with the empirically estimated reliabilities and, therefore, provided an accurate description of how the reliability increased with the number of TIP events. The calculation of the split-half reliability of the TIP performance for 20, 50, 100, or 345 TIP events per half year is shown in Figure 6. The latter number of TIP events (345) corresponds to the average number of TIP events inspected by a screener per half year. However, reliability decreased over time. This means that considering 92 TIP images for performance evaluation was sufficient to achieve a minimum reliability of 0.7 in the first half year, whereas 205 TIP images were necessary to obtain an equal reliability in the eighth half year.

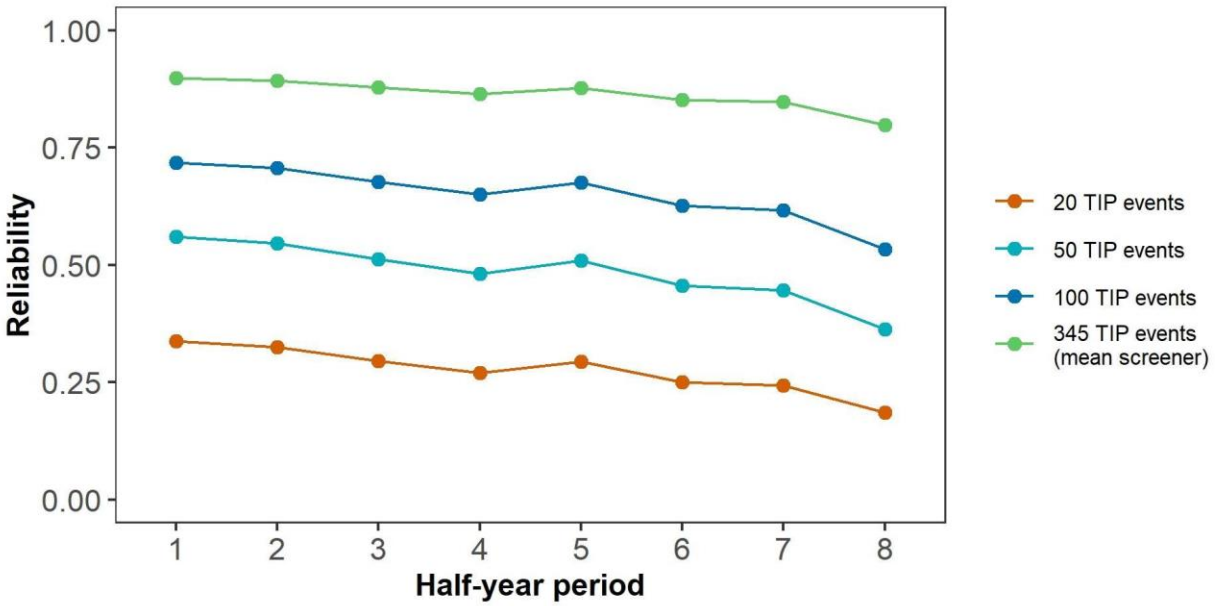


Figure 6. Reliability values for 20, 50, 100, and 345 TIP events (mean number of TIP events per screener per half-year period) for eight half-year periods.

Decomposing the reliability into standard error and true variance (Figure 7A and B) shows that the decrease in reliability was not attributable to an increasing standard error (which also decreased over time). It is more likely that the reliability declined due to an over-proportionate decrease in the true variance of the TIP performance between screeners. Figure 7C shows that the average TIP performance, in terms of the hit rate, increased over time, which might have led to a limited room for true variance (i.e., a ceiling effect).

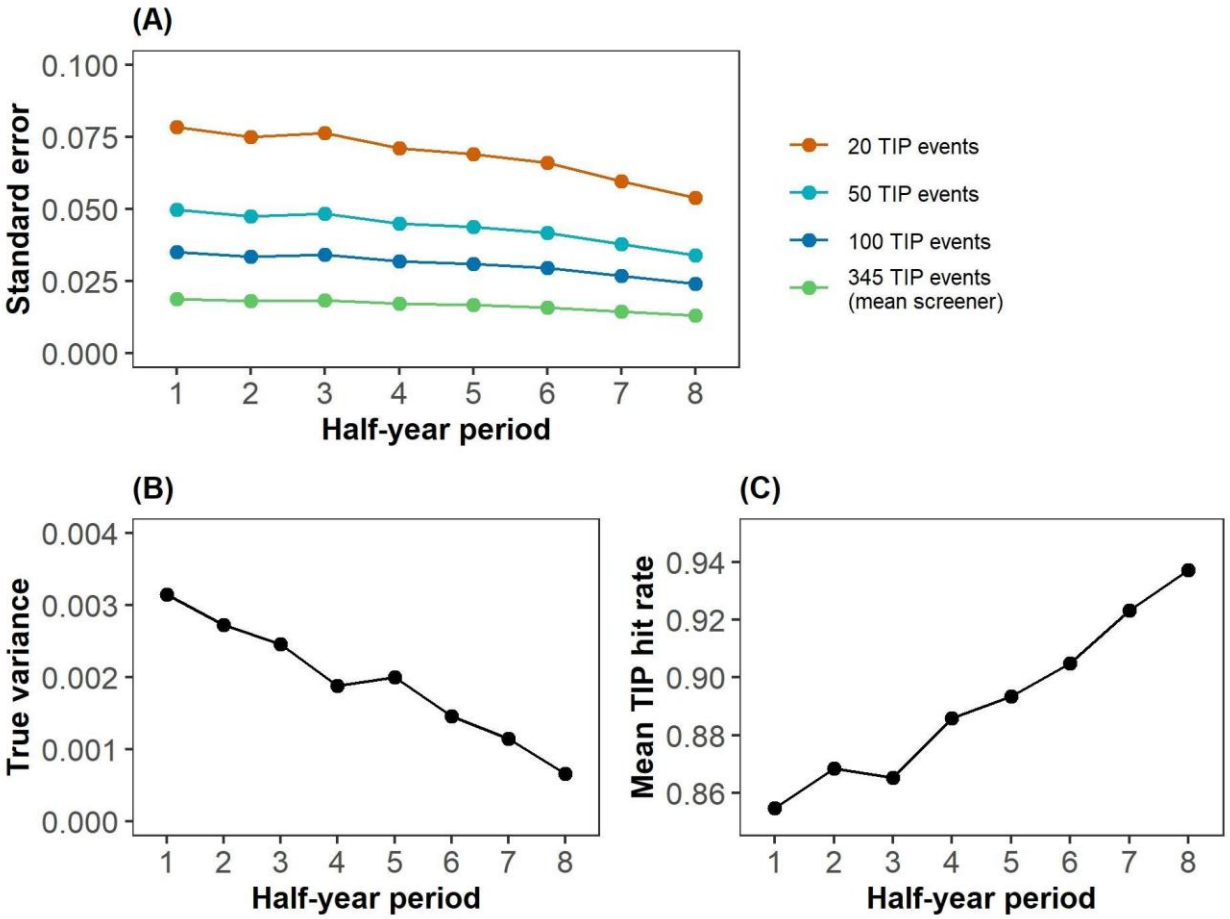


Figure 7. Mean standard error (A), variance between screeners (B), and the hit rate (C) for eight half-year periods.

The average covert test hit rate over all screeners was 79.50 % (*SD* = 0.40). The GEE revealed that screeners with a better TIP performance also showed higher covert test performance. In total, 1,194 covert tests were considered; and, on average, 826 TIP events were considered per covert test per screener (*SD* = 323.58). Figure 8 depicts the estimated relationship between the TIP hit rate and covert test performance.

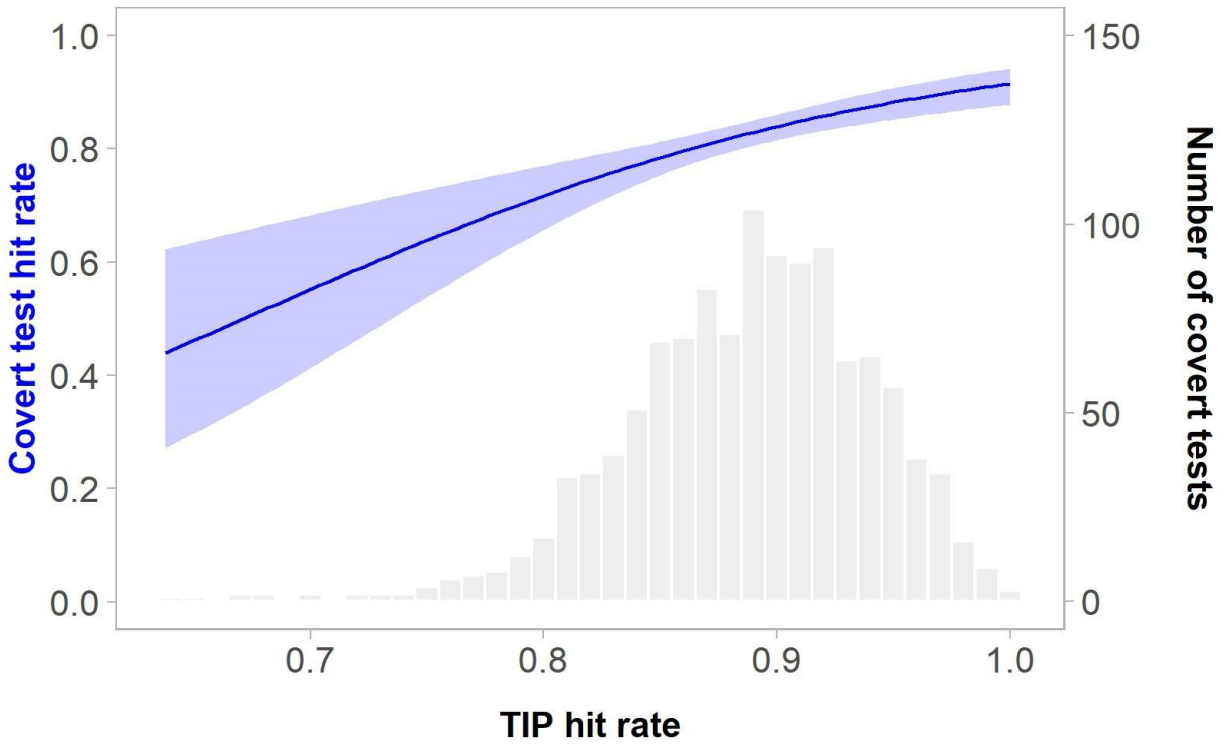


Figure 8. Relationship between covert test and TIP hit rate (blue line). 95% confidence band indicated by the blue area around the blue line. The histogram shows the distribution of the number of conducted covert tests.

Discussion and conclusion. TIP data are widely used for quality control in airport security as well as in research. We investigated whether TIP data provide a reliable and valid measure of detection performance by analyzing 4 years of data from an international airport. We found that reliability increased with the number of TIP events following the Spearman–Brown prediction (Brown, 1910; Spearman, 1910). This finding is important, because it means that the reliability of a TIP system can be estimated for a specific number of TIP events per screener (e.g., 50 TIP events), and the data can be extrapolated to calculate the necessary number of TIP events to achieve a desired reliability. For the investigated TIP library, approximately 100 TIP events were sufficient to achieve a minimum reliability of 0.7. This reliability value is recommended if the measure is to be used as a first indication (e.g., dividing the screeners into two performance groups) or for group diagnostics (Kline, 2000; Murphy & Davidshofer, 2014). However, if performance measures have consequences for screeners (e.g., mandatory remedial training), it is highly recommended to achieve higher reliability values of at least 0.8 (Brough, 2019; Murphy & Davidshofer, 2014). To achieve high reliability values, the dependency on the difficulty of the TIP images should be considered. Our results showed a decrease in reliability over time for a constant number of TIP events. This was probably due to an increase in the average hit rate. To avoid this, a proportion of TIP images should

be exchanged regularly (e.g., 10% every year) to prevent overlearning of images. Moreover, TIP libraries should be large enough to ensure that screeners do not view the same prohibited items too often.

Our analysis found that TIP performance was associated significantly with detection performance in covert tests. In other words, screeners who performed better in detecting TIP were more likely to detect prohibited items in covert tests. This study thereby provides the first validation of TIP performance. However, our results do not indicate that the TIP is perfectly realistic. For instance, we found that the hit rate was higher in the TIP test than in the covert test. Further, one should be aware that the TIP hit rate does not fully reflect the screeners' detection ability. In detection tasks such as X-ray image inspection, the hit rate depends not only on searchers' target-detection ability but also on their response tendency (Green & Swets, 1966; Macmillan & Creelman, 2005). A limitation of our study is that we could analyze the reliability and validity of TIP data from only one airport. It would be interesting to continue this research with TIP data from other airports using different TIP systems and screening technologies.

.

General Discussion

This thesis provides important theoretical, methodological, and practical contributions relevant to performance measurement and work design in X-ray image inspection at security checkpoints. Manuscript 1 confirmed that d_a with a slope parameter of around 0.6 is a more valid measure of detection performance compared to d'. Screeners maintained performance for one hour, without breaks affecting performance. However, a shift in response tendency with time on task was apparent. Evidence that longer screening durations may cause more distress highlights the importance of long-term studies in the field. Manuscript 2 provided additional evidence that screeners can maintain performance for longer than 20 min under real working conditions. When screeners could inspect X-ray images for up to 60 min, they screened an average of 35 min, which corresponds to the screening duration they reported as ideal. Those screeners did not report more distress. However, individual differences in performance, preferred screening durations, and conducted screening durations were observed. Manuscript 1 and, in part, Manuscript 2 provide evidence that time on task in X-ray image inspection results in a shift in response tendency and does not lead to a decline in sensitivity, thereby confirming Rubinstein's (2020) observation that time on task results in a different vigilance pattern compared to classical vigilance tasks. On this basis, an extension of the 20-min rule should be discussed. Manuscript 3 provided evidence that TIP performance data, which are frequently used for performance monitoring, can provide a reliable and valid measurement of operational threat detection. To obtain a reliable measurement of performance, at least 100 TIP events should be considered per screener. Manuscript 3 was the first to show that TIP performance and covert test performance are correlated.

This thesis is based on studies with a high ecological validity. All were conducted with professional screeners and real X-ray baggage images. The study in Manuscript 1 contributes to ecological validity through the similarity to remote screening conditions. In Manuscript 2 and Manuscript 3, TIP data from airports were analyzed. Additionally, the laboratory setting in Manuscript 1 makes a high contribution to internal validity. The findings have significant theoretical and methodological implications. Specifically, this work provides insight into how performance changes with time on task in X-ray baggage inspection, and it distinguishes this performance pattern from the classical vigilance decrement. Likewise, it provides methods and measures that allow a reliable and valid measurement of screener performance, which is of high practical value for airports, security companies, regulators, and researchers. This is particularly because the duration of X-ray image inspection and the number of images used for performance measurement are set by regulations that could be affected by this work. In addition to presentations at several conferences, the results of

this work have been shared and discussed with various airports and the EU Commission, who found this research very interesting and valuable.

Reliable and valid measurement of screener performance

This work provided important insights that contribute to reliable performance measurement at security checkpoints. In practice, the TIP hit rate is often used as an indicator of how well screeners detect prohibited items. Until now, it has been unclear whether TIP provides a reliable measure of screener performance. With Manuscript 3, we were able to show that TIP provides a reliable and valid measure of threat detection. We found that the reliability of the performance measurement depends on the number of TIP events considered for performance calculation. This dependency followed the Spearman–Brown prediction based on the assumptions of CTT. This implies that CTT methods such as the Spearman–Brown prediction can be applied to TIP data. This is an important theoretical and methodological contribution relevant to future analyses of TIP data. The reliability of performance measurement can also be influenced by the difficulty of TIP events. In Manuscript 3, we observed a decrease in reliability over time, probably due to an increase in the average hit rate of screeners, which, in turn, might be attributed to the familiarity of prohibited items used in TIP. Several practical conclusions can be drawn from these results for ensuring adequate reliability of performance measurement. One should consider a minimum of 100 TIP events per screener to measure performance reliably. Ideally, more TIP events per screener should be used, especially if reliability values above 0.7 are to be achieved. To avoid image overlearning, a portion of the fictional threat items (FTIs) should be exchanged regularly, and TIP libraries should be large enough. Further, TIP images should be checked for artifacts to prevent them from being too easy (Riz à Porta et al., 2022). Also, if the average TIP hit rate reaches very high values, the TIP library should be exchanged. It should be noted that reliability depends not only on the amount of measurement error but also on the true differences between individuals (Brough, 2019; Murphy & Davidshofer, 2014). Therefore, if one is interested in the absolute TIP performance, rather than the comparison of individuals, confidence intervals and standard errors of the measurement should be considered. These measures can be derived from the estimated reliability.

Manuscript 3 was the first to show that the TIP hit rate can predict performance in covert tests, indicating that it is a valid measure of operational threat detection. This means that screeners who were better at detecting TIP were more likely to detect real prohibited items in covert tests. Yet, our results do not suggest that TIP is fully realistic. It was still easier to recognize TIP images compared to covert tests, as evidenced by higher TIP hit rates. This aligns with previous findings indicating that TIP produces a share of unrealistic and easy images (Bassetti, 2018; Riz à Porta et al., 2022).

Nonetheless, TIP allows discrimination between screeners with high and low recognition performance and thus has predictive validity. Whereas the TIP hit rate can be very informative, one should keep in mind that it depends on a person's response tendency (Green & Swets, 1966; Macmillan & Creelman, 2005) and therefore does not fully reflect a screener's detection ability. Although it is not immediately relevant from a security perspective whether threats are found due to detection ability (sensitivity) or response tendency, this differentiation can be important when it comes to, for example, maintaining operational efficiency, optimizing training, or answering research questions. To this end, in addition to the hit rate, the false alarm rate of the screeners should be considered. Unfortunately, many FTI TIP systems do not allow measurement of the false alarm rate due to technical limitations (only the reject rate, as seen in Manuscript 2). However, with combined threat image (CTI) technology, images of fully prepared baggage, including the prohibited item, can be projected onto the screener's workstation (Schwaninger, 2006). Thus, images of an entire baggage without prohibited items can also be projected, which allows the false alarm rate to be calculated. Accordingly, the response tendency and measures independent of it, such as the sensitivity, can be calculated, thereby providing more valid measures of detection ability in the field. This technological advance provides practitioners with a more comprehensive understanding of screener performance, and this, in turn, allows for more targeted interventions.

Manuscript 1 showed that for X-ray image inspection, d_a with a slope of around 0.6 is a more valid measure of threat detection than *d'*. In agreement with other studies, we found that the signal-plus-noise and noise ratio do not follow a normal distribution with equal variances as assumed for the measure *d'* (Godwin, Menneer, Cave, & Donnelly, 2010; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010). Future research might shed more light on the determinants of the slope parameter and the causes of the unequal noise and signal-plus-noise distributions. One assumption is that prohibited items vary greatly in how well they are recognized (Sterchi et al. 2019). What is clear is that using an invalid measure when there is a large difference in response tendency will lead to erroneous conclusions—that is, it will falsely indicate a significant difference in sensitivity or a lack thereof.

Effects of time on task on screener performance in X-ray image inspection

As opposed to many vigilance tasks, we did not find a decline in detection performance with time on task after 15 to 20 min (Davies & Parasuraman, 1982; Mackworth, 1948; See et al., 1995; Teichner, 1974; Warm, 1984). In our laboratory study in Manuscript 1, detection performance d_a did not decline over the entire 60-min test. It even increased at the beginning of the test. In Manuscript 2, screeners maintained the TIP hit rate up to 60 min in the field when task load was low or average—

which was the case for 85% of all inspected images. A decrease in the hit rate was apparent only when task load was high. Moreover, unlike other studies (Arrabito et al., 2015; Balci & Aghazadeh, 2003; Galinsky et al., 2000; Kopardekar & Mital, 1994; Lim & Kwok, 2016; Steinborn & Huestegge, 2016), we did not find that breaks had a positive effect on performance. In Manuscript 1, the group screening continuously for 60 min showed a comparable performance to the group taking 10-min breaks every 20 min of screening. Similarly, in the field, the performance of the group conducting longer screening sessions (on average 35 min) and the group screening for 20 min was comparable. Because no decline in performance was evident, there was also no possibility of recuperating performance losses by taking breaks. This is consistent with the results of Chavaillaz et al. (2019), who found no performance differences between different break regimes in a baggage inspection task. Not only did we find no decrease in performance over time, but screeners worked more efficiently with increasing task duration. This was revealed by a decline in the false alarm rate for the low target prevalence condition in Manuscript 1 and a decline in the reject rate in Manuscript 2, resulting in fewer manual bag searches as well as declining processing times in both studies. This has a positive impact on operation as it increases passenger throughput.

These observed performance patterns with increasing time on task do not correspond to the classical vigilance decrement, and they cannot be explained through underload (Robertson et al., 1997) or resource theory (Helton & Warm, 2008; Matthews et al., 2010). Rather than getting worse at discriminating between targets and nontargets with time on task, it seems that screeners shift their response tendency. Rubinstein's DART suggests that when task demands are high (increasing time on task, high task load), screeners actively switch to a resource-efficient response pattern (Rubinstein, 2020). In Manuscript 1, we were able to confirm a shift in response tendency that depended on the target prevalence. This shift was based largely on changes in the false alarm rate, which occurred at the beginning of the task. Interestingly, we saw a shift in the response tendency from the first to the second 20-min block, which suggests that participants needed some time to adjust to the prevailing prevalence. In addition to confirming the target prevalence effect (Wolfe et al., 2007), Manuscript 1 therefore supports the notion that the target prevalence effect is caused by implicit learning as opposed to explicit instructions (Ishibashi & Kita, 2014; Lau & Huang, 2010). Whether the decrease in the reject rate and processing time observed in the field study are truly a shift in response tendency to save resources is something we cannot deduce conclusively. Yet, the fact that we observed a decrease in hit rate only when the task load was high supports the concept that resources are conserved when task demand is high. This resource-conserving behavior has also been found in other studies in which, for example, individuals rely more on automation when task load is high (Dixon & Wickens, 2006; C. D. Wickens & Dixon, 2007).

What can be deduced from our findings for practice, and can a relaxation of the 20-min regulation be recommended? Manuscript 2 showed that for a majority of the cases (low and average task load), longer screening durations did not affect detection negatively. Moreover, with increasing time on task, screeners worked more efficiently, and this provides operational benefits. In addition, Manuscript 2 illustrated that individual differences were much more crucial as a determining factor for high detection performance compared to time on task. This aligns with research that individuals differ in visual cognitive abilities and vigilance and working memory capacity—findings that are relevant for the recognition and detection of prohibited items in X-ray baggage images (Hardmeier & Schwaninger, 2008; Hättenschwiler et al., 2019; Mitroff et al., 2018; Peltier & Becker, 2020; Rusconi et al., 2015; Schwaninger et al., 2005). Looking only at the effects of time on task on performance, our studies suggest that a relaxation of the 20-min regulation can be discussed.

In addition to performance, screeners' well-being should be considered. Whereas Manuscript 1 suggested that longer screening durations might lead to more distress, this could not be confirmed in the field study. There, higher engagement was observed for the group conducting longer screening durations of around 35 min. This could be due to greater autonomy, because they were allowed to end screening sessions after 20 min in a self-determined manner. However, we cannot rule out other factors that led to higher engagement. One should also consider that there was significant variation in conducted screening durations between screeners in Manuscript 2. Considering the average duration of screening sessions performed and the preferred screening duration reported by screeners, a duration of 30-40 min appears to be suitable for the majority of screeners. Ideally, screeners would be able to decide for themselves how long they continue screening after 20 min (with upper limit), because Manuscript 2 revealed large individual differences in performance and preference for screening duration. Moreover, this would be a way to grant screeners more autonomy in what is a highly regulated job. Autonomy is considered to be a crucial factor in theories of work design and has been linked to job satisfaction (Hackman & Oldham, 1980; Ryan & Deci, 2000). The generalizability of our conclusions to airports with fixed screening sessions might be limited, because the screeners in our field study could decide to end screening sessions in a self-determined way. To make definitive statements on the extension of screening durations, it is highly advisable to conduct further field studies, because airports differ in throughput, passenger types, screening technologies, and many other aspects. These studies should consider and monitor screeners' well-being, because it may be an important component for the long-term acceptance and success of longer screening durations.

Conclusion

The detection of prohibited items in cabin baggage is one of the key aspects ensuring safe air travel. To prevent mistakes, it is vital to establish regulations that support performance maintenance and to monitor performance of detection in X-ray baggage screening with methods and measures that are reliable and valid. The findings of this thesis contribute to a better understanding of how task duration affects performance in X-ray baggage inspection, and whether current means of monitoring screener performance at checkpoints provide reliable measurements. Finally, this thesis offers theoretical and practical contributions to work design at security checkpoints.

The findings from Manuscripts 1 and 2 indicate that screeners can maintain performance even during extended screening durations up to 60 min. These results suggest that time on task primarily influences the response tendency in X-ray baggage inspection and should be distinguished from the classical vigilance decrement. Based on performance and survey results, the current 20-min regulation (European Commission, 2015) could be designed more flexibly and an extension to 30 to 40 min can be considered. Manuscript 3 shows that TIP, which is used by airports around the world, allows reliable measurement of screener performance. It also demonstrated, for the first time, that there is a correlation between threat detection in TIP and covert tests, indicating that TIP is a valid measure of operational threat detection. From a practical standpoint, it was established that a minimum of 100 TIP images should be considered to reliably measure performance, and additional recommendations are provided to increase the reliability of TIP data.

Future research may shed light on whether similar results regarding time on task can be found for screeners who do not work in remote screening rooms and for longer screening durations with a fixed duration. If regulators and airports contemplate implementing longer screening durations, especially those exceeding 30 min, it is crucial to investigate their effects on factors such as eyestrain (Kaur et al., 2022; Mehra & Galor, 2020) and perceived mental workload (Teo & Szalma, 2011; Warm, Matthews & Finomore, 2008). Some airports are discussing dedicated screening positions in which certain screeners would conduct solely X-ray image inspection for several hours a day. In such cases, the significance of task rotation at the checkpoint should be explored—not only for providing breaks from screening but also for enhancing task variety (Hackman & Oldham, 1980). Further research with CTI TIP would be of interest for both performance monitoring with time on task and TIP data reliability. The implementation of CTI TIP allows to measure false alarm rates at checkpoints, which, along with hit rates, permits the calculation of the sensitivity and response tendency. This contributes to a more accurate measurement and comprehensive understanding of screener performance under real working conditions. Moreover, CTI TIP promises to be a more reliable measure of TIP hit rate compared to FTI TIP, because images can be prescreened to remove

low-quality images. Investigating differences in reliability between FTI and CTI TIP could help to further improve performance measurement.

As in other industries, Artificial Intelligence (AI)-based processes and automation are increasingly being developed and deployed at security checkpoints. However, screeners will retain an important role in X-ray image analysis for some time to come, although their role may shift to tasks involving oversight, system control, and resolution of complex alarm scenarios (Harris, 2002; Wetter, 2013). The results of this work remain relevant to the use of such newer technologies and may even help to guide their implementation. Finally, our findings and practical contributions extend to related domains and may shape the work design of other jobs that involve the inspection of baggage for prohibited items in, for example, prisons, stadiums, or hotels.

References

Allport, A., & Wylie, G. (1999). Task-switching: Positive and negative priming of task-set. In G. W. Humphreys, J. Duncan, & A. Treisman (Eds.), *Attention, space, and action: Studies in cognitive neuroscience* (pp. 273–296). Oxford University Press.

Arrabito, G. R., Ho, G., Aghaei, B., Burns, C., & Hou, M. (2015). Sustained attention in auditory and visual monitoring tasks: Evaluation of the administration of a rest break or exogenous vibrotactile signals. *Human Factors*, *57*(8), 1403–1416. https://doi.org/10.1177/0018720815598433

Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources model: State of the art. *Journal of Managerial Psychology*, *22*(3), 309–328. https://doi.org/10.1108/02683940710733115

Balci, R., & Aghazadeh, F. (2003). The effect of work-rest schedules and type of task on the discomfort and performance of VDT users. *Ergonomics*, *46*(5), 455–465. https://doi.org/10.1080/0014013021000047557

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*(2), 127–150. https://doi.org/10.1177/1094428104263672

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

Bassetti, C. (2018). Airport security contradictions: Interorganizational entanglements and changing work practices. *Ethnography*, *19*(3), 288–311. https://doi.org/10.1177/1466138117696513

Bassetti, C. (2021). The tacit dimension of expertise: Professional vision at work in airport security. *Discourse Studies*, *23*(5), 597–615. https://doi.org/10.1177/14614456211020141

Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica*, *152*, 158–165. https://doi.org/10.1016/j.actpsy.2014.08.005

Biggs, A. T., Kramer, M. R., & Mitroff, S. R. (2018). Using cognitive psychology research to inform professional visual search operations. *Journal of Applied Research in Memory and Cognition*, *7*(2), 189–198. https://doi.org/10.1016/j.jarmac.2018.04.001

Brough, P. Ed. (2019). *Advanced Research Methods for Applied Psychology*. Routledge. https://doi.org/10.4324/9781315517971

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Buser, D., & Merks, S. (2020). Centralised image processing: Challenges, trends and time on task. *Aviation Security International*, 33–35.

Buser, D., Schwaninger, A., Sauer, J., & Sterchi, Y. (2023). Time on task and task load in visual inspection: A four-month field study with X-ray baggage screeners. *Applied Ergonomics*, *111*, 103995. https://doi.org/10.1016/j.apergo.2023.103995

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2019). Work design for airport security officers: Effects of rest break schedules and adaptable automation. *Applied Ergonomics*, *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. *Human 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. *43rd Annual 2009 International Carnahan Conference on Security Technology, Zurich*, 46–51. https://doi.org/10.1109/CCST.2009.5335565

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

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

Drury, C. G., & Watson, J. (2002). Good practices in visual inspection. *Human Factors in Aviation Maintenance-Phase Nine, Progress Report, FAA/Human Factors in Aviation Maintenance*, 1–90.

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.* Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R1998

Galinsky, T. L., Swanson, N. G., Sauter, S. L., Hurrell, J. J., & Schleifer, L. M. (2000). A field study of supplementary rest breaks for data-entry operators. *Ergonomics*, *43*(5), 622–638. https://doi.org/10.1080/001401300184297

Ghylin, K. M., Drury, C. G., Batta, R., & Lin, L. (2007). Temporal effects in a security inspection task: Breakdown of performance components. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(2), 93–97. https://doi.org/10.1177/154193120705100209

Godwin, H. J., Menneer, T., Cave, K. R., & Donnelly, N. (2010). Dual-target search for high and low prevalence X-ray threat targets. *Visual Cognition*, *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. (2010). The impact of relative prevalence on dual-target search for threat items from airport X-ray screening. *Acta Psychologica*, *134*(1), 79–84. https://doi.org/10.1016/j.actpsy.2009.12.009

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. In *Wiley & Sons, Inc.* https://doi.org/10.1901/jeab.1969.12-475

Hackman, J. R., & Oldham, R. G. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, *60*(2), 159–170. https://doi.org/10.1037/h0076546

Hackman, J. R., & Oldham, R. G. (1980). *Work redesign*. Addison-Wesley.

Hardmeier, D., & Schwaninger, A. (2008). Visual cognition abilities in X-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT, June 2008*, 1–4. https://doi.org/10.13140/RG.2.1.4335.7924

Harris, D. H. (2002). How to really improve airport security. *Ergonomics in Design*, *10*(1), 17–22. https://doi.org/10.1177/106480460201000104

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? *Frontiers in Psychology*, *10*(MAR), 1–17. https://doi.org/10.3389/fpsyg.2019.00525

Helton, W. S. (2004). Validation of a short stress state questionnaire. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/154193120404801107

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

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *Proceedings - International Carnahan Conference on Security Technology*, *November 2004*, 303–308. https://doi.org/10.1109/ccst.2004.1405409

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. *WIT Transactions on the Built Environment*, *82*, 417–426. https://doi.org/10.2495/SAFE050411

Ishibashi, K., & Kita, S. (2014). Probability cueing influences miss rate and decision criterion in visual searches. *I-Perception*, *5*(3), 170–175. https://doi.org/10.1068/i0649rep

Ishibashi, K., Kita, S., & Wolfe, J. M. (2012). The effects of local prevalence and explicit expectations on search termination times. *Attention, Perception, and Psychophysics 74*(1), 115–123. https://doi.org/10.3758/s13414-011-0225-4

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. *Ophthalmology and Therapy*, *11*(5), 1655–1680. https://doi.org/10.1007/s40123-022-00540-9

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge.

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

Kopardekar, P., & Mital, A. (1994). The effect of different work-rest schedules on fatigue and performance of a simulated directory assistance operator's task. *Ergonomics*, *37*(10), 1697–1707. https://doi.org/10.1080/00140139408964946

Kuhn, M. (2017). Centralised image processing: The impact on security checkpoints. *Aviation Security International*, *23*(5), 28–30.

Kundel, H. L. (2000). Disease prevalence and the index of detectability: A survey of studies of lung cancer detection by chest radiography. In E. A. Krupinski (Ed.), Proc. *SPIE 3981, Medical Imaging 2000: Image Perception and Performance* (pp. 135–144). https://doi.org/10.1117/12.383100

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

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lim, J., & Kwok, K. (2016). The effects of varying break length on attention and time on task. *Human Factors*, *58*(3), 472–481. https://doi.org/10.1177/0018720815617395

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *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. *Psychological Science*, *21*(6), 829–839. https://doi.org/10.1177/0956797610371339

Macmillan, N. A., & Creelman, D. C. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. a, Huggins, J., Gilliland, K., Grier, R., & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry. *Emotion*, *2*(4), 315–340. https://doi.org/10.1037/1528-3542.2.4.315

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 diagnostic monitoring for sustained attention. *Journal of Experimental Psychology: Applied*, *16*(2), 187–203. https://doi.org/10.1037/a0019572

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science*, *15*(5), 302–306. https://doi.org/10.1111/j.0956-7976.2004.00673.x

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

Menneer, T., Cave, K. R., & Donnelly, N. (2009). The cost of search for multiple targets: Effects of practice and target similarity. *Journal of Experimental Psychology: Applied*, *15*(2), 125–139. https://doi.org/10.1037/a0015331

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. *Human Factors*, *58*(2), 218–228. https://doi.org/10.1177/0018720815616306

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., & Schwaninger, A. (2014). A multi-method approach towards identifying situational factors and their relevance for X-ray screening.

Proceedings of the 48thIEEE- International Carnahan Conference on Security Technology, Rome, Italy (2014), pp. 208–213. - https://doi.org/10.1109/CCST.2014.6987001

Mitroff, S. R., George, T., Ericson, J. M., Sharpe, B., & Company, K. (2018). Predicting airport screening officers' visual search competency with a rapid assessment. *Human Factors*, *60*(2), 201–211. https://doi.org/10.1177/0018720817743886

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Murphy, K. R., & Davidshofer, C. O. (2014). Psychological testing: Principles and applications (6th ed.). Pearson Education Limited.

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. *Theoretical Issues in Ergonomics Science*, *21*(2), 239–258. https://doi.org/10.1080/1463922X.2019.1682712

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569. https://doi.org/10.3758/BF03196517

Peltier, C., & Becker, M. W. (2020). Individual differences predict low prevalence visual search performance and sources of errors: An eye-tracking study. *Journal of Experimental Psychology: Applied*, *26*(4), 646–658. https://doi.org/10.1037/xap0000273

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1–12), 125–126. https://doi.org/10.3758/BF03342823

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ttps://www.r-project.org/

Riz à Porta, R., Sterchi, Y., & Schwaninger, A. (2022). How realistic is threat image projection for X-ray baggage screening? *Sensors*, *22*(6), 2220. https://doi.org/%0A10.3390/s22062220

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. *Journal of Experimental Psychology: Human Perception and Performance*, *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. *Frontiers in Human Neuroscience*, *9*(439). https://doi.org/10.3389/fnhum.2015.00439

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Schwaninger, A. (2006). Threat image projection: Enhancing performance? *Aviation Security International* (pp. 36–41).

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems*, *20*(6), 29–35. https://doi.org/10.1109/MAES.2005.1412124

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*(2), 230–249. https://doi.org/10.1201/9780203872512.ch23

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, *80*(6), 481–488. https://doi.org/10.1037/h0035203

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

Skorupski, J., & Uchroński, P. (2018). Evaluation of the effectiveness of an airport passenger and baggage security screening system. *Journal of Air Transport Management*, *66*, 53–64.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Spitz, G., & Drury, C. G. (1978). Inspection of sheet materials—test of model predictions. *Human Factors*, *20*(5), 521–528. https://doi.org/10.1177/001872087802000502

Stanislaw, H. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *3*, 37–149.

Steinborn, M. B., & Huestegge, L. (2016). A walk down the lane gives wings to your brain. Restorative benefits of rest breaks on cognition and self-control. *Applied Cognitive Psychology*, *30*(5), 795–805. https://doi.org/10.1002/acp.3255

Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (2019). Detection measures for visual inspection of X-ray images of passenger baggage. *Attention, Perception, and Psychophysics, 81*(5), 1297–1311. https://doi.org/10.3758/s13414-018-01654-8

Teichner, W. H. (1974). The detection of a simple visual signal as a function of time of watch. *Human 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. *Proceedings of the Human Factors and Ergonomics Society*, *January 2014*, 1180–1184. https://doi.org/10.1177/1071181311551246

Tiwari, T., Singh, A., & Singh, I. (2009). Task demand and workload: Effects on vigilance performance and stress. *Journal of the Indian Academy of Applied Psychology*, *35*(2), 265–275. http://medind.nic.in/jak/t09/i2/jakt09i2p265.pdf?origin=publication_detail

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tucker, P. (2003). The impact of rest breaks upon accident risk, fatigue and performance: A review. *Work & Stress*, *17*(2), 123–137. https://doi.org/10.1080/0267837031000155949

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). *Even in correctable search, some types of rare targets are frequently missed*. *71*(3), 541–553. https://doi.org/10.3758/APP.71.3.541

Wales, A. W. J., Anderson, C., Jones, K. L., Schwaninger, A., & Horne, J. A. (2009). Evaluating the two-component inspection model in a simplified luggage search task. *Behavior Research Methods*, *41*(3), 937–943. https://doi.org/10.3758/BRM.41.3.937

Walter, S., Hofer, F., Dolder, Z., & Ghelfi-Waechter, S. (2021). The why and how of security drills at the security checkpoint. *Journal of Airport Management*, *15*(2), 147–159.

Warm, J. S. (1984). *Sustained attention in human performance*. Wiley.

Warm, J. S., Matthews, G., & Finomore, V. S. (2008). Workload and stress in sustained attention. In In P.A. Hancock and J.L. Szalma (Eds.), *Performance under stress* (pp. 115–141). Ashgate Publishing.

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

Wetter, O. E. (2013). Imaging in airport security: Past, present, future, and the link to forensic and clinical radiology. *Journal of Forensic Radiology and Imaging*, *1*(4), 152–160. https://doi.org/10.1016/j.jofri.2013.07.002

Wetter, O. E., Hardmeier, D., & Hofer, F. (2008). Covert testing at airports. *Proceedings - International Carnahan Conference on Security Technology*, 357–363.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

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

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

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

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

Acknowledgments

- First and foremost, I would like to thank Adrian Schwaninger for granting me the opportunity to embark on this academic endeavor, for his valuable insights, for sharing his expertise and knowledge. His infectious enthusiasm has played a crucial role in driving me forward.
- I am very grateful to Klaus Opwis, my thesis supervisor, for enabling the realization of this thesis, for providing insightful feedback on my work, for his support, and for his expertise that broadened my perspectives.
- I would like to express my appreciation and gratitude to Yanik Sterchi for his valuable contributions, his continuous support, and for his guidance and mentorship throughout this journey.
- Immense thanks go to my research team, who have collaborated with me on various projects and studies. Their contributions have been invaluable in achieving the goals of this research.
- I thank the doctoral committee for evaluating this work.
- I thank my parents for always supporting me, for their guidance and motivation, and, most of all, for always believing in me.
- Last but not least, thank you Joris Jessen, for your patience and mental support.

Complete Manuscripts

Buser, D., Sterchi, Y., & Schwaninger, A. (2020). Why stop after 20 min? Breaks and target prevalence in a 60-min X-ray baggage screening task. *International Journal of Industrial Ergonomics*, *76*, 102897. https://doi.org/10.1016/j.ergon.2019.102897

Buser, D., Schwaninger, A., Sauer, J., & Sterchi, Y. (2023). Time on task and task load in visual inspection: a four-month field study with X-ray baggage screeners. *Applied Ergonomics*, *111*, 103995. https://doi.org/10.1016/j.apergo.2023.103995

Buser, D., Schwaninger, A., Rehor, V., & Sterchi, Y. (under review). Reliability and validity of threat image projection data on X-ray baggage screening.

International Journal of Industrial Ergonomics 76 (2020) 102897

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

journal homepage: http://www.elsevier.com/locate/ergon

Why stop after 20 minutes? Breaks and target prevalence in a 60-minute X-ray baggage screening task

Daniela Buser [*,1], Yanik Sterchi [1], Adrian Schwaninger

School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland

ARTICLE INFO

Keywords:
X-ray image inspection
Visual search
Time on task
Breaks
Detection measures
Target prevalence effect

ABSTRACT

Current EU regulation restricts continuously reviewing X-ray images of passenger baggage to 20-min duration as a precautionary measure to prevent performance decrements in airport security officers (screeners). However, this 20-min limit is not based on clear empirical evidence on how well screeners can sustain their performance over time. Our study tested screeners in a 60-min simulated X-ray cabin baggage screening task. One group took 10-min breaks after 20 min of screening; the other group worked without breaks. We found no decrease in performance over 60 min in either group. Breaks did not affect performance, but they did reduce the amount of subjective distress. By varying target prevalence, we found that d_a with a slope of about 0.6 is a more valid measure of detection performance than d'. Target prevalence caused a criterion shift. Our results provide a basis for conducting field studies of prolonged screening durations, and open the discussion on whether more flexible break policies and work schedules should be considered.

1. Introduction

Throughout the world, X-ray technology is used to scan passenger baggage at airports, and security officers (screeners) inspect the X-ray images. The current European regulation defines a maximum of 20 min of continuously reviewing X-ray images as a precautionary measure to prevent any decrease in detection performance of screeners (European Commission, 2015). Therefore, after 20 min of screening passenger baggage, screeners usually rotate to another position at the airport security checkpoint where they carry out other tasks such as assisting passengers with divesting, alarm resolution of the walk-through metal detector or person scanner, and secondary bag search (Michel et al., 2014). A new technology called *remote cabin baggage screening* (RCBS), which is being employed increasingly by airports, creates operational challenges for this 20-min rule. With RCBS, security personnel visually inspect X-ray images in an office-like environment separated from the checkpoint. RCBS allows for a higher utilization of X-ray machines and screeners while also providing a quieter workplace for X-ray screeners without the distractors at the checkpoint (Kuhn, 2017). However, relocating image inspection away from the checkpoint into a remote room makes rotating between X-ray image inspection and other tasks at the checkpoint more costly and difficult to coordinate. One way to alleviate such concerns would be to introduce screening durations longer than 20 min. Our study investigated how performance changes over time (i.e., as a function of time on task) by instructing screeners to review X-ray images continuously for 60 min without breaks and comparing their performance with screeners in another condition who took 10-min breaks after each 20 min of screening. The following sections summarize previous research on X-ray screening, performance over time, breaks, and the measurement of screener performance.

1.1. Research on X-ray screening

To prevent passengers from carrying prohibited articles (guns, explosives, knives, etc.) onto an airplane, passenger bags are screened at airport security checkpoints using X-ray machines (Harris, 2002). By visually inspecting these X-ray images, screeners are engaging in visual search and decision making (Koller et al., 2009; McCarley et al., 2004; Wales et al., 2009). Visually searching for prohibited articles among distractors in X-ray images is a cognitively demanding task (for recent reviews, see Biggs et al., 2018; Biggs and Mitroff, 2014). Moreover, X-ray image inspection requires different visual cognitive abilities to

* Corresponding author. University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Riggenbachstrasse 16, CH-4600, Olten, Switzerland.
E-mail address: daniela.buser@fhnw.ch (D. Buser).
[1] Joint first authorship between Daniela Buser and Yanik Sterchi.

https://doi.org/10.1016/j.ergon.2019.102897
Received 17 December 2018; Received in revised form 20 September 2019; Accepted 3 December 2019
Available online 22 January 2020

those in traditional visual search (Hättenschwiler et al., 2019). To prevent a decrease in performance, European regulation limits continuously reviewing X-ray images to 20 min (European Commision, 2015). The introduction of this limitation more than two decades ago was probably based on research into vigilance that had been based on other tasks than X-ray image inspection (personal communication with airport security expert, March 2019). To our knowledge, only two published studies have investigated effects of time on task in X-ray baggage screening with professional screeners. Another study (Chavaillaz et al., 2019) investigated time on task with a student sample and will be discussed in a later paragraph on breaks.

Meuter and Lacherez (2016) examined screener performance in the field for durations up to 30 min. They analyzed 4 months of threat image projection (TIP) data from an Australian airport. TIP is a technology that projects prerecorded X-ray images of prohibited articles onto real X-ray images of passenger bags during baggage screening at airports (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Skorupski and Uchroński, 2016). The TIP hit rate (or percent detected) refers to the proportion of projected fictional threat items that screeners detect. Meuter and Lacherez (2016) found a small decrease of approximately 2 percentage points in the hit rate with time on task when workload was high (operationalized as more than 5.4 X-ray images screened per min). No decrease in performance was found when workload was low. A closer examination of high workload sessions revealed that performance started to decrease after 10 min. Although this is a very interesting and valuable study, there are some limitations regarding its practical implications. First, the observed decrease in the hit rate was very small. Because only screening durations up to 30 min were examined, it is unclear how performance would evolve over longer screening durations. It should also be noted that Meuter and Lacherez (2016) analyzed data from conventional airport security checkpoints. It is therefore unclear whether the results would also apply to RCBS in which screeners work in an office-like environment with much less noise and distraction (Kuhn, 2017). Another limitation of their study is that they were unable to analyze screeners' false alarm rate with the TIP system and the screening process at the tested airport. When measuring only the hit rate, one cannot determine whether an observed change in hit rate is due to a change in response tendency and/or whether it reflects a change in detection performance in terms of sensitivity (Green and Swets, 1966; Macmillan and Creelman, 2005).

Ghylin et al. (2007) examined effects of time on task on hit rates, false alarm rates, and sensitivity for longer screening durations. In their study, airport security screeners completed a simulated X-ray cabin baggage screening task over the course of 4 h. Results were aggregated for each of the 4 h. The authors found a decrease in the hit rate and false alarm rate over time, but no significant change in the sensitivity measure *A'* (Pollack and Norman, 1964). This suggests a shift in response tendency (i.e., a criterion shift as defined in signal detection theory, Green and Swets, 1966; Macmillan and Creelman, 2005). Also reaction times decreased over time. Ghylin et al. (2007) concluded that vigilance decrements occurred—a conclusion that we shall address in the next section. Whereas this study provides very interesting results, it compared only full hours and did not report on whether and how hit rate, false alarm rate, sensitivity, and response tendency change within 1 h. This limits the derivation of conclusions regarding performance changes within the first hour of screening.

1.2. Research on vigilance

The effect of time on task has been investigated quite extensively for vigilance tasks that share some similarities with X-ray image inspection. Both are characterized by long search periods and require the searcher to stay alert to few targets appearing (Davies and Parasuraman, 1982). In both tasks, the infrequent appearance of targets causes more misses (Wolfe et al., 2007). For difficult vigilance tasks, performance decrements can already be observed after as early as 5 min (Nuechterlein et al., 1983; Rose et al., 2002). Most studies have revealed decreases in vigilance within the first 15–30 min of the task (Mackworth, 1948; Teichner, 1974; Warm, 1984). Nonetheless, it is not clear whether the performance decrement within the first 15–30 min often found in vigilance tasks can also be expected for X-ray image inspection in RCBS, because the tasks differ in certain aspects. In vigilance tasks, a short distraction can lead to missing a target, whereas in an X-ray image inspection in RCBS, screeners have to actively declare that no target is present in an image, as is the case in experiments on X-ray image inspection and visual search (e.g., Koller et al., 2009; McCarley et al., 2004). Traditional vigilance tasks use a single target, whereas X-ray image inspection entails a visual search for multiple targets (Godwin et al., 2010; Mitroff et al., 2015). Moreover, in X-ray image inspection, certain types of targets are very rare (e.g., bombs), whereas other targets occur more frequently in carry-on baggage (e.g., liquids and gels).

1.3. Breaks and performance

Further insight into the effect that time on task has on performance in detection tasks can be gained from research on the effect of breaks. Several studies have reported mainly positive effects of breaks on performance in a variety of different detection tasks (Arrabito et al., 2015; Colquhoun, 1959; Kopardekar and Mital, 1994). Breaks have been found to decrease perceived workload (Arrabito et al., 2015). In a different task, breaks reduced perceived fatigue and discomfort (Galinsky et al., 2000). However, positive effects of the frequency and length of breaks depend on the type, difficulty, and duration of the task (Tucker, 2003). To our knowledge, only one study investigated different types of breaks in X-ray image inspection of passenger bags (Chavaillaz et al., 2019). Student participants were tested in a simulated X-ray baggage screening task during 1 h with and without adaptable automation as support system. They could either take spontaneous breaks, 5 min breaks every 20 min, or 10 min breaks every 20 min of continuous X-ray image inspection. No performance differences between break regimes were found, which suggests that a more flexible regulation on breaks providing more autonomy could be considered. However, a limitation of this study is that it was conducted with student participants and a task adapted to students. It remains unclear whether professional screeners (airport security officers) can maintain their performance over 60 min of continuous X-ray image inspection without breaks.

1.4. Measuring performance in X-ray image inspection

Some challenges emerge when investigating screener performance. Common measures for X-ray image inspection are the hit rate (HR, the percentage of prohibited items detected) and the false alarm rate (FAR, percentage of harmless baggage falsely sent to secondary search). Because the hit rate and false alarm rate depend on the response tendency, it is recommended to use detection measures that are considered to be independent of response tendency (Macmillan and Creelman, 2005). Signal detection theory (Green and Swets, 1966) provides a general framework for defining detection performance, called *sensitivity*, that is independent from response tendency, called *criterion*. Research in X-ray image inspection often uses *d'* as such a measure of sensitivity (for a recent discussion see Sterchi et al., 2019). This is calculated as follows:

$$d' = z(HR) - z(FAR)$$

whereby z is the inverse of the cumulative distribution function of the standard normal distribution (Green and Swets, 1966). The resulting measure of response tendency, the criterion c, is calculated as

$$c = -0.5\,[z(HR) + z(FAR)].$$

However, recent research has questioned the validity of *d'* for X-ray image inspection. Several studies examining the effect that the hit rate decreases when targets are rare—the so-called *target prevalence*

effect—have found that *d'* increases as targets become less frequent (Godwin et al., 2010; Wolfe et al., 2007; Wolfe and Van Wert, 2010). This is paradoxical, especially when it is considered that response times are usually faster when targets appear infrequently (low prevalence) compared to when they are frequent (high prevalence). Moreover, signal detection theory assumes that target prevalence affects only the criterion and not the sensitivity (Green and Swets, 1966). Instead of assuming that sensitivity actually increases when target prevalence decreases, Wolfe et al. (2007) have argued that visual search in X-ray images does not fulfil the assumptions that underlie *d'*. Signal detection theory assumes a decision process in which target-present and target-absent trials result in Gaussian distributions of some measure of evidence for the presence of a target, and *d'* assumes that these distributions have equal variance (Macmillan and Creelman, 2005). If this assumption of equal variance is not met, one can use d_a, an index of detectability proposed by Simpson and Fitter (1973). This offers an extension of *d'* with the slope *s* as an additional open parameter that is the ratio of the two standard deviations:

$$d_a = \sqrt{\frac{2}{1+s^2}}\,[z(HR) - sz(FAR)].$$

The corresponding measure of response tendency, the criterion c_a, is calculated as

$$c_a = \frac{-\sqrt{2}\,s}{\sqrt{(1+s^2)}\,(1+s)}\,[z(HR) + z(FAR)\,]$$

whereby *z* is the inverse of the cumulative distribution function of the standard normal distribution.

Wolfe et al. (2007) have argued that d_a is more appropriate (in line with Kundel, 2000, who found the same target prevalence effect for the inspection of medical X-ray images), and they estimated the slope parameter to be around 0.6 (again in line with Kundel, 2000). Following this approach, several other studies have found the slope parameter to be around 0.6 when investigating the effect of target prevalence in X-ray image inspection (Godwin et al., 2010; Van Wert, Horowitz and Wolfe, 2009; Wolfe and Van Wert, 2010). Consistent with these findings, Sterchi et al. (2019) have found slope parameters around 0.6 based on an experiment manipulating the criterion through instruction and another experiment using confidence ratings.

1.5. Present study

This study investigated the effects of time on task and breaks on screener performance when X-ray images were analyzed for 60 min. Whereas one group screened for 60 min continuously, the other group took 10-min breaks between 20-min screening blocks. Based on current evidence, we cannot formulate clear hypotheses on performance decrements or perceived stress depending on time on task over a period of 60 min. Perceived stress was monitored by asking screeners to complete the Short Stress State Questionnaire (SSSQ; Helton, 2004).

In order to measure screener performance independently from response tendency, we varied target prevalence in order to determine which detection measure is valid when analyzing the effect that time on task has on detection performance. In line with previous research (Godwin et al., 2010; Sterchi et al., 2019; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe and Van Wert, 2010), we assumed that *d'*, which implies a slope parameter of 1, would be an invalid measure of detection performance for this task and might be affected by target prevalence. We expected the slope parameter to be around 0.6, and that d_a based on that slope would be more appropriate.

2. Methods

2.1. Participants

A total of 71 screeners working at a European airport completed the study (four additional participants were unable to attend the second test date and were therefore excluded from analyses). All had been recruited by the airport's security service provider and participated during their regular working hours. Screeners were aged between 20 and 67 years ($M = 32.01$, $SD = 12.82$), had 0.3–12 years of working experience ($M = 2.08$, $SD = 2.23$), and 46% of them were female.[2] The study complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board of the School of Applied Psychology of the University of Applied Sciences and Arts Northwestern Switzerland. Informed consent was obtained from all screeners prior to their participation.

2.2. Design

A 3 (time on task*: 0–20 min, 20–40 min, 40–60 min*; within-subject factor) × 2 (breaks condition: *with breaks, without breaks*; between-subject factor) × 2 (prevalence condition: *high prevalence, low prevalence*; within-subject factor) mixed factorial design was employed. To analyze the effect of time on task, the 60-min X-ray baggage screening task was split into three 20-min screening blocks: *0–20 min*, *20–40 min*, and *40–60 min*. Participants were divided into two groups: The group *with breaks* had a 10-min break after each 20-min screening block; the group *without breaks* screened for 60 min without breaks. All screeners completed the task twice, once in the *low prevalence* condition and once in the *high prevalence* condition. The order of the two prevalence conditions was counterbalanced across subjects.

The following performance measures served as dependent variables: hit rate, false alarm rate, sensitivity (*d'*, d_a), criterion (*c*, c_a), and processing time. We also investigated the influence of the breaks condition and prevalence condition on the three factors of the SSSQ (distress, worry, and engagement; Helton, 2004).

2.3. Materials

For the experiment, 864 single view X-ray images of passenger cabin (carry-on) baggage were used to create a simulated X-ray baggage screening task. For a subset of the images, prohibited items were merged into the bags using a validated X-ray image merging algorithm (Mendes et al., 2011). Prohibited items belonged to one of three categories: guns, knives, and improvised explosive devices (IEDs). Each image contained a maximum of one prohibited item. To create enough content for the task, each image of a passenger bag and each prohibited item was used twice. For the passenger baggage, one of the two images was presented in a mirrored version in order to reduce recognition. For prohibited items, both an easy and a difficult rotation (as defined by X-ray image inspection experts) of each prohibited item was projected into different bag images. Fig. 1 shows four X-ray images from two bags as examples. The complexity of the bag images and the superposition of the prohibited items, which are both known to affect difficulty in detecting the prohibited item (Bolfing et al., 2008; Hardmeier et al., 2005), were held at a medium level and not varied systematically.

In the high prevalence condition, one out of two bags (50%) contained a prohibited item. A target prevalence of 50% is typically employed by studies investigating target prevalence effects (e.g., Godwin et al., 2010; Ishibashi et al., 2012; Wolfe and Van Wert, 2010). Furthermore, it matches the prevalence of the screeners' training (Koller et al., 2008; Schwaninger, 2004). In the low prevalence condition, one out of eight bags (12.5%) contained a prohibited item. Although this was

[2] Two participants did not report their demographics.

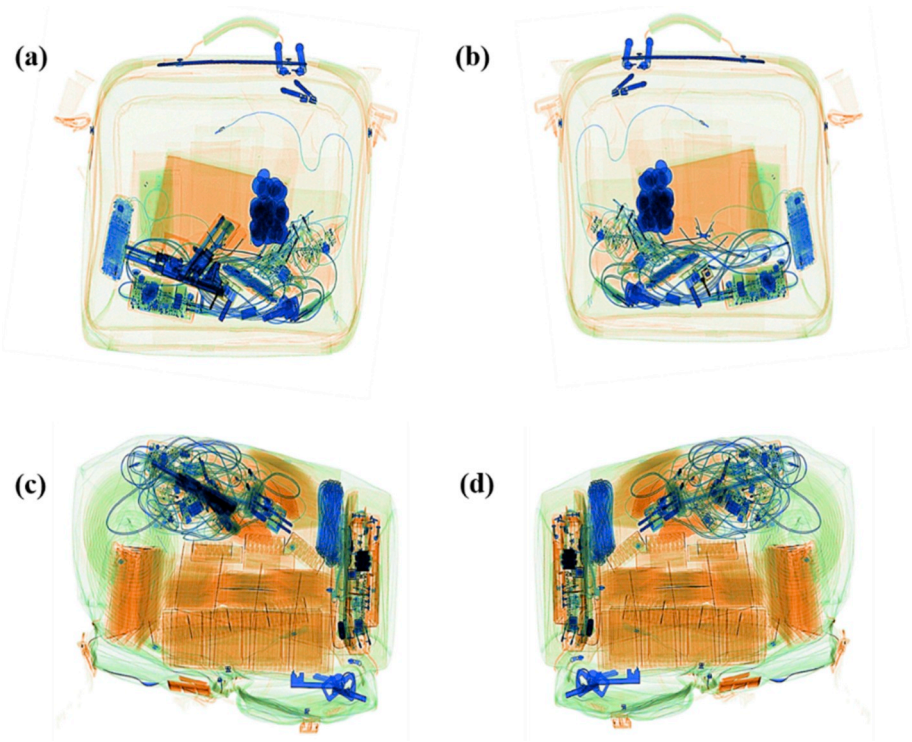


Fig. 1. Examples of X-ray images of passenger bags. (a) Bag with a prohibited item (gun) in easy rotation, (b) mirror reversed bag without prohibited item, (c) different bag with the same gun in difficult rotation, (d) mirror reversed bag without prohibited item.

higher than in practice, it was necessary in order to collect enough target-present trials within the experiment to calculate reliable hit rates.

To allow all participants to be compared on the basis of the same images, regardless of their processing time and the total number of images analyzed during the task, the system automatically synchronized the progress between all participants every 5 min. Hence, the total of 864 images was split into 12 sets of 72 images. Because the images of each set were in a fixed order and progress was synchronized, all participants inspected at least the first 24 images of each set, and these first 24 images were then used to calculate performance. The order of the 12 image sets was counterbalanced across participants with a Latin square design ensuring that the difficulty did not vary systematically with time on task.

We measured perceived stress levels with the SSSQ (Helton, 2004). This 24-item questionnaire is a valid measure of task-related stress. It taps three different factors of stress: *distress*, *worry,* and *engagement*. The three factors address the motivational, cognitive, and affective aspects of task-related stress: *Engagement* refers to the willingness to act; *worry*, to self-regulation; and *distress*, to negative emotions. Items were rated on 5-point scales ranging from 1 (*not at all*) to 5 (*extremely*).

2.4. Procedure

Testing sessions took place in a normally lit room at the airport using single view X-ray machine simulators on HP ProOne 400 computer workstations with 20-inch (50 cm) TFT monitors and a screen resolution of 1600 × 900 pixels. Each screener sat approximately 50 cm away from the monitor. The X-ray images covered about two-thirds of the computer screen. Six to twelve participants performed the task in each session while working individually, quietly, and under supervision. This is a typical working condition in RCBS (Kuhn, 2017). Screeners were randomly assigned to either the group *with breaks* or the group *without breaks*. Each participant completed the task twice, once with *low prevalence* and once with *high prevalence*. The order of the prevalence conditions was counterbalanced across participants. The two test sessions were separated by an interval of 3–5 weeks.

Each test session lasted about 1.5 h. Screeners were informed about the test procedure and instructed to analyze images as quickly and accurately as possible as if they were working. Because screeners are used to a target prevalence of 50% in training and certification, instructions also informed them about the target prevalence to avoid confusion. Screeners had to press a button labeled *OK* if they perceived an image as harmless. If they thought the image contained a prohibited item, they had to locate the prohibited item by double clicking on it (marking); select whether it was a *gun, knife,* or *IED* (categorizing); and then press a button labeled *NOT OK*. Feedback was given in the same manner as that provided by the TIP system operational at their airport: immediate feedback for images containing a prohibited item informing about the correctness of the final decision between *OK* and *NOT OK*, the marking, and the categorizing. Screeners did not receive feedback if the image did not contain a prohibited item.

After the instructions, screeners completed practice trials containing 16 images to familiarize themselves with the simulator interface and the procedure. They first completed the 16 trials without time limit, and then repeated the same trials with the 12-s limit per image used in the actual task. This time limit was employed to match the time limit per image for X-ray screening at this airport. Screeners then completed 60 min of X-ray image inspection. The group *with breaks* had a 10-min break after each 20 min of screening, whereas the group *without breaks* analyzed X-ray images for 60 min continuously and had a 20-min break thereafter. After completing the X-ray baggage screening task, screeners filled out the SSSQ and provided information on their shift schedule, work experience, age, and gender.

2.5. Analyses

To ensure that the same images were used to measure performance in all participants, only responses for the first 24 images of each of the 12 image sets (as explained in section 2.3 Materials) and only for images that appeared in both the high- and low-prevalence conditions were

analyzed. For the calculation of the dependent variables, responses were aggregated for each of the three 20-min screening blocks and each participant separately.

The hit rate was calculated as the share of images correctly declared as *NOT OK* and the false alarm rate as the share of images wrongly declared as *NOT OK* without taking marking and categorizing into account. This corresponds to operations at the checkpoint where all bags declared as *NOT OK* are sent to secondary search. The detection measures d' and d_a as well as the slope parameter were based on the *z*-transformed hit rate and false alarm rate, whereby *z* refers to the inverse of the cumulative distribution function of the standard normal distribution (Green and Swets, 1966). Because this function is undefined for extreme proportions (e.g., a hit rate of one or false alarm rate of zero), the hit rate and false alarm rate were corrected with the log-linear rule (Hautus, 1995) when calculating d', d_a, and the slope parameter. The slope parameter was estimated by calculating the difference in *z*-transformed hit rate and false alarm rate between the two target prevalence conditions for each participant. The average slope parameter then corresponded to the average difference in *z*-transformed hit rate divided by the difference in *z*-transformed false alarm rate. For the slope estimation, we report bootstrapped BCa-CIs (Efron, 1987) based on 20,000 resamples. The processing time refers to the time from image appearance until the *OK* or *NOT OK* button was pressed. For images with a prohibited item, this included the marking and categorizing of the prohibited item. This processing time is therefore not directly comparable to conventional reaction times.

All ANOVAs were carried out in R version 3.5.1 (R Core Team, 2018). The Greenhouse–Geisser correction (Greenhouse and Geisser, 1959) was used where applicable and effect sizes are reported with η_p^2 (partial eta squared). In case of significant effects of time on task, post hoc analyses were calculated comparing the first screening block (*0*–20 min) with the second screening block (*20*–40 min) and the second screening block with the third screening block (*40*–60 min). In case of no significant interactions with target prevalence, both prevalence conditions were averaged for each participant. In case of significant interactions with target prevalence, these comparisons were calculated for both levels of target prevalence separately. Post hoc tests were Holm–Bonferroni corrected (Holm, 1979).

3. Results

We report results on the hit rate and false alarm rate, the sensitivity (d' and d_a) and the criterion (c and c_a), the processing time, and the three factors of the SSSQ. We computed 2 (*with breaks* and *without breaks*) × 2 (*high prevalence* and *low prevalence*) × 3 (*0*–20 min, *20*–40 min, *40*–60 min) ANOVAs with hit rate, false alarm rate, d', d_a, c, c_a, and processing time as dependent variables.

3.1. Hit rate and false alarm rate

Fig. 2 shows the hit rate and false alarm rate for the two groups *with breaks* and *without breaks* in both prevalence conditions as a function of time on task. The ANOVA for the hit rate revealed a significant main effect of prevalence, $F(1, 69) = 37.99$, $p < .001$, $\eta_p^2 = .36$; no effect of time on task, $F(1.93, 133.17) = 1.78$, $p = .174$, $\eta_p^2 = .03$; and no effect of breaks, $F(1, 69) = 1.84$, $p = .180$, $\eta_p^2 = .03$. None of the two-way interactions were significant: Breaks × Prevalence, $F(1, 69) = 0.25$, $p = .621$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.93, 133.17) = 1.75$, $p = .179$, $\eta_p^2 = .02$; and Prevalence × Time on task, $F(1.96, 134.94) = 3.06$, $p = .051$, $\eta_p^2 = .04$. The three-way interaction was also not significant, $F(1.96, 134.94) = 0.31$, $p = .731$, $\eta_p^2 = .00$.

The ANOVA with false alarm rate as dependent variable revealed a significant main effect of prevalence, $F(1, 69) = 118.53$, $p < .001$, $\eta_p^2 = .63$; no effect of time on task, $F(1.87, 129.37) = 0.24$, $p = .776$, $\eta_p^2 = .00$; and no effect of breaks, $F(1, 69) = 0.00$, $p = .957$, $\eta_p^2 = .00$. The two-way interaction between Prevalence × Time on task was significant, $F(1.97, 136.18) = 17.9$, $p < .001$, $\eta_p^2 = .21$. No other interactions attained significance: Breaks × Prevalence, $F(1, 69) = 0.01$, $p = .917$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.87, 129.37) = 1.23$, $p = .294$, $\eta_p^2 = .02$; Breaks × Prevalence × Time on task $F(1.97, 136.18) = 0.30$, $p = .737$, $\eta_p^2 = .00$. Post hoc analyses for the significant interaction of Prevalence × Time on task revealed a significant increase in the false alarm rate from *0*–20 min to *20*–40 min in the high-prevalence condition ($p = .004$) and a significant decrease from *0*–20 min to *20*–40 min in the low-prevalence condition ($p = .004$). No significant difference was found between *20*–40 min and *40*–60 min in either the high-prevalence ($p = .811$) or low-prevalence condition ($p = .649$).

3.2. Sensitivity and criterion

Fig. 3 shows detection performance in terms of sensitivity d' and d_a for both break and prevalence conditions as a function of time on task. The ANOVA with d' as a dependent variable revealed a significant main effect of prevalence, $F(1, 69) = 12.83$, $p < .001$, $\eta_p^2 = .16$; a significant main effect of time on task $F(1.99, 136.97) = 3.62$, $p = .030$, $\eta_p^2 = .05$; and no effect of breaks, $F(1, 69) = .80$, $p = .375$, $\eta_p^2 = .01$. All interactions were nonsignificant: Prevalence × Time on task, $F(1.93, 133.34) = 1.08$, $p = .340$, $\eta_p^2 = .02$; Breaks × Prevalence, $F(1, 69) = 0.15$, $p = .697$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.99, 136.97) = 2.77$, $p = .067$, $\eta_p^2 = .04$; and Breaks × Prevalence × Time on task, $F(1.93, 133.34) = 1.83$, $p = .166$, $\eta_p^2 = .03$. Post hoc analyses for the main effect of time on task in d' revealed a significant increase from *0*–20 min to *20*–40 min ($p = .039$), but no significant difference between *20*–40 min and *40*–60 min ($p = .856$).

The estimated slope parameter was 0.65 (95% BCa-CI [0.41, 0.89]) and thereby lower than the slope of 1.0 assumed by d'. Fig. 3b shows the sensitivity measure d_a based on this slope estimation as a function of time on task. The ANOVA for d_a revealed a main effect of time on task, $F(1.97, 135.91) = 3.43$, $p = .036$, $\eta_p^2 = .05$; no significant main effects of prevalence, $F(1, 69) = 0.65$, $p = .423$, $\eta_p^2 = .01$ (whereby the main effect of prevalence has no informative value, because this main effect was used to estimate the slope parameter); or breaks, $F(1, 69) = 1.03$, $p = .314$, $\eta_p^2 = .01$. No interaction attained significance: Breaks × Prevalence, $F(1, 69) = 0.22$, $p = .638$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.97, 135.91) = 2.49$, $p = .088$, $\eta_p^2 = .03$; Prevalence × Time on task, $F(1.95, 134.72) = 0.11$, $p = .895$, $\eta_p^2 = .00$; and Breaks × Prevalence × Time on task, $F(1.95, 134.72) = 1.53$, $p = .221$, $\eta_p^2 = .02$. Post hoc analyses for the main effect of time on task in d_a revealed a significant increase from *0*–20 min to *20*–40 min ($p = .034$), but no significant difference between *20*–40 min and *40*–60 min ($p = .754$).

Fig. 4 displays the criterion measures c and c_a for both break and prevalence conditions as a function of time on task. In accordance with calculating d_a, a slope of 0.65 was used to determine the criterion c_a. Because c_a is a linear transformation of c that does not affect significance testing, ANOVA and post hoc results were identical for both c and c_a and are therefore reported only once. The ANOVA with c and c_a as a dependent variable revealed a significant main effect of prevalence, $F(1, 69) = 141.58$, $p < .001$, $\eta_p^2 = .67$; but no effect of time on task, $F(1.93, 133.02) = 0.40$, $p = .665$, $\eta_p^2 = .01$; or breaks, $F(1, 69) = 0.96$, $p = .329$, $\eta_p^2 = .01$. The interaction Prevalence × Time on task, $F(1.95, 134.28) = 11.82$, $p < .001$, $\eta_p^2 = .15$, was significant. No significant effects were found for Breaks × Prevalence, $F(1, 69) = .25$, $p = .619$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.93, 133.02) = .24$, $p = .782$, $\eta_p^2 = .00$; or Breaks × Prevalence × Time on task, $F(1.95, 134.28) = 0.02$, p = .977, $\eta_p^2 = .00$. Post hoc analyses for the significant interaction of Prevalence × Time on task revealed a significant increase in c and c_a from *0*–20 min to *20*–40 min for high prevalence ($p = .002$) and a significant decrease in c and c_a *for* low prevalence ($p = .038$). The criterion (c and c_a) did not change significantly from *20*–40 min to *40*–60 min for either high prevalence ($p = .995$) or low prevalence ($p = .995$).

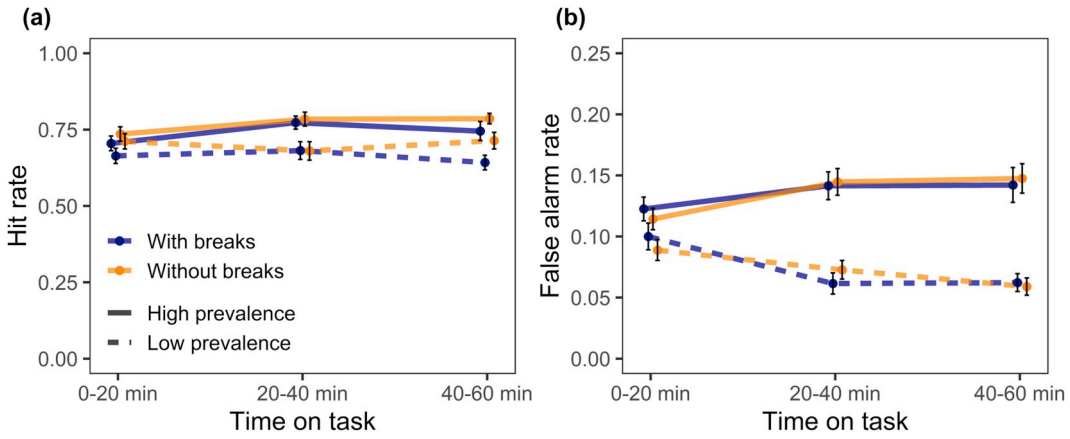


Fig. 2. Mean hit rate (a) and false alarm rate (b) for both break and prevalence conditions as a function of time on task. Error bars represent standard errors.

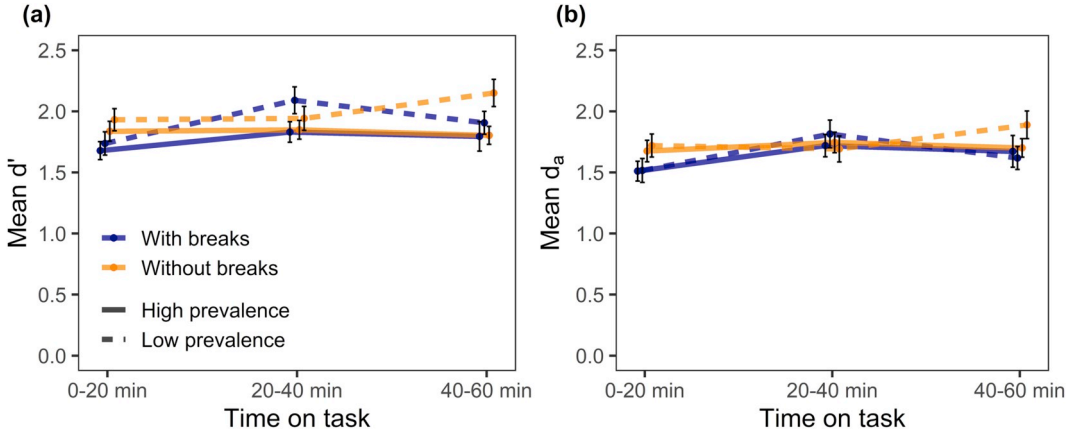


Fig. 3. Mean sensitivity measure *d'* (a) and sensitivity measure d_a (b) for both break and prevalence conditions as a function of time on task. Error bars represent standard errors.

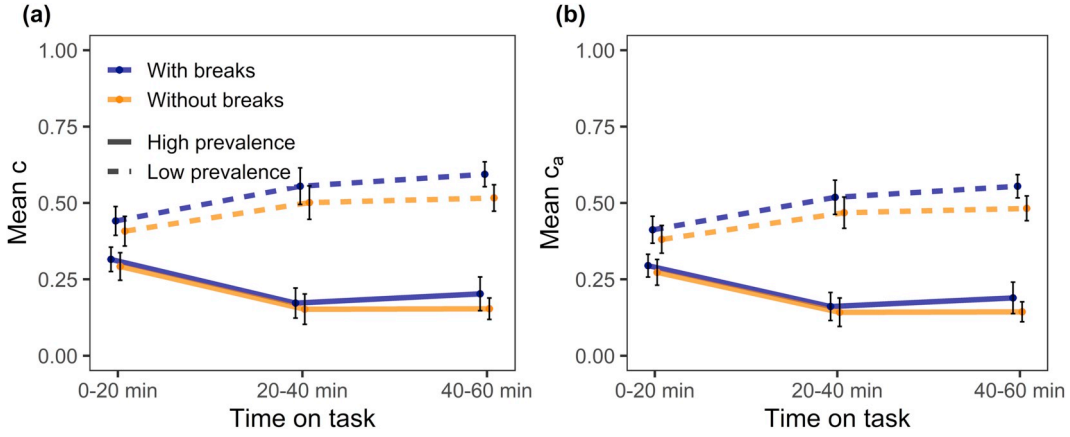


Fig. 4. Mean criterion measures *c* (a) and c_a (b) for both break and prevalence conditions as a function of time on task. Error bars represent standard errors.

3.3. Processing time

Fig. 5 shows screeners' processing time for target-absent and target-present trials for both break and prevalence conditions as a function of time on task. The ANOVA for target-absent trials revealed a significant main effect of prevalence, $F(1, 69) = 89.01$, $p < .001$, $\eta_p^2 = .56$; and time on task, $F(1.69, 116.40) = 127.51$, $p < .001$, $\eta_p^2 = .65$. The main effect of breaks was not significant, $F(1, 69) = 1.27$, $p = .264$, $\eta_p^2 = .02$. The interaction Prevalence × Time on task was significant, $F(1.54, 105.92) = 37.07$, $p < .001$, $\eta_p^2 = .35$. The other interactions did not attain significance, Breaks × Prevalence, $F(1, 69) = 0.34$, $p = .560$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.69, 116.40) = 2.93$, $p = .066$, $\eta_p^2 = .04$; Breaks × Prevalence × Time on task, $F(1.54, 105.92) = .53$, $p = .543$, $\eta_p^2 = .01$. Post hoc tests for the interaction of Prevalence x Time on task revealed a significant decrease from *0*–20 min to *20*–40 min for the high prevalence ($p = .010$) and low prevalence condition ($p < .001$). The decrease was also significant from *20*–40 min to *40*–60 min for the high prevalence ($p = .001$) and the low prevalence condition ($p < .001$).

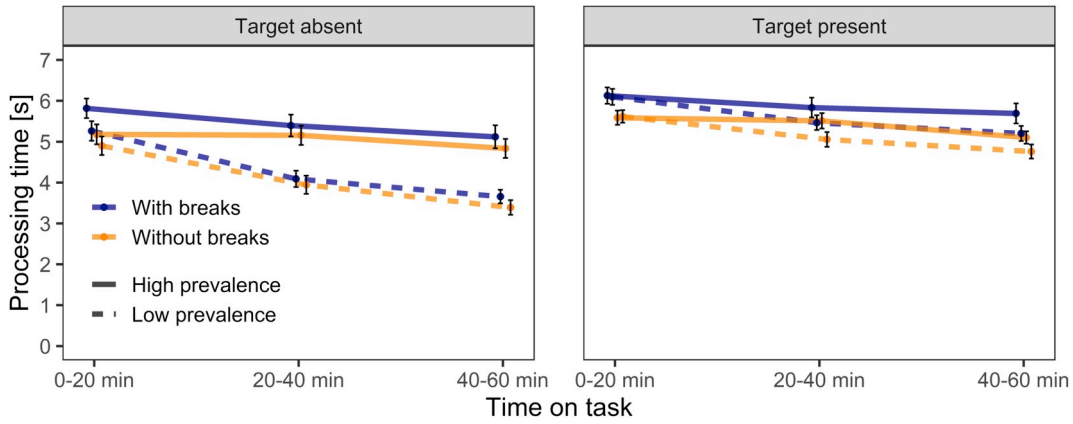


Fig. 5. Mean processing time for target-absent and target-present trials for both break and prevalence conditions as a function of time on task. Error bars represent standard errors.

For target-present trials, the ANOVA revealed significant main effects of prevalence, $F(1, 69) = 6.67$, $p = .012$, $\eta_p^2 = .09$; time on task, $F(1.97, 136.22) = 26.98$, $p < .001$, $\eta_p^2 = .28$; and breaks, $F(1, 69) = 5.28$, $p = .025$, $\eta_p^2 = .07$. The interaction Prevalence × Time on task also attained significance, $F(1.97, 135.76) = 4.61$, $p = .012$, $\eta_p^2 = .06$. All other interactions were not significant, Breaks × Prevalence, $F(1, 69) = 0.04$, $p = .851$, $\eta_p^2 = .00$; Breaks × Time on task, $F(1.97, 136.22) = .43$, $p = .649$, $\eta_p^2 = .01$; and Breaks × Prevalence × Time on task, $F(1.97, 135.76) = 0.29$, $p = .745$, $\eta_p^2 = .00$. Post hoc tests for the significant interaction of Prevalence × Time on task revealed no significant difference in reaction time between *0*–20 min and *20*–40 min for high prevalence ($p = .161$) but a significant decrease from *0*–20 min to *20*–40 min for the low prevalence condition ($p < .001$). Again, between *20*–40 min and *40*–60 min, there was no significant decrease for the high prevalence condition ($p = .071$) but a significant decrease from *0*–20 min to *20*–40 min for the low prevalence condition ($p = .016$).

3.4. Subjective measures of distress, worry, and engagement

Fig. 6 shows the reported levels of *distress, worry*, and *engagement* for both break and prevalence conditions.[3] We calculated separate 2 (with vs. without breaks) × 2 (high vs. low prevalence) ANOVAs for each of the three measures of subjective stress. For *distress*, the ANOVA revealed a significant main effect of breaks, $F(1, 66) = 9.17$, $p = .004$, $\eta_p^2 = .12$.[4] The main effect of prevalence, $F(1, 66) = 1.44$, $p = .234$, $\eta_p^2 = .02$, and the interaction Breaks × Prevalence, $F(1, 66) = 1.59$, $p = .212$, $\eta_p^2 = .02$, were not significant. For *worry*, the ANOVA revealed no significant effects: breaks, $F(1, 66) = 2.35$, $p = .13$, $\eta_p^2 = .03$; prevalence, $F(1, 66) = .58$, $p = .449$, $\eta_p^2 = .01$; or Breaks × Prevalence, $F(1, 66) = .04$, $p = .847$, $\eta_p^2 = .00$. For *engagement*, the ANOVA also revealed no significant effects for either breaks, $F(1, 66) = 0.70$, $p = .406$, $\eta_p^2 = .01$; prevalence, $F(1, 66) = 0.56$, $p = .455$, $\eta_p^2 = .01$; or for the interaction Breaks × Prevalence, $F(1, 66) = 0.04$, $p = .847$, $\eta_p^2 = .00$.

4. Discussion

To examine the effects of time on task and breaks on screener performance, two groups of airport security officers (screeners) performed an X-ray baggage screening task for 60 min. Whereas one group took breaks in line with the 20-min rule in the EU regulation, the other group worked for 60 min without breaks. Performance did not decrease over the course of 60 min of X-ray baggage screening. Moreover, breaks had no effect on performance. However, screeners without breaks reported more distress. Target prevalence was varied to determine the valid detection measure for this task. For X-ray image inspection, the detection measure d_a with a slope of approximately 0.6 seems to be a more valid measure of detection than *d'*. We confirmed the typical prevalence effect to be a criterion shift, and found that it developed at the beginning of the task.

Because our findings on the effects of time on task and breaks depend on selecting an appropriate detection measure, we first discuss the main effects of target prevalence and the change of hit rate, false alarm rate, sensitivity, criterion, and processing time in relation to the target prevalence effect. Then, we discuss the screeners' ability to maintain performance over time and the effect of breaks.

4.1. Detection measures for X-ray image inspection

Screeners showed a lower hit rate and a lower false alarm rate in the low target prevalence condition compared to the high target prevalence condition. This is the typical effect of target prevalence: People adjust their response tendency (criterion in signal detection theory) depending on the base rate with which targets occur (Godwin et al., 2010; Ishibashi and Kita, 2014; Ishibashi et al., 2012; Lau and Huang, 2010; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe and Van Wert, 2010). When comparing *d'* between the two target prevalence conditions over the full length of the task (i.e., the main effect of target prevalence), we found higher *d'* values for the low target prevalence condition in line with previous research on X-ray image inspection (Godwin et al., 2010; Wolfe et al., 2007; Wolfe and Van Wert, 2010). Consistent with these studies, we also found that screeners needed less time to inspect an image in this condition. In line with Kundel (2000) and Wolfe et al. (2007), we would argue that it is implausible for screeners to become faster and better at detection when fewer targets occur. It is more plausible that the equal variance assumption of *d'* (Green and Swets, 1966) is not met, and that the observed change in hit rate and false alarm rate is a mere change in response tendency (criterion *c* and c_a) as assumed in signal detection theory (Macmillan and Creelman, 2005). Comparing the *z*-transformed hit rate and false alarm rate between the two target prevalence conditions resulted in an average slope parameter of 0.65. This is close to the slope of around 0.6 that previous studies have found for the task of X-ray image inspection of passenger baggage (Godwin et al., 2010; Sterchi et al., 2019; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe and Van Wert, 2010). Therefore, in line with these previous studies, d_a seems to be the appropriate detection measure here. A comparison of the criterion across the two target prevalence conditions again showed a clear prevalence effect. As mentioned, screeners needed less time to inspect an

[3] Values are missing for two participants who did not fill out the SSSQ.

[4] Because the data did not always meet the assumptions of normal distribution or homoscedasticity, a Wilcoxon rank sum test was computed that also revealed a significant difference between the break conditions ($W = 1616$, $p = .003$).

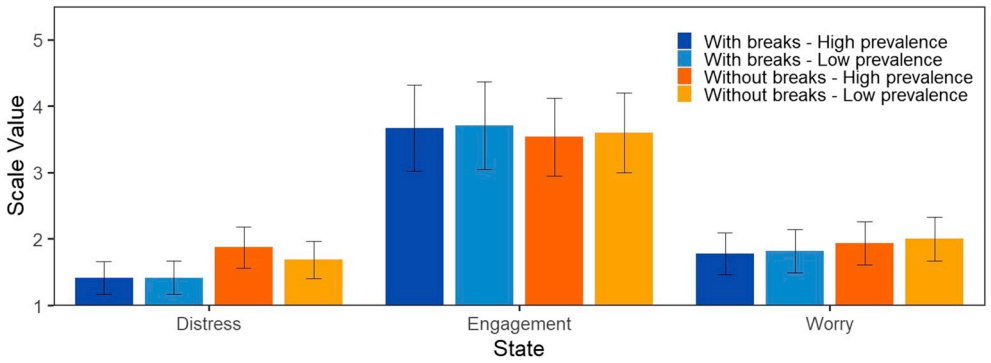


Fig. 6. Mean of reported levels of *distress, worry*, and *engagement* broken up by break and target prevalence conditions. Error bars represent standard errors.

X-ray image in the low target prevalence condition. This was especially the case for target-absent trials. Also previous research found shorter reaction times for target-absent trials when target prevalence was lower (Godwin et al., 2010; Wolfe et al., 2007; Wolfe and Van Wert, 2010).

In summary, consistent with previous studies, we found that a lower target prevalence leads to a criterion shift resulting in a lower hit rate and a lower false alarm rate. Moreover, for X-ray image inspection, our results confirm that d_a with a slope of about 0.6 is a more valid measure of detection performance than *d'*.

4.2. Interaction between target prevalence and time on task

Previous studies have found that the target prevalence effect depends on implicit learning and experienced prevalence rather than on explicit instruction, and that it therefore takes some time until searchers adapt to the prevailing target prevalence by shifting their criterion (Ishibashi et al., 2012; Lau and Huang, 2010). For the false alarm rate, we found a significant interaction between target prevalence and time on task. More specifically, the false alarm rate increased from the first (*0*–20 min) to the second (*20*–40 min) screening block in the high target prevalence condition and decreased in the low target prevalence condition. This is consistent with the criterion shift (c_a) that we found. However, for the hit rate, the interaction between target prevalence and time on task did not attain significance. Considering the *p* value was close to significance, this could have been due to insufficient statistical power. The hit rate was calculated from fewer images than the false alarm rate, and this led to higher standard errors. Our analysis of the criterion, which takes the hit and the false alarm rate into account, clearly confirms that the effect of target prevalence increased from the first (*0*–20 min) to the second (*20*–40 min) screening block of the task. In the high target prevalence condition, participants increased their tendency to declare that an X-ray image contained a prohibited item. In the low target prevalence condition, they increasingly reported images to be harmless (target absent). In general, our results are consistent with previous studies showing that participants first have to experience the prevalence of the targets for the target prevalence effect to fully develop (Ishibashi et al., 2012; Lau and Huang, 2010). In addition, consistent with findings reported by Lau and Huang (2010), we found that instructions alone were not sufficient to evoke the target prevalence effect.

4.3. Effect of time on task on screener performance

As mentioned in the previous section, we found a criterion shift at the beginning of the task that depended on the target prevalence condition. To discuss the effect of time on task on detection performance, it therefore makes sense to focus on the sensitivity measure d_a (with a slope of 0.65 in our study) that is not affected by this criterion shift. We found a small increase in d_a from the first screening block (*0*–20 min) to the second screening block (*20*–40 min) of the task and no change thereafter. This is consistent with the results of Chavaillaz et al. (2019), who also found a small increase in detection performance in the first 20 min of X-ray image inspection. It is possible that there is a warm-up phase in X-ray image inspection during which the cognitive processes necessary for this task become fully activated—as can be observed in other recognition tasks (Allport and Wylie, 1999; Monsell, 2003). Nonetheless, it is also possible that the observed ramp-up in performance was an accustomization to the specifics of the task employed in our experiment.

Whereas our study found no decline in performance over the course of 60 min, Meuter and Lacherez (2016) found a small decrease of two percentage points in hit rate after 10 min of screening under high workload (i.e., when screeners analyzed more than 5.4 baggage images per min). There are several possible explanations for this difference. The decrease Meuter and Lacherez found was quite small but based on a large amount of data. Our statistical power would not allow us to confirm a decrease in the hit rate of two percentage points. We further found that screeners adapted to the target prevalence by shifting their criterion at the beginning of the task. The change found by Meuter and Lacherez might also have been a criterion shift. However, this cannot be determined, because it was not possible to measure false alarm rate in their study. Finally, whereas their study analyzed data from a conventional checkpoint at which screening was performed in the lane, our study investigated RCBS. It may well be more difficult to maintain performance in an environment with more noise and distractors (Michel et al., 2014; Mocci et al., 2001; Yu et al., 2015).

As already argued in the introduction, X-ray image inspection shares certain similarities with vigilance tasks, but it also reveals clear differences. Whereas performance decreases within the first 15–30 min (Mackworth, 1948; Teichner, 1974; Warm, 1984) on most vigilance tasks, our participants were able to maintain their performance over the course of 60 min. This also argues against classifying X-ray baggage screening as a typical vigilance task. One could argue that our study contrasts more strongly with vigilance tasks than the conventional X-ray baggage screening task, because we used higher target prevalence levels. However, whereas certain threats such as IEDs are rare in practice, other prohibited articles such as liquids and gels left in baggage still provide quite common targets.

Regarding processing times as well, screeners were able to maintain their performance throughout the full duration of the task. Processing times even decreased throughout the task with the exception of target-present trials in the high prevalence condition, where no change was found (for a similar effect, see Chavaillaz et al., 2019). For target-absent trials, processing times decreased throughout the task in the low and the high prevalence condition, but more strongly in the low prevalence condition. These decreases in processing times cannot be associated with a speed–accuracy tradeoff because there was no decrease in the performance measure d_a. It is more likely that screeners adapted to the task, its conditions, or the interface settings. We cannot be sure whether this effect would also occur in practice after screeners become familiar with the X-ray machine interface. Beyond the general speed–accuracy

tradeoff, the nature of the investigated task, which included the marking and categorizing of targets, is not well suited for a detailed discussion of processing times.

4.4. The effects of breaks on performance

Closely linked to how performance changes over time is the question regarding what effect breaks have on performance. We did not find effects of breaks on the hit rate, false alarm rate, sensitivity, or response tendency (criterion). Likewise, Chavaillaz et al. (2019) found no performance differences between different break regimes (spontaneous breaks and 5 or 10 min breaks every 20 min) in a 60-min simulated X-ray baggage screening task with student participants. Whereas breaks have often had a positive effect on performance in previous studies conducted in other detection tasks (Arrabito et al., 2015; Colquhoun, 1959; Kopardekar and Mital, 1994), they are mainly thought to offer rest, recuperation, and prevention of fatigue (Tucker, 2003). Although our participants who performed 60 min of continuous X-ray screening had no possibility of recuperation during breaks, they still did not show a decrease in performance. We found a main effect of breaks on processing times for target-present trials, but not for target-absent trials. However, the effect was already present in the first 20 min screening block and did not increase thereafter, indicating that it was not the result of the breaks themselves. Maybe knowing that there will be no breaks induced some stress, and the associated arousal, in turn, led to faster processing times. This is related to the effects we found in terms of well-being or distress. The screeners in the condition without breaks reported more distress in the SSSQ. Hence, whereas screeners were able to maintain detection performance over 60 min without breaks, this led to increased distress. In the long term, this could have an effect on performance.

4.5. Limitations and future research

Whereas our study has shown that screeners can maintain detection performance over 60 min without breaks, we also found that this caused more distress. Considering that participants only did 60 min of X-ray screening twice with 3–5 weeks in between, it is unclear how prolonged screening would affect performance and well-being if it were to be repeated multiple times a day and over months. On the one hand, distress levels might decrease due to increased practice. On the other hand, distress levels might increase further over time. This, in turn, could have a negative impact on well-being and on performance in the long term. Therefore, field studies are needed to determine the long-term effect of longer screening durations on performance and well-being. Such field studies would also tackle other limitations of our study. In our laboratory experiment, poor performance did not have any consequences, whereas a miss can be disastrous in practice. This might make prolonged screening time more stressful. Furthermore, target prevalence is lower in practice, and this could make it more difficult to sustain attention and performance. It is quite possible that people react differently to prolonged working sessions. Future studies could investigate interindividual differences and test whether flexible break schedules would provide a solution, although this might be difficult in practice.

5. Conclusions

Our study showed that screener performance did not decrease in continuous X-ray inspection over the course of 60 min. Moreover, breaks did not influence performance. However, breaks did seem to have an effect on well-being, in the sense that screeners without breaks reported more distress. Our results open the discussion on whether more flexible break policies and work schedules should be considered. They provide a basis for conducting field studies of prolonged screening durations. This should include a careful monitoring of screeners' performance and well-being. If field trials succeed, relaxing the 20-min rule would provide additional flexibility that could be helpful when implementing new technologies such as remote cabin baggage screening. In addition, this study provides further evidence that d_a with a slope of approx. 0.6 is a more valid measure of detection performance than d' for the X-ray image inspection of cabin baggage and should be considered in future studies on this task.

References

Allport, A., Wylie, G., 1999. Task-switching: positive and negative priming of task-set. In: Humphreys, G.W., Duncan, J., Treisman, A. (Eds.), Attention, Space, and Action: Studies in Cognitive Neuroscience. Oxford University Press, New York, NY, pp. 273–296.

Arrabito, G.R., Ho, G., Aghaei, B., Burns, C., Hou, M., 2015. Sustained attention in auditory and visual monitoring tasks: evaluation of the administration of a rest break or exogenous vibrotactile signals. Hum. Factors 57 (8), 1403–1416. https://doi.org/10.1177/0018720815598433.

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., 2014. Improving the efficacy of security screening tasks: a review of visual search challenges and ways to mitigate their adverse effects. Appl. Cognit. Psychol. (29), 142–148. https://doi.org/10.1002/acp.3083.

Bolfing, A., Halbherr, T., Schwaninger, A., 2008. Image based factors and human factors contribute to threat detection performance in X-ray aviation security screening. In: HCI and Usability for Education and Work. 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008, pp. 419–438. https://doi.org/10.1007/978-3-540-89350-9_30. Proceedings.

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.

Colquhoun, W.P., 1959. The effect of a short rest-pause on inspection efficiency. Ergonomics 2 (4), 367–372. https://doi.org/10.1080/00140135908930451.

Cutler, V., Paddock, S., 2009. Use of threat image projection (TIP) to enhance security performance. Proc. 43rd IEEE Int. Carnahan Conf. Secur. Technol. 29–35. https://doi.org/10.1109/CCST.2009.5335565.

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

Efron, B., 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82 (397), 171–185. https://doi.org/10.1080/01621459.1987.10478410.

European Commision, 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. Off. J. Eur. Union.

Galinsky, T.L., Swanson, N.G., Sauter, S.L., Hurrell, J.J., Schleifer, L.M., 2000. A field study of supplementary rest breaks for data-entry operators. Ergonomics 43 (5), 622–638. https://doi.org/10.1080/001401300184297.

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., 2010. 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.

Green, D.G., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley & Sons, Inc. https://doi.org/10.1901/jeab.1969.12-475.

Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24 (2), 95–112. https://doi.org/10.1007/BF02289823.

Hardmeier, D., Hofer, F., Schwaninger, A., 2005. The X-ray object recognition test (X-ray ORT) - a reliable and valid instrument for measuring visual abilities needed in X-ray screening. In: Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology, pp. 189–192. https://doi.org/10.1109/CCST.2005.1594876 (c).

Harris, D.H., 2002. How to really improve airport security. Ergon. Des 10 (1), 17–22. https://doi.org/10.1177/106480460201000104.

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 https://doi.org/10.3389/fpsyg.2019.00525.

Hautus, M.J., 1995. Corrections for extreme proportions and their biasing effects on estimated values of d′. Behav. Res. Methods Instrum. Comput. 27 (1), 46–51. https://doi.org/10.3758/BF03203619.

Helton, W.S., 2004. Validation of a short stress state questionnaire. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 48 (11), 1238–1242. https://doi.org/10.1177/154193120404801107.

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.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6 (2), 65–70. https://doi.org/10.2307/4615733.

Ishibashi, K., Kita, S., 2014. Probability cueing influences miss rate and decision criterion in visual searches. I-Perception 5 (3), 170–175. https://doi.org/10.1068/i0649rep.

Ishibashi, K., Kita, S., Wolfe, J.M., 2012. The effects of local prevalence and explicit expectations on search termination times. Atten. Percept. Psychophys. 74 (1), 115–123. https://doi.org/10.3758/s13414-011-0225-4.

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.

Koller, S.M., Hardmeier, D., Michel, S., Schwaninger, A., 2008. Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-Ray image interpretation. J. Transp. Secur. 1 (2), 81–106. https://doi.org/10.1007/s12198-007-0006-4.

Kopardekar, P., Mital, A., 1994. The effect of different work-rest schedules on fatigue and performance of a simulated directory assistance operator's task. Ergonomics 37 (10), 1697–1707. https://doi.org/10.1080/00140139408964946.

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

Kundel, H.L., 2000. Disease prevalence and the index of detectability: a survey of studies of lung cancer detection by chest radiography. In: SPIE 3981, Medical Imaging 2000: Image Perception and Performance, vol. 3981, pp. 135–144. https://doi.org/10.1117/12.383100.

Lau, J.S.H., Huang, L., 2010. The prevalence effect is determined by past experience, not future prospects. Vis. Res. 50 (15), 1469–1474. https://doi.org/10.1016/j.visres.2010.04.020.

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.

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, second ed. Lawrence Erlbaum Associates, Mahwah, New Jersey.

McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., Boot, W.R., 2004. Visual skills in airport-security screening. Psychol. Sci. 15 (5), 302–306. https://doi.org/10.1111/j.0956-7976.2004.00673.x.

Mendes, M., Schwaninger, A., Michel, S., 2011. Does the application of virtually merged images influence the effectiveness of computer-based training in x-ray screening? Proc. 45th IEEE Int. Carnahan Conf. Secur. Technol. 1–8. https://doi.org/10.1109/CCST.2011.6095881.

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 (2), 218–228. https://doi.org/10.1177/0018720815616306.

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., Schwaninger, A., 2014. A multi-method approach towards identifying situational factors and their relevance for X-ray screening. Proc. 48th IEEE Int. Carnahan Conf. Secur. Technol. 208–213. https://doi.org/10.1109/CCST.2014.6987001.

Mitroff, S.R., Biggs, A.T., Cain, M.S., 2015. Multiple-target visual search errors: overview and implications for airport security. Policy Insights Behav. Brain Sci. 2 (1), 121–128. https://doi.org/10.1177/2372732215601111.

Mocci, F., Serra, A., Corrias, G.A., 2001. Psychological factors and visual fatigue in working with video display terminals. Occup. Environ. Med. 58 (4), 267–271. https://doi.org/10.1136/oem.58.4.267.

Monsell, S., 2003. Task switching. Trends Cogn. Sci. 7 (3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7.

Nuechterlein, K.H., Parasuraman, R., Jiang, Q., 1983. Visual sustained attention: image degradation produces rapid sensitivity decrement over time. Science 220 (4594), 327–329. https://doi.org/10.1126/science.6836276.

Pollack, I., Norman, D.A., 1964. A non-parametric analysis of recognition experiments. Psychon. Sci. 1 (12), 125–126. https://doi.org/10.3758/bf03342823.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. http://www.R-project.org.

Rose, C.L., Murphy, L.B., Byard, L., Nikzad, K., 2002. The role of the big five personality factors in vigilance performance and workload. Eur. J. Personal. 16 (3), 185–200. https://doi.org/10.1002/per.451.

Schwaninger, A., 2004. Computer based training: a powerfull tool to the enhancement of human factors. Aviat. Secur. Int. 2, 31–36.

Simpson, A.J., Fitter, M.J., 1973. What is the best index of detectability? Psychol. Bull. 80 (6), 481–488. https://doi.org/10.1037/h0035203.

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.

Sterchi, Y., Hättenschwiler, N., Schwaninger, A., 2019. Detection measures for visual inspection of X-ray images of passenger baggage. Atten. Percept. Psychophys. 81 (5), 1297–1311. https://doi.org/10.3758/s13414-018-01654-8.

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

Tucker, P., 2003. The impact of rest breaks upon accident risk, fatigue and performance: a review. Work Stress 17 (2), 123–137. https://doi.org/10.1080/0267837031000155949.

Van Wert, M.J., Horowitz, T.S., Wolfe, J.M., 2009. Even in correctable search, some types of rare targets are frequently missed. Atten. Percept. Psychophys. 71 (3), 541–553. https://doi.org/10.3758/APP.71.3.541.

Wales, A.W.J., Anderson, C., Jones, K.L., Schwaninger, A., Horne, Ja A., 2009. Evaluating the two-component inspection model in a simplified luggage search task. Behav. Res. Methods 41 (3), 937–943. https://doi.org/10.3758/BRM.41.3.937.

Warm, J.S., 1984. Sustained Attention in Human Performance. Wiley, Chichester, UK.

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.

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

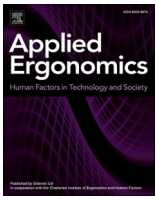
Yu, R., Yang, L., Liu, C., 2015. Effects of target prevalence and speech intelligibility on visual search performance. Meas. Control 48 (3), 87–91. https://doi.org/10.1177/0020294015569265.

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.

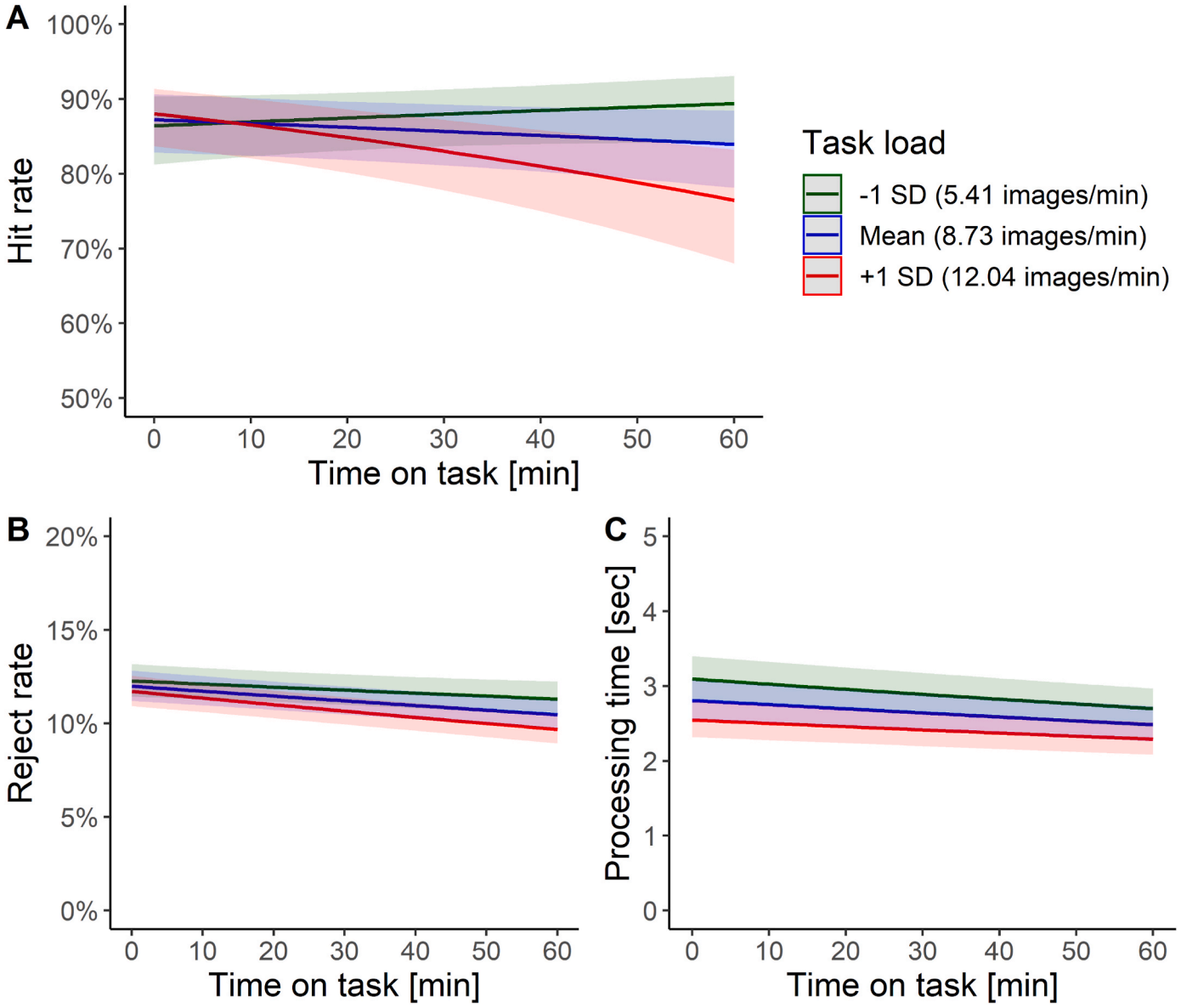


Fig. 1. Effects of time on task on hit rate (A), reject rate (B), and processing time (C) for three levels of task load: M - 1 SD, M, and M + 1 SD.

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

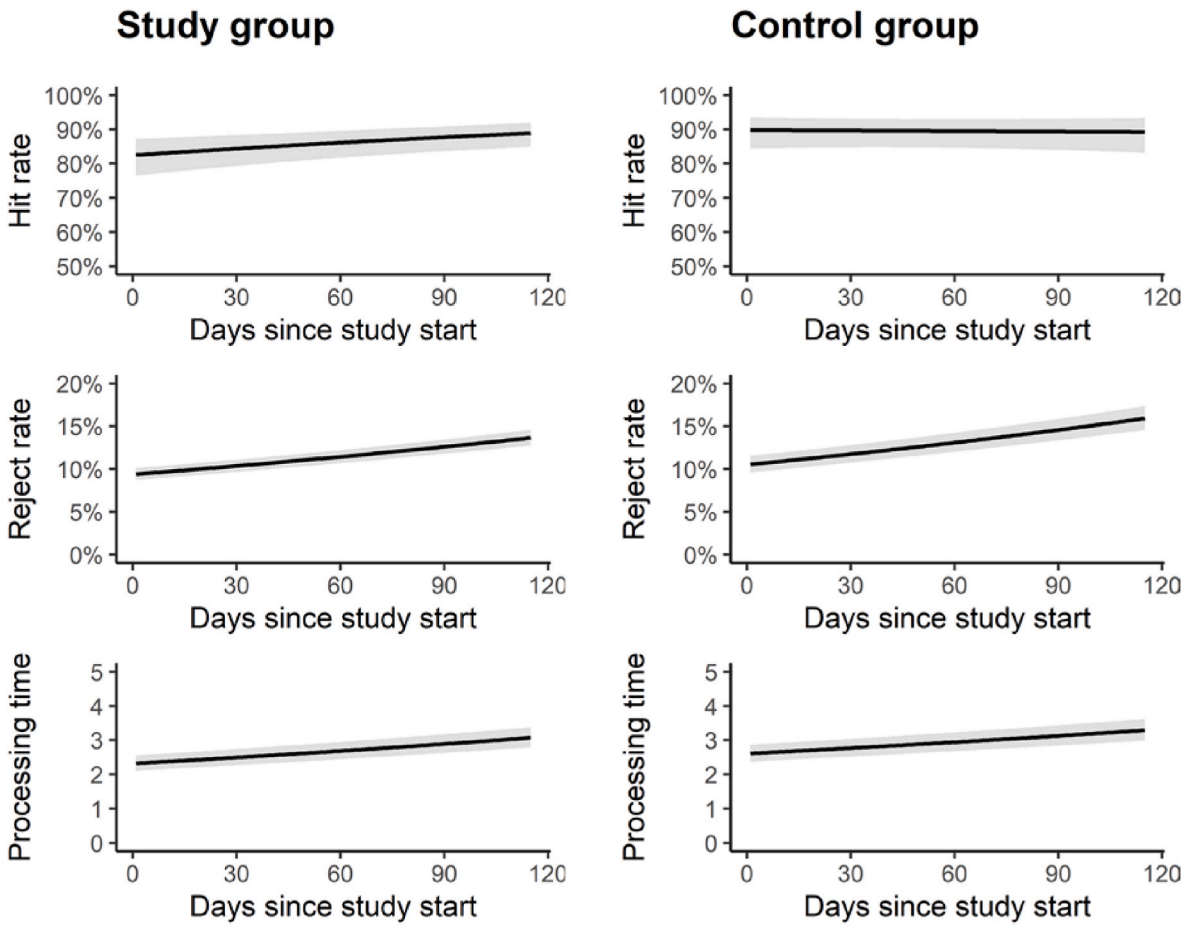


Fig. 2. Change in hit rate, reject rate, and processing time over the course of the study for the study group (left), and the control group (right).

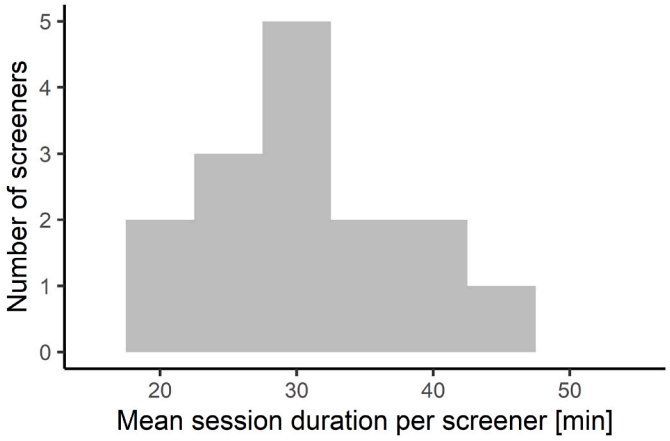


Fig. 3. Distribution of the mean screening duration per screener of the study group for sessions that were ended on the screeners' own terms.

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

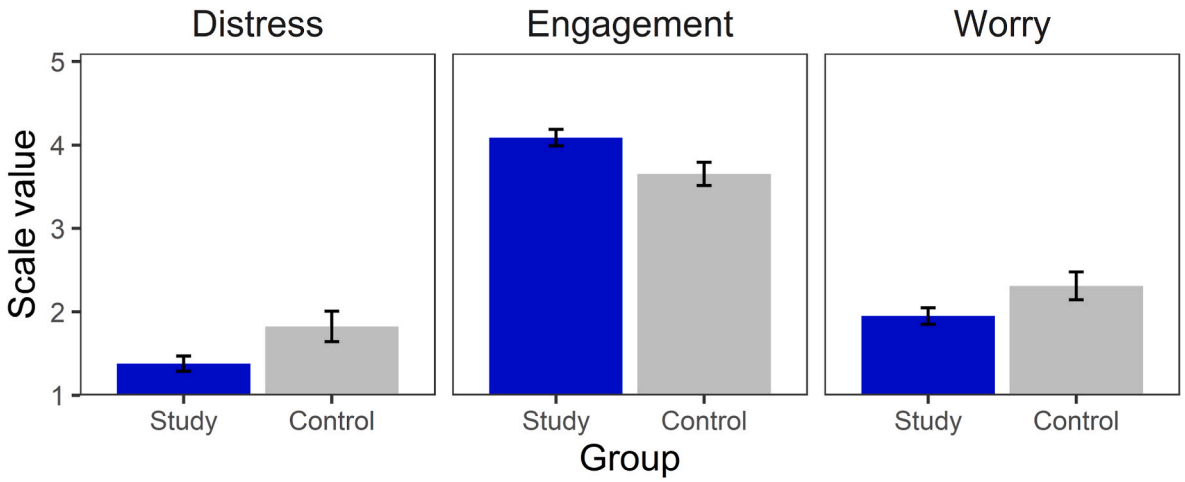


Fig. 4. Mean of Distress, Engagement, and Worry for the study and control group. Error bars represent standard errors.

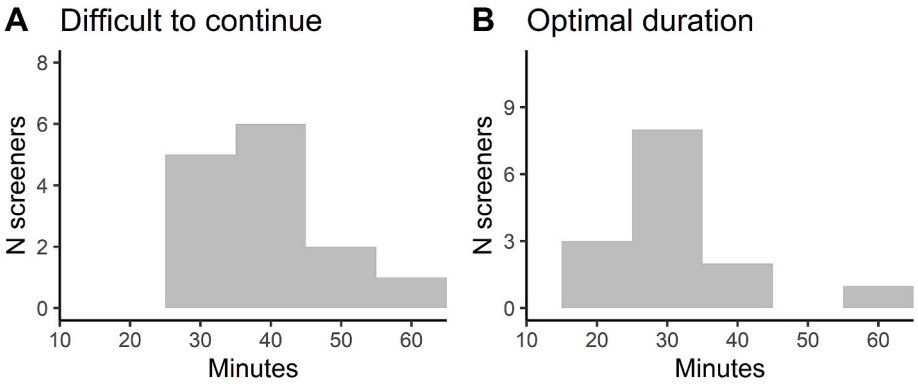


Fig. 5. Distribution of the study group's responses to the survey questions (A) "After what time did it get difficult to continue screening?", and (B) "What do you consider to be an optimal screening duration?"

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.

Journal of Air Transport Management

Reliability and validity of threat image projection data on X-ray baggage screening

--Manuscript Draft--

Manuscript Number:	JATM-D-22-00516
Article Type:	Research Paper
Keywords:	Aviation security; X-ray baggage screening; Threat image projection; Covert test; Key performance indicator; Socio-technical system
Corresponding Author:	Yanik Sterchi FHNW University of Applied Sciences and Arts Northwestern Switzerland Olten, SWITZERLAND
First Author:	Daniela Buser
Order of Authors:	Daniela Buser
	Adrian Schwaninger
	Vaclav Rehor
	Yanik Sterchi
Abstract:	Many airports use threat image projection (TIP) to measure how well security officers (screeners) detect prohibited items during passengers' baggage screening. However, research on the reliability and validity of TIP data is scarce. To address this research gap, we analyzed a large dataset of cabin baggage TIP data (1,206,076 TIP events from 728 screeners over four years). We found reliability to increase with the number of TIP events in accordance with the Spearman–Brown prediction and that approximately 100 TIP events were sufficient to achieve a minimum reliability value of 0.7 during periods when TIP was difficult. TIP data predicted the outcome of covert tests (wherein instructed people tried to smuggle real prohibited articles through the checkpoint; 1,194 covert tests from 474 screeners), indicating that TIP is a valid measure of operational threat detection. The results imply that TIP provides a reliable and valid measure of threat detection if it is difficult enough and 100 or more TIP events are considered per screener.
Suggested Reviewers:	Hayward Godwin hayward.godwin@soton.ac.uk Hayward Godwin has published several articles on X-ray baggage screening and experience with threat image projection data.
	Renata Meuter r.meuter@qut.edu.au Renata Meuter has published an article with Pilippe Lacherez based on threat image projection data with policy implications for airport security.
	Pilippe Lacherez p.lacherez@qut.edu.au Philippe Lacherez has published an article with Renata Meuter based on threat image projection data with policy implications for airport security.

Reliability and validity of threat image projection data on X-ray baggage screening

Buser[1], Schwaninger[1], Rehor[2] & Sterchi[1]

[1]Institute Humans in Complex Systems, School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland

[2]Department of Air Transport, Faculty of Transportation Sciences, Czech Technical University in Prague, Prague 1, Czech Republic

Conflict of Interest: None

Funding: This work was funded by the Swiss Federal Office of Civil Aviation and the University of Applied Sciences and Arts Northwestern Switzerland.

Reliability and validity of threat image projection data on X-ray baggage screening

Abbreviations
CBS: cabin baggage screening
CI: confidence interval
CTT: classical test theory
EU: European Union
FTI: fictional threat items
GEE: generalized estimation equations
HBS: hold baggage screening
IED: improvised explosive device
TIP: threat image projection

Abstract

Many airports use threat image projection (TIP) to measure how well security officers (screeners) detect prohibited items during passengers' baggage screening. However, research on the reliability and validity of TIP data is scarce. To address this research gap, we analyzed a large dataset of cabin baggage TIP data (1,206,076 TIP events from 728 screeners over four years). We found reliability to increase with the number of TIP events in accordance with the Spearman–Brown prediction and that approximately 100 TIP events were sufficient to achieve a minimum reliability value of 0.7 during periods when TIP was difficult. TIP data predicted the outcome of covert tests (wherein instructed people tried to smuggle real prohibited articles through the checkpoint; 1,194 covert tests from 474 screeners), indicating that TIP is a valid measure of operational threat detection. The results imply that TIP provides a reliable and valid measure of threat detection if it is difficult enough and 100 or more TIP events are considered per screener.

Keywords: airport security, cabin baggage screening, detection performance, threat image projection, covert tests

1. Introduction

Most airports worldwide use threat image projection (TIP) technology to measure how well security officers (screeners) detect threats during passengers' baggage screening. With TIP, pre-recorded X-ray images of prohibited items (bombs, guns, knives, etc.) are projected onto the X-ray images of passengers' baggage (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Skorupski and Uchroński, 2016). The TIP system records whether the screeners detect them. These data are used for quality control by airports, governments, and security companies (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Riz à Porta et al., 2022; Skorupski and Uchroński, 2016). Although widely used, only one study has investigated the reliability of TIP data (Hofer and Schwaninger, 2005). Furthermore, to our knowledge, no study has examined whether TIP data provide a valid measure of real prohibited item detection. Therefore, we aimed to examine the reliability of TIP by analyzing a large set of data from an international airport. We also assessed the validity of the TIP data by analyzing whether it can predict how well screeners detect prohibited items in covert tests, wherein instructed people attempted to smuggle prohibited items in their baggage (e.g., knives, inert bombs, or guns) past the checkpoint (Walter et al., 2021; Wetter et al., 2008).

2. Background and theory

In many countries, TIP is mandated and used to ensure the minimum detection performance of screeners (Bassetti, 2021). Therefore, pre-recorded images of threats, known as fictional threat items (FTIs), are projected onto 1%–4% of all passenger baggage during screening (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Meuter and Lacherez, 2016; Skorupski and Uchroński, 2018). When screeners suspect a prohibited item, they press a designated button and the system immediately provides feedback on whether an FTI is present or not (Bassetti et al.,

2015; Schwaninger, 2006). The screeners' responses to such TIP images are recorded by the system (TIP events), providing data on the TIP performance of screeners, typically calculated as the hit rate (the percentage of projected TIP images detected by the screener) (Hofer and Schwaninger, 2005; Meuter and Lacherez, 2016). If screeners do not achieve a minimum hit rate, they must undergo remedial training and successfully complete a recertification process before they are allowed to continue screening during the operation. Some airports also reward or penalize screeners depending on their TIP performance (Bassetti, 2018). Researchers analyze TIP data to answer various research questions related to visual search and human factors at security checkpoints (Buser et al., 2022; Meuter and Lacherez, 2016; Skorupski and Uchroński, 2016). Other than performance measurement, TIP also addresses the fact that most threats, such as real bombs or guns, occur very rarely at checkpoints. Research has shown that people do not recognize rare targets well, which is known as the target prevalence effect (Godwin et al., 2010; Menneer et al., 2010; Wolfe et al., 2007, 2005). By increasing the number of threats to be detected, TIP aims to mitigate this effect while also increasing motivation by providing screener feedback on detection performance (Cutler and Paddock, 2009; Harris, 2002; Riz à Porta et al., 2022; Schwaninger, 2009).

TIP data must be reliable and valid for effective use. Reliability refers to the extent to which results are reproducible with repeated measurements, which corresponds to its accuracy or consistency (Murphy and Davidshofer, 2014). In other words, the measured TIP performance is reliable if a screener shows similar TIP performances when measured repeatedly under similar conditions. Ensuring reliability is of particular importance when the measurement of a skill or construct has consequences for the tested subjects (Murphy and Davidshofer, 2014). The reliability of a measurement method can be quantified by adopting a statistical model, like the classical test theory (CTT). CTT was originally conceptualized to test the accuracy and sensitivity of

questionnaires or tests. It assumes that a single underlying dimension, such as threat detection skills, is being measured and that every person has a single true score (T) on that dimension (Murphy and Davidshofer, 2014; Rindskopf, 2015). Therefore, all test items (in this case, TIP images) should measure the same skill (threat detection). It is assumed that a person's observed test score, X, is equal to the sum of their true score, T, and the measurement error, e. Therefore, the basic equation of CTT is as follows:

$$X = T + e$$

CTT thereby assumes that the errors have a mean of zero, are independent of the true score and that errors on different measures are independent. A consequence of this model is that the variation across individuals or the variance in the observed test scores is the sum variance of true scores, S_T^2, and the variance in the error, S_e^2.

$$S_X^2 = S_T^2 + S_e^2$$

Based on this equation, reliability is defined as the proportion of total variance in scores attributable to true variance, S_T^2, rather than error variation, S_e^2.

$$Reliability = \frac{S_T^2}{(S_T^2 + S_e^2)} = \frac{S_T^2}{S_X^2}$$

This statistical model leads to several possible methods for estimating reliability (for an overview see: Murphy and Davidshofer, 2014; Rindskopf, 2015); however, in principle, two or more tests are used at different time points and the outcomes of such correlation result in reliability values that lie between zero and one. Values close to one indicate a very reliable test, whereas a reliability of zero indicates that the measured scores are only random and are not because of the measured construct. To determine the reliability of the TIP data, we employed the split-half

reliability method. In this method, items of a test (in this case, responses to TIP images over a certain period) are split into two groups; the score of one half is related to that of the other half by correlating the test scores of both halves. The more similar the test scores of the two halves, the higher the reliability. A disadvantage of using split-half reliability is that it only determines the test score for half of the available items because they are split into two groups. As reliability increases with the number of considered items, the reliability for half of the items is lower than the reliability for the full set. Under the CTT assumptions, changes in reliability based on the number of items can be estimated using the Spearman–Brown prediction (Brown, 1910; Spearman, 1910). Therefore, this formula is commonly used to correct the split-half reliability for full tests or to calculate how long a test would have to be to achieve a certain reliability.

$$r_2 = \frac{k\, r_1}{1 + (k-1) r_1}$$

Therefore, r_1 is the reliability of the original test and r_2 is the predicted reliability of a test, which is a factor k longer than the original test. Reliability should at least reach a minimum value of 0.7 (Kline, 2000; Murphy and Davidshofer, 2014). This applies if the test results are used as a first indication (e.g., dividing the screeners into two performance groups) or for group diagnostics. However, based on the results, if measures or decisions are taken with consequences for the individual (e.g., getting hired or undergoing remedial training), it is highly recommended to achieve higher reliability values of at least 0.8 (Brough, 2019; Murphy and Davidshofer, 2014).

As introduced above, reliability coefficients indicate the proportion of the variance in score that is attributable to true variance and not error (e.g., a reliability of 0.8 indicates that 80% of the variance stems from variation in the true score and 20% stems from error), and therefore, provides a measure of accuracy relative to the score's variance. Under the CTT assumptions, absolute

measures of accuracy, or the standard error and confidence intervals (CI), can be derived from the reliability coefficient of the test, r, and the variability of test scores, S^2.

$$SE = S^2 * \sqrt{(1 - r)}$$

Standard error indicates the variability in test scores attributable to measurement errors in absolute terms (Murphy and Davidshofer, 2014). The standard error can then be used to compute CI, which inform the range in which the true score of an individual lies with a certain level of confidence in a normal distribution.

It is unclear whether CTT assumptions can be applied to TIP data because of the differences between TIP and standardized tests. With FTI TIP, every image and test item is different because FTIs are always projected onto a different X-ray image of passengers' baggage at a random position. Furthermore, different TIP libraries composed of different images can be used at one airport, and for each library, TIP images are exchanged and updated regularly. In addition, screeners differ in the number of TIP images they analyze, and responses to TIP images are collected over a long period of time (six months), during which the performance of screeners can change. Therefore, it is important to conduct research on TIP data reliability and how well TIP meets the CTT assumptions and whether estimates based on CTT can be applied to TIP.

To our knowledge, only one study has examined the reliability of TIP data (Hofer and Schwaninger, 2005). The researchers split TIP data from cabin baggage screening (CBS) and hold baggage screening (HBS) into two groups based on odd and even days and calculated multiple split-half reliabilities using different data aggregations. Reliability values above 0.7 were achieved for HBS data; however, reliability values were insufficient for CBS data (0.58 and lower). The researchers hypothesized that high TIP hit rates in the CBS data caused ceiling effects in

performance and that very small inter-individual differences in performance could have led to these low reliability results. However, differences in reliability between CBS and HBS TIP data may have also arisen because of differences in the TIP libraries and systems used; FTI TIP was implemented for CBS, whereas combined threat image TIP was used for HBS where pre-recorded X-ray images of baggage, including prohibited items, were projected into the stream of X-ray images during screening. Another limitation of the study was that the number of TIP events was not directly considered; the two halves of the test were unequal in size, and the reduced number of TIP images because of splitting of the data was not corrected. Moreover, the study did not examine the validity of the TIP data.

TIP performance should not only be reliable, but also a valid measure of the detection performance. Validity is defined as the extent to which a test measures what it claims to measure, which, in this case, is the detection of real prohibited items. Several methods are used to assess validity (Murphy and Davidshofer, 2014). One method is to analyze the degree to which a performance measure can predict the true performance of the skill being measured. If the TIP performance is a valid measure of threat detection, it should be able to predict how likely a screener is to detect real prohibited items in an X-ray image. However, a study by Riz a Porta et al. (2022) showed that screeners at an international airport considered approximately one-third of all TIP images to contain artifacts and look unrealistic. Therefore, it is unclear whether TIP truly measures the detection of real prohibited items, and it is important to investigate whether TIP performance and the detection of prohibited items are correlated. Covert tests are conducted at airports and by governments and police staff as quality control measures (Schwaninger, 2009; Skorupski and Uchroński, 2016; Walter et al., 2021; Wetter et al., 2008). By correlating TIP data with covert test results, the validity of TIP can be investigated, which is another important contribution of our study.

3. Materials and Methods

2.1 Participants and data

The participants were professional airport security screeners who were selected, qualified, trained, and certified according to the standards set by the appropriate national authority (civil aviation administration) in compliance with the relevant regulations in the European Union (European Commission, 2015). We analyzed TIP and covert test data from an international airport covering four years of CBS. The data included all responses to the TIP events of the screeners, which were recorded as either hit (TIP detected) or missed (TIP missed). To examine the reliability of the TIP data, we analyzed 1,206,076 TIP events from 728 screeners over four years. To analyze the validity of the TIP data, we performed correlational analyses with 1,194 covert test results from 474 screeners from the same participant sample. 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.

2.2 Procedure

Similar to other airports (Michel et al., 2014), the screeners worked at four positions at the checkpoint and rotated among themselves. At the X-ray screening point, each screener logged into the workstation with a unique user ID and inspected the X-ray images of the passengers' baggage for prohibited items. The TIP systems at this airport projected FTIs onto X-ray images of passengers' baggage with a target prevalence of 2.9%. The screeners were aware that the TIP was operational and that their detection performance was monitored. When screeners suspected a prohibited item, they indicated this by pressing a specific button, and the TIP system provided

immediate feedback on whether an FTI was present or not. If FTI was present, the X-ray image had to be analyzed again without FTI. If no FTI was present, the relevant piece of baggage was further inspected (through manual search and explosive trace detection). After X-ray image inspection, screeners logged out from the workstation and continued working at another checkpoint position or took a break.

At the target airport, covert tests were conducted regularly by the staff recruited by the airport, which was rotated to avoid recognition by screeners. The staff were then instructed to smuggle a real threat through security control, either in their baggage or on themselves. The selection and placement of prohibited items followed the protocol defined by the quality control team. The prohibited items corresponded to the same categories as those in the TIP (guns, bombs, knives, etc.). For each test, the prohibited item category, where the item was placed, and the difficulty of the test were protocoled. After the covert test, the outcome was discussed with the involved screener(s) and the difficulty of each test was evaluated again by the quality team by reviewing the X-ray image of the baggage recorded during the test. The outcome was documented as either pass (item found) or fail (item not found). Further information such as which checkpoint the test was conducted, if and which X-ray machine type was used, as well as the date and time were protocoled.

2.3 Analyses

2.3.1. Reliability of TIP data

To investigate the reliability of the TIP data, we assessed split-half reliability using the following procedure: to control for changes in TIP performance over time, TIP events were first sorted by date and time of occurrence, and every two consecutive TIP events per screener were paired. To estimate the reliability for *n* number of TIP events, *n* pairs of TIP events were randomly selected

(without replacement) and the TIP events of each pair were randomly split into two groups. This ensured that both groups had a comparable distribution of TIP events over time for each screener. For each screener and each of the two groups of TIP events, the TIP hit rate (proportion of detected TIP events) was calculated, and the Pearson correlation between the hit rates of the two groups was computed. To estimate reliability, all random steps (i.e., sampling and splitting of pairs) were repeated 10,000 times and the resulting correlation coefficients were averaged.

As airports and authorities usually consolidate performance on a half-year basis, our analyses were conducted for half-year periods from July 2015 to June 2019. In the first step, we determined whether the Spearman–Brown prediction accurately described how the reliability varied as a function of the number of TIP events considered for TIP performance evaluation. To estimate the reliabilities for the various numbers of TIP events using the same sample of screeners, only screeners with at least 100 TIP events per six months were included. We calculated the split-half reliabilities (as described above) considering 5–50 TIP events for performance evaluation, in increments of five. The reliability of the corresponding number of TIP events was then estimated using the Spearman–Brown prediction. For this purpose, the reliability of 25 TIP events (per split) was used as the baseline. We then compared the reliabilities of the two calculations. To determine the reliability of each half-year, we wanted to retain as many screeners as possible for analysis. Therefore, 10 TIP events per screener and per split were used to calculate the split-half reliability for each half-year. Consequently, screeners that did not have a minimum of 20 TIP events within the respective half-year period were excluded from the analysis. Table 1 shows the number of data points excluded per half-year based on these requirements. The split-half reliability was

determined for 20 TIP events (per split), and based on the Spearman–Brown prediction, the reliability for higher numbers of TIP events (50, 100, 345) was calculated[1].

Table 1. Number of TIP events and screeners excluded from reliability analysis per half-year

Time period	Screeners excluded (N)	Screeners excluded (%)	TIP excluded (N)	TIP events excluded (%)
1	15	3.72	106	0.08
2	16	3.90	168	0.12
3	13	3.05	96	0.06
4	16	3.59	138	0.08
5	24	5.19	258	0.13
6	34	7.05	305	0.18
7	41	8.20	373	0.27
8	49	9.86	355	0.36

Note. The excluded screeners did not meet the minimum of 20 TIP events per half-year.

2.3.2 Validity of TIP data

To assess the validity of the TIP data, we used binomial generalized estimation equations (GEE) (Ballinger, 2004; Liang and Zeger, 1986) with TIP and covert test data from July 2015 to June 2019.

As was the case for TIP, only CBS covert test data from the X-ray positions were included. Furthermore, only covert tests for which it was possible to link half a year of TIP data before and after the covert test were included. In total, 1,194 covert tests from 474 screeners were used for statistical analyses. An average of 2.52 covert tests were analyzed per screener (*SD* = 1.79). To

[1] The data originate from two separate TIP systems. An additional analysis revealed that the reliabilities of the two systems were comparable, and that the data can be analyzed jointly. Individual analysis of the libraries resulted in a reliability value of 0.45 and jointly of 0.43 using 50 TIP events. Standard deviations were between 0.33–0.34 and the standard error was 0.0003.

examine the effect of the TIP hit rate on the covert test performance, a binomial GEE was calculated, as GEEs are suited to fit generalized linear models with longitudinal and clustered data.[2] The TIP hit rate was included by aggregating TIP events that occurred within half a year before or after the covert test. To ensure a better model convergence, the TIP hit rate was z-transformed. The model also controlled for the prohibited item category (gun, knife, improvised explosive device, etc.), different checkpoints within the airport, X-ray machine type, and complexity of the covert test as factors, and the screener as a random variable. The binomial GEE was estimated using the R-package GEE, with an exchangeable correlation structure. All analyses were performed using R (R Core Team, 2020).

4. Results

3.1 Reliability of TIP data

The Spearman–Brown prediction corresponded well with the empirically estimated reliabilities and, therefore, provided an accurate description of how the reliability increased with the number of TIP events, as illustrated in Figure 1 for the first three of the eight half-year periods (the other five half-year periods showed the same level of correspondence).

[2] Additionally, a generalized mixed model was estimated, which resulted in very similar coefficient estimates but showed singularity problems.

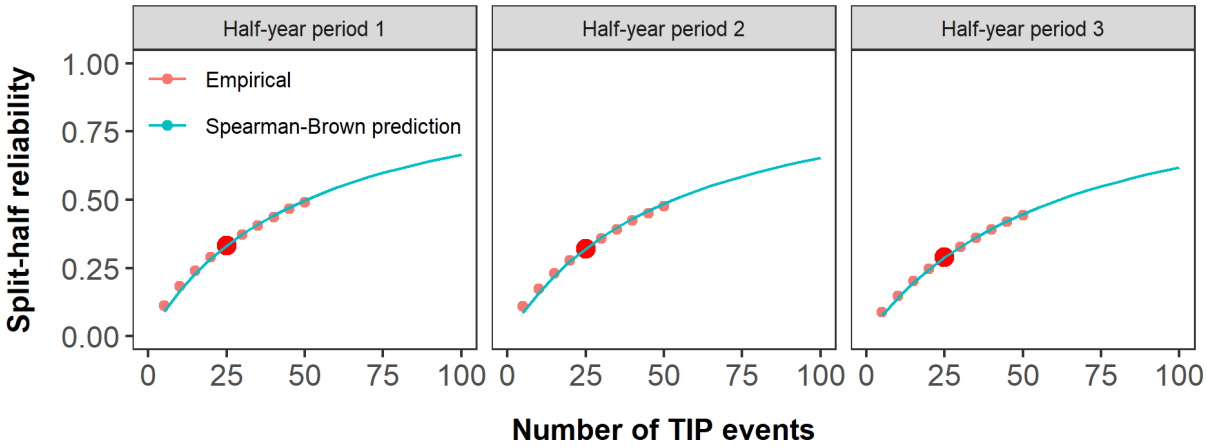


Figure 1. Split-half reliability based on the number of TIP events considered for calculating TIP performance for the first three half-year periods. The red dots indicate the empirically estimated split-half reliabilities. The blue lines show the predicted reliabilities based on the split-half reliability for 25 TIP events (red solid dot).

Figure 2 shows the split-half reliability of the TIP performance for 20, 50, 100, or 345 TIP events per half-year, with the latter corresponding to the average number of TIP events inspected by a screener per half-year.

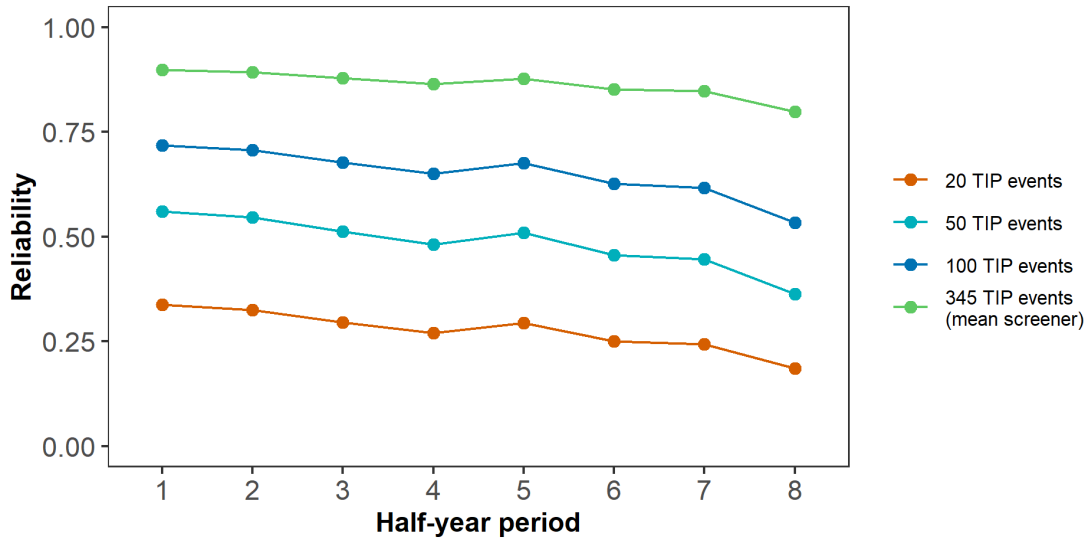


Figure 2. Reliability values for 20, 50, 100, and 345 TIP events (mean number of TIP events per screener per half-year period) for eight half-year periods.

As shown, reliability decreased over time. Decomposing the reliability into true variance and standard error (Figure 3B and C) shows that the decrease in reliability was not attributable to an increasing standard error (which also decreased over time); rather, the reliability decreased because of an over-proportionate decrease in the true variance of the TIP performance between the screeners. As shown in Figure 3C, the average TIP performance increased over time, which might have led to a limited room for true variance (i.e., a ceiling effect).

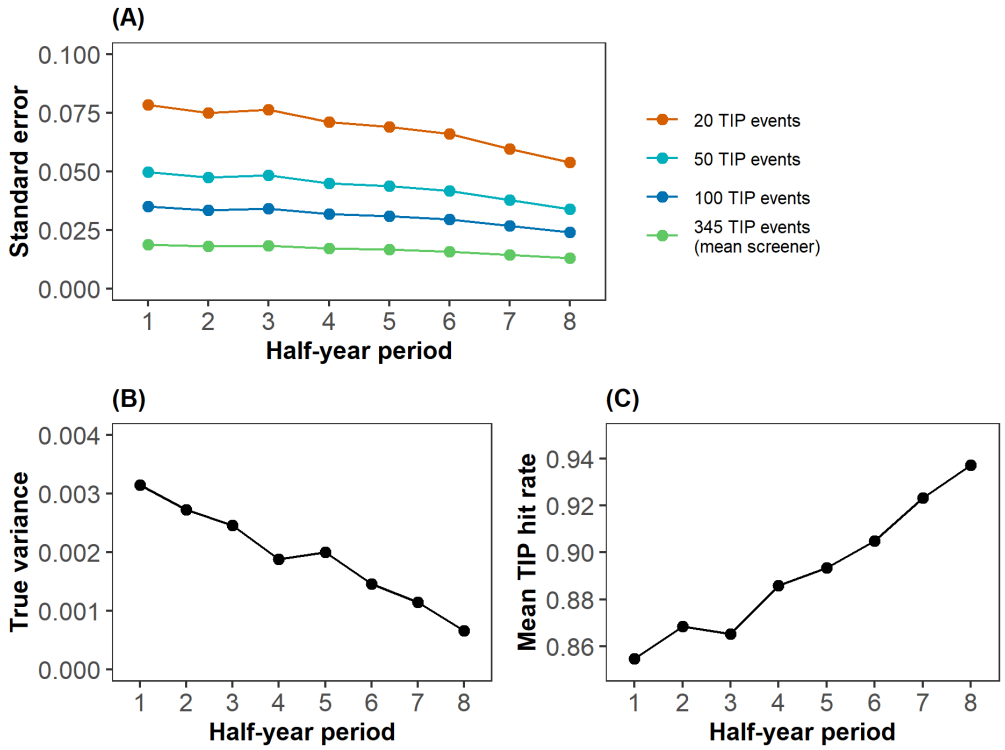


Figure 3. Mean hit rate (A),variance between screeners (B), and the standard error (C) for eight half-year periods.

Table 1 shows the necessary number of TIP events to reach a reliability of either 0.70, 0.75, or 0.80 for each half-year period. Although considering 92 TIP images for performance evaluation was sufficient to achieve a minimum reliability of 0.7 in the first half-year, 205 TIP images were necessary to obtain an equal reliability in the eighth half-year.

Time period	N TIP required to achieve a reliability of 0.70 (% of screeners who did not analyze this number of images or more)	N TIP required to achieve a reliability of 0.75 (% of screeners who did not analyze this number of images or more)	N TIP required to achieve a reliability of 0.80 (% of screeners who did not analyze this number of images or more)
1	92 (16.38%)	118 (20.60%)	158 (25.06%)
2	97 (15.12%)	125 (19.76%)	167 (22.68%)
3	111 (16.20%)	143 (19.25%)	191 (23.47%)
4	126 (18.16%)	162 (21.52%)	216 (26.91%)
5	112 (16.23%)	144 (18.83%)	192 (21.21%)
6	140 (19.71%)	180 (23.86%)	239 (30.50%)
7	145 (27.00%)	187 (34.20%)	249 (46.40%)
8	205 (56.34%)	263 (70.62%)	351 (84.91%)
Average over time frames	129	165	220

Table 1. Number of TIP events required per screener to achieve reliable performance measurement for each time frame. The brackets indicate the percentage of screeners who did not analyze the number of required TIP image or more.

Figure 4 illustrates the 95% confidence intervals (95%-CIs) for a reliability of .7 and .8 and a hit rate of .80, based on the average true variance across the eight half-year periods. The 95%-CI includes the true TIP performance with a 5% probability of error. As can be seen, a higher reliability results in a smaller 95%-CI (at constant true variance).

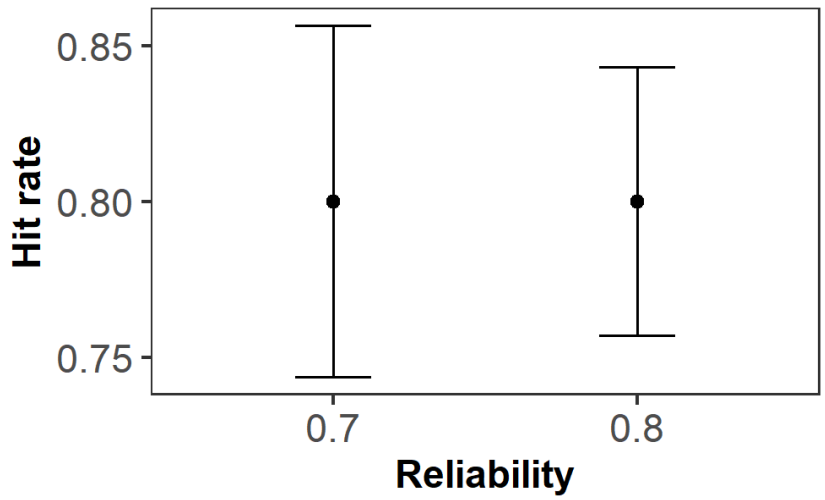


Figure 4. 95% CI for an average hit rate of 0.8 and a reliability of 0.7 (left) or 0.8 (right).

3.2 Validity of TIP data

The average covert test hit rate over all screeners was 79.50 % (*SD* = 0.40). The GEE with covert test performance as a dependent measure revealed that screeners with a better TIP performance also showed higher covert test performance (Table 2). Figure 5 shows the estimated relationship between the TIP hit rate and covert test performance. In total, 1,194 covert tests were considered, and on average, 826 TIP events were considered per covert test per screener (*SD* = 323.58). As shown in Figure 5, screeners with a higher TIP performance also showed a higher covert test performance. TIP data predicted the covert test performance well (odd ratio = 1.481; the odds of passing a covert test increase by about 48% for an increase in the TIP hit rate by one standard deviation).

Table 2. GEE results for covert test performance

Term	β	Robust SE	Robust z	Odd ratio	p
Intercept	1.477	0.247	5.939	4.378	< 0.001
TIP hit rate scaled	0.393	0.084	4.678	1.481	< 0.001
Complexity: 2	-1.080	0.213	-5.073	0.339	< 0.001
Complexity: 3	-2.085	0.279	-7.469	0.124	< 0.001
Complexity: unknown	-0.211	0.237	-0.892	0.810	0.372
Category: bomb	0.114	0.224	0.507	1.120	0.612
Category: knife	-0.633	0.298	-2.128	0.531	0.034
Category: other	-0.682	0.247	-2.759	0.506	0.006
Checkpoint: gates	0.465	0.201	2.314	1.592	0.021
Checkpoint: staff & VIP	0.582	0.217	2.683	1.790	0.007
Checkpoint: transfer	1.362	0.273	4.976	3.903	< 0.001
Machine type 2	0.327	0.199	1.645	1.386	0.100

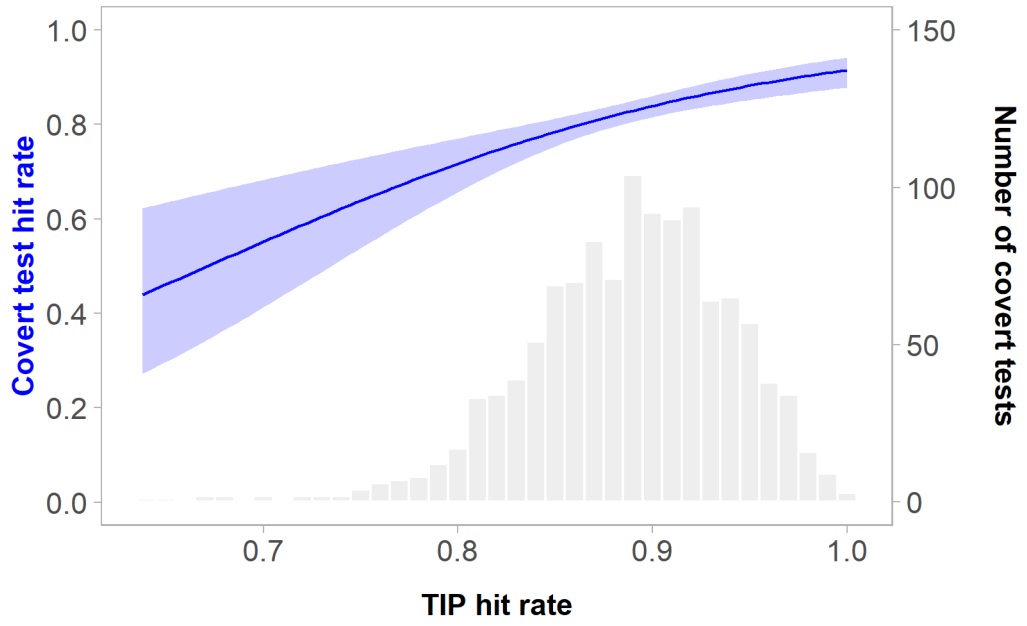


Figure 5. Relationship between covert test and TIP hit rate (blue line). 95% confidence band indicated by the blue area around the blue line. The histogram shows the distribution of the number of conducted covert tests.

5. Discussion

TIP data are used worldwide for quality control at airports and governments and security companies. At airports, if screeners do not achieve a minimum hit rate, they must undergo remedial training and successfully complete a recertification process before being allowed to continue screening during operation (Bassetti, 2018; Riz à Porta et al., 2022). We investigated whether TIP data provides a reliable and valid measure of detection performance by analyzing data from an international airport covering four years. We found that reliability increased with the number of TIP events following the Spearman–Brown prediction. This finding is important because the reliability of a TIP system can be estimated for a certain number of TIP events per screener (e.g., 20 events), and the data can be extrapolated to calculate the necessary number of TIP events to achieve the desired reliability. Based on this method, we found that with a good TIP library, approximately 100 TIP events were sufficient to achieve a minimum reliability of 0.7. This reliability value is recommended if the measure is used as a first indication (e.g., dividing the screeners into two performance groups) or for group diagnostics. However, if performance measures have consequences for screeners (e.g., mandatory remedial training), it is highly recommended to achieve higher reliability values of at least 0.8 (Brough, 2019; Murphy and Davidshofer, 2014). To achieve this, our results suggested that approximately 170 TIP events were required, assuming that a good-quality TIP library is used. When estimating reliability based on a randomized split of the TIP data, we found resampling to be important because a single estimate can deviate strongly from the mean estimate. Therefore, we recommend resampling for a randomized split-half reliability estimation.

Our analyses also showed decreasing reliability throughout the four-year period. Although the error variance also decreased over time, there was a disproportionate decrease in the true variance between screeners, causing a decrease in reliability. It seems likely that the decline in the true variance was caused by an increase in the average hit rate, leading to a ceiling effect. With an increasing number of screeners reaching a TIP performance close to the possible maximum, TIP performance can no longer be distinguished between these high-performing screeners. Ceiling effects are a known reliability issue in performance measurement; for example, in competency assessment tests. The increase in the average hit rate may have been caused by improvements in screeners' detection ability or, perhaps more likely, by an increase in familiarity with FTIs.

Reliability informs the share of variance attributable to true variance as opposed to the variance because of measurement error and, therefore, not only depends on the amount of measurement error (error variance) but also on how much individuals differ (true variance). If one is more interested in the absolute TIP performance than in the comparison between individuals or groups, more informative standard errors and CIs can be derived from the estimated reliability and variance across screeners. For example, Figure 4 shows the 95% CI for a reliability of 0.7 or 0.8 and the average true variance across all periods of our data set. These 95% CI show the interval in which the true TIP performance lies when accepting a 5% probability of error. As can be seen in Figure 4, the 95% CI decreases when reliability increases.

To achieve high reliability, its strong dependency on the difficulty of the TIP images should be considered. Our results showed a decrease in reliability over time for a constant number of TIP events, which was likely because of an increase in the hit rate. To avoid this, a proportion of FTIs should be exchanged regularly (e.g., 10% every year) to prevent overlearning of images. TIP libraries should be large enough to ensure that screeners do not view the same prohibited items too

often. If the average hit rate of screener reaches very high values (such as 95%), then the TIP library should be exchanged. Furthermore, it is advisable to check them for artifacts (Riz à Porta et al., 2022) to avoid TIP images from being too easy. For a given TIP system, reliability can be improved by increasing the number of TIP events used for performance assessment by extending the evaluation period or by increasing the TIP rate (the percentage of baggage images selected for TIP).

For TIP data to provide a useful assessment of detection performance, it must be reliable and valid; TIP performance must reflect the performance in detecting real prohibited articles. Bassetti (2018) reported that screeners sometimes recognize TIP images because they appear artificial. Riz à Porta et al. (2022) found that a third of TIP images from an international airport look unrealistic. However, with two-thirds of the images looking realistic, these researchers concluded that TIP performance should still achieve its purpose and, to a large extent, reflect the performance in detecting real prohibited articles. Consistent with this view, our analysis found that TIP performance was significantly associated with detection performance in covert tests. In other words, screeners who performed better in detecting TIP were more likely to detect prohibited items in covert tests. It must be noted that this finding provides the first validation of TIP performance. However, our results do not indicate that the TIP is perfectly realistic. For instance, we found that the hit rate was higher in the TIP test than in the covert test. This is consistent with previous findings that TIP produces a share of unrealistic and easy images (Riz à Porta et al., 2022). Despite these differences in difficulty, our results showed that TIP has predictive validity and distinguishes between screeners with high and low detection performance.

When discussing TIP performance, one must be aware that it does not fully reflect the screeners' detection ability. In a detection task such as X-ray image inspection, the hit rate depends

not only on searchers' target-detection ability, but also on their response tendency (Green and Swets, 1966; Macmillan and Creelman, 2005). Screeners can increase their TIP hit rate by indicating that they suspect a threat at the cost of a higher false alarm rate more frequently. Therefore, assessing the screener detection ability should ideally consider both hit and false alarm rates. For many TIP systems, including those used in our study, the false alarm rate is unavailable (only the rejection rate is measured, which includes all bags selected for manual search because of prohibited items and false alarms). Therefore, studies based on TIP data are often limited to the hit rate (Hofer and Schwaninger, 2005; Meuter and Lacherez, 2016; Skorupski and Uchroński, 2016). However, even though the TIP hit rate does not provide complete information about the ability to detect threats, it is not immediately relevant whether threats are found owing to high detection ability or response tendency from a security perspective, and the hit rate is sufficient to evaluate the achieved security performance (Hofer and Schwaninger, 2005; Macmillan and Creelman, 2005).

A limitation of our study is that we could only analyze the reliability and validity of TIP data from one airport. It would be interesting to continue this research with TIP data from other airports, using different TIP systems and screening technologies. In addition to the analysis of FTI TIP (as in our study), it would be interesting to investigate the use of combined threat images (images of baggage with integrated prohibited items) (Hofer and Schwaninger, 2004). As such images consist of both baggage and prohibited items, they can be carefully prepared, and unrealistic images can be excluded beforehand. It can therefore also be expected that reliability in TIP using combined threat images is higher when compared to FTI TIP. By showing fully prepared images, images without prohibited items can be projected to assess the false alarm rate (Hofer and Schwaninger, 2004).

Despite these limitations, our results provide valuable information for the calculation of the reliability of TIP data. We showed that the Spearman–Brown formula can be used to calculate the number of TIP events required to achieve the desired reliability. Our results suggest that with a good TIP library, approximately 100 TIP events are sufficient to achieve a reliability of 0.7 and approximately 70% more TIP events are sufficient for a reliability of 0.8. Care should be taken when TIP performance increases over years; exchanging TIP images regularly is recommended. When interpreting absolute TIP performance, CIs provide useful information; they decrease with increasing reliability. Our study is the first to investigate the validity of TIP data. We found clear evidence that higher TIP scores are associated with better covert test performance. Both operational performance assessments complement each other; TIP allows exposing screeners to prohibited items frequently, whereas covert tests are important to ensure that screeners react appropriately. Our study also highlights the importance of maintaining a good TIP library and avoiding ceiling effects by regularly exchanging TIP images to draw reliable conclusions about detection performance.

Acknowledgements

The authors wish to thank the representatives of the airport who provided the data and Silvie Hrbková for her support in data preparation and analysis.

References

Ballinger, G.A., 2004. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organ. Res. Methods 7, 127–150. https://doi.org/10.1177/1094428104263672

Bassetti, C., 2021. The tacit dimension of expertise: Professional vision at work in airport security. Discourse Stud. 23, 597–615. https://doi.org/10.1177/14614456211020141

Bassetti, C., 2018. Airport security contradictions: Interorganizational entanglements and changing work practices. Ethnography 19, 288–311. https://doi.org/10.1177/1466138117696513

Bassetti, C., Ferrario, R., Campos, M.L.M., 2015. Airport security checkpoints: An empirically-grounded ontological model for supporting collaborative work practices in safety critical environments. ISCRAM 2015 Conf. Proc. - 12th Int. Conf. Inf. Syst. Cris. Response Manag. 2015-Janua.

Brough, P., 2019. Advanced Research Methods for Applied Psychology, Advanced Research Methods for Applied Psychology. Routledge. https://doi.org/10.4324/9781315517971

Brown, W., 1910. Some Experimental Results in the Correlation of Mental Abilities. Br. J. Psychol. 1904 - 1920 3, 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Buser, D., Schwaninger, A., Sauer, J., Sterchi, Y., 2022. Time on task and task load in visual inspection: a four-months field study with baggage screeners. Appl. Ergon.

Cutler, V., Paddock, S., 2009. Use of Threat Image Projection (TIP) to enhance security performance, in: 43rd Annual 2009 International Carnahan Conference on Security Technology, Zurich. pp. 46–51. https://doi.org/10.1109/CCST.2009.5335565

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. [WWW Document]. Off. J. Eur. Union. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R1998

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

Green, D.M., Swets, J.A., 1966. Signal detection theory and psychophysics, Wiley & Sons, Inc. https://doi.org/10.1901/jeab.1969.12-475

Harris, D.H., 2002. How to really improve airport security. Ergon. Des. https://doi.org/10.1177/106480460201000104

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

Hofer, F., Schwaninger, A., 2004. Reliable and valid measures of threat detection performance in X-ray screening. Proc. - Int. Carnahan Conf. Secur. Technol. 303–308. https://doi.org/10.1109/ccst.2004.1405409

Kline, P., 2000. The Handbook of Psychological Testing, Second. ed. Routledge.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Macmillan, N.A., Creelman, D.C., 2005. Detection Theory: A User's Guide., 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Menneer, T., Donnelly, N., Godwin, H.J., Cave, K.R., 2010. High or low target prevalence increases the dual-target cost in visual search. J. Exp. Psychol. Appl. 16, 133–144.

https://doi.org/10.1037/a0019569

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

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., Schwaninger, A., 2014. A multi-method approach towards identifying situational factors and their relevance for X-ray screening. Proc. 48th IEEE Int. Carnahan Conf. Secur. Technol. 208–213. https://doi.org/10.1109/CCST.2014.6987001

Murphy, K.R., Davidshofer, C.O., 2014. Psychological testing. Principles and Applications, Sixth. ed, British Library Cataloguing-in-Publication Data. Pearson Education Limited.

Rindskopf, D., 2015. Reliability: Measurement, Second Edi. ed, International Encyclopedia of the Social & Behavioral Sciences: Second Edition. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44050-X

Riz à Porta, R., Sterchi, Y., Schwaninger, A., 2022. How Realistic Is Threat Image Projection for X-ray Baggage Screening ? Sensors 22, 2220.

Schwaninger, A., 2009. Why do airport security screeners sometimes fail in covert tests? Proc. - Int. Carnahan Conf. Secur. Technol. 41–45. https://doi.org/10.1109/CCST.2009.5335568

Schwaninger, A., 2006. Threat Image Projection: enhancing performance? Aviat. Secur. Int. 36–41.

Skorupski, J., Uchroński, P., 2018. Evaluation of the effectiveness of an airport passenger and baggage security screening system. J. Air Transp. Manag. 66, 53–64.

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

Spearman, C., 1910. Correlation Calculated From Faulty Data. Br. J. Psychol. 1904‐1920 3, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Walter, S., Hofer, F., Dolder, Z., Ghelfi-Waechter, S., 2021. The why and how of security drills at the security checkpoint. J. Airpt. Manag. 15, 147–159.

Wetter, O.E., Hardmeier, D., Hofer, F., 2008. Covert testing at airports. Proc. - Int. Carnahan Conf. Secur. Technol. 357–363.

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

Wolfe, J.M., Horowitz, T.S., Van Wert, M.J., Kenner, N.M., Place, S.S., Kibbi, N., 2007. Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks. J. Exp. Psychol. Gen. 136, 623–638. https://doi.org/10.1037/0096-3445.136.4.623