Universität Basel

Fakultät für Psychologie

Human Factors in X-ray Image Inspection of Passenger Baggage – Basic and Applied Perspectives

Inaugural Dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy to the Department of Psychology,
of the University of Basel
by
Yanik Sterchi
from Lützelflüh, BE, Switzerland

Basel, 2019

Approved by the Department of Psychology

At the request of

Prof. Dr. Klaus Opwis (First Reviewer)

Prof. Dr. Adrian Schwaninger (Second Reviewer)

Basel, Switzerland, _____

Prof. Dr. Alexander Grob, Dean of the Faculty of Psychology

Statement of Authorship

1. I, Yanik Sterchi, hereby declare that I have written the submitted doctoral thesis “Human Factors in X-ray Image Inspection of Passenger Baggage – Basic and Applied Perspectives” without assistance from third parties not indicated.
2. I only used the indicated resources.
3. I marked all citations.
4. My cumulative dissertation is based on six manuscripts. Manuscripts 1, 2, 3, 5, and 6 have already been published. Manuscript 4 has been submitted and it has received positive reviews. It is currently under revision for resubmission to the same journal. For Manuscripts 2, 3, 4 and 6 I had the leading role in conceptualization, project management, review, data analysis, article writing and manuscript preparation. In Manuscript 4, the lead was shared with Daniela Buser. For Manuscript 2 and 3, the co-authors significantly contributed to project management, article writing and manuscript preparation. For Manuscript 1, Nicole Hättenschwiler and Sarah Merks had the leading role while I contributed significantly to the conceptualization, the data analysis, article writing and manuscript preparation. For Manuscript 5, Nicole Hättenschwiler had the leading role and I contributed significantly to the literature review, data analysis and interpretation, article writing and manuscript preparation.

Inhalt

Abstract

The X-ray image inspection of passenger baggage contributes substantially to aviation security and is best understood as a search and decision task: Trained security officers – so called screeners – search the images for threats among many harmless everyday objects, but the recognition of objects in X-ray images and therefore the decision between threats and harmless objects can be difficult. Because performance in this task depends on often difficult recognition, it is not clear to what extent basic research on visual search can be generalized to X-ray image inspection. Manuscript 1 of this thesis investigated whether X-ray image inspection and a traditional visual search task depend on the same visual-cognitive abilities. The results indicate that traditional visual search tasks and X-ray image inspection depend on different aspects of common visual-cognitive abilities. Another gap between basic research on visual search and applied research on X-ray image inspection is that the former is typically conducted with students and the latter with professional screeners. Therefore, these two populations were compared, revealing that professionals performed better in X-ray image inspection, but not the visual search task. However, there was no difference between students and professionals regarding the importance of the visual-cognitive abilities for either task.

Because there is some freedom in the decision whether a suspicious object should be declared as a threat or as harmless, the results of X-ray image inspection in terms of hit and false alarm rate depend on the screeners' response tendency. Manuscript 2 evaluated whether three commonly used detection measures – d', A', and d_a – are a valid representation of detection performance that is independent from response tendency. The results were consistently in favor of d_a with a slope parameter of around 0.6. In Manuscript 3 it was further shown that screeners can change their response tendency to increase the detection of novel threats. Also, screeners with a high ability to recognize everyday objects detected more novel threats when their response tendency was manipulated.

The thesis further addressed changes that screeners face due to technological developments. Manuscript 4 showed that screeners can inspect X-ray images for one hour straight without a decrease in performance under conditions of remote cabin baggage screening, which means that X-ray image inspection is performed in a quiet room remote from the checkpoint. These screeners did not show a lower performance, but reported more distress, compared to screeners who took a 10 min break after every 20 min of screening.

Manuscript 5 evaluated detection systems for cabin baggage screening (EDSCB). EDSCB only increased the detection of improvised explosive devices (IEDs) for inexperienced screeners if alarms by the EDSCB were indicated on the image and the screeners had to decide whether a threat was present or not. The detection of mere explosives, which lack the triggering device of IEDs, was only increased if the screeners could not decide against an alarm by the EDSCB. Manuscript 6 used discrete event simulation to evaluate how EDSCB impacts the throughput of passenger baggage screening. Throughput decreased with increasing false alarm rate of the EDSCB. However, fast alarm resolution processes and screeners with a low false alarm rate increased throughput.

Taken together, the present findings contribute to understanding X-ray image inspection as a task with a search and decision component. The findings provide insights into basic aspects like the required visual-cognitive abilities and valid measures of detection performance, but also into applied research questions like for how long X-ray image inspection can be performed and how automation can assist with the detection of explosives.

Introduction

The screening of passenger baggage is a crucial element in aviation security and it is typically done by recording X-ray images of the baggage, which are then analyzed by security officers (screeners). This screening is not a trivial task and mistakes can have fatal consequences. The terrorist attacks of 11 September 2001 painfully brought these consequences to public awareness and as a result a growing body of research on X-ray image inspection of passenger baggage emerged (for recent reviews see for example Biggs, Kramer, & Mitroff, 2018; Biggs & Mitroff, 2015). When inspecting X-ray images, screeners search for prohibited items among many harmless everyday objects. In that sense, X-ray image inspection might be seen as visual search task: the act of looking for targets amongst an array of distractors (e.g. Treisman & Gelade, 1980). When visual search aims at finding familiar targets, it relies on object recognition (Wolfe, 1998). Compared to other visual search tasks, performance in X-ray image inspection might depend even more strongly on object recognition. The recognition of objects in our everyday life is often effortless and almost always accurate (Kosslyn, 1980). In an X-ray image, the color and transparency of objects depend on the effective atomic number and material density (Singh & Singh, 2003). Therefore, objects can look very different in an X-ray image compared to everyday life (Halbherr, Schwaninger, Budgell, & Wales, 2013). In addition, some of the threat items are not known from everyday life (e.g. an improvised explosive device, IED) and other threat items can look similar to harmless objects (e.g. a knive can resemble a pen; Schwaninger, 2005). Objects in X-ray images of passenger bags are therefore more ambiguous and difficult to detect for untrained individuals (Halbherr et al., 2013; Koller, Hardmeier, Michel, & Schwaninger, 2008). Because performance in X-ray image inspection depends heavily on the recognition of ambiguous targets, it is better described as consisting of a *search component* and a *decision component* that comprises the decision whether a target object is a threat or harmless, which includes recognition processes (Koller, Drury, & Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, & Horne, 2009).

Considering that performance in X-ray image detection likely depends more strongly on the recognition of ambiguous objects, it is not clear to what extent the vast body of research on traditional visual search generalizes to X-ray image inspection (for reviews see e.g. Carrasco, 2011, 2014, 2018; Chan & Hayward, 2013; Eckstein, 2011; Humphreys & Mavritsaki, 2012; Nakayama & Martini, 2011). In Manuscript 1, we therefore investigated whether traditional visual search and X-ray image inspection rely on the same visual-cognitive abilities. To this end, potentially relevant visual-cognitive abilities were derived from the literature and it was investigated whether these abilities are comparably relevant for X-ray image inspection and a traditional visual search task. Research on traditional visual search is typically conducted with student samples, whereas participants in research on X-ray image inspection are mostly professionals. Therefore, the study of Manuscript 1 was conducted with students and professionals to allow a direct comparison.

Early psychophysical experiments that investigated the detection of weak auditory or visual signals encountered a challenge: How study participants responded did not only depend on sensory factors, but also on their decision process (Green & Swets, 1966). Under the term of signal detection theory (SDT), Green and Swets (1966) established a theoretical foundation with methods to conduct experiments and analyze data that aimed at separating detection performance and response tendency. Since then, SDT has been extended and broadly applied to problems beyond psychophysics (Macmillan & Creelman, 2005). Also X-ray image inspection includes a decision component and screeners can shift their response tendency, i.e. they can change the frequency of *target present* responses in relation to *target absent* responses. Thereby, they unidirectionally change their hit rate and false alarm rate. This can make it difficult to compare the performance between individuals, groups, or conditions that have different response tendencies, which is not unlikely when evaluating different conditions of X-ray image inspection (e.g. Hattenschwiler, Merks, & Schwaninger, 2018). It seems plausible that study participants can feel less confident in their decisions when they are tested with images or technology that are unfamiliar. Such a change in confidence can affect response tendency. In addition,

other factors like the relative frequency with which targets occur – the *target prevalence* – and the respective cost of hits and false alarms can have an impact (Macmillan & Creelman, 2005). Performance in X-ray image inspection is therefore often compared based on *detection measures* that are supposed to be independent of response tendency, like *d′* or *A′* (e.g., Brunstein & Gonzalez, 2011; Halbherr, Schwaninger, Budgell, & Wales, 2013; Ishibashi, Kita, & Wolfe, 2012; Madhavan, Gonzalez, & Lacson, 2007; Mendes, Schwaninger, & Michel, 2013; Menneer, Donnelly, Godwin, & Cave, 2010; Rusconi, Ferri, Viding, & Mitchener-Nissen, 2015; Schwaninger, Hardmeier, Riegelnig, & Martin, 2010; Yu & Wu, 2015). However, several studies raise doubt on the validity of *d′* and *A′* (Godwin, Menneer, Cave, & Donnelly, 2010; Hofer & Schwaninger, 2004; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010). We investigated the validity of these detection measures with two experiments in Manuscript 2 and one experiment in Manuscript 4.

That response tendency can shift could also be useful for practical purposes. If screeners can willingly shift their response tendency, they could e.g. inspect the baggage of high-risk passengers more thoroughly. It could also be useful for detecting novel threats. Maybe screeners are able to shift their response tendency to a point where bags are only declared as harmless if the screener is certain that the bag only contains harmless everyday objects. How effective such an inspection strategy is, likely depends on a screener's ability to recognize everyday objects. Manuscript 3 therefore investigated how well screeners can detect known and novel threats by applying different inspection strategies. It was further investigated how performance depends on the ability to recognize everyday objects in X-ray images and whether an e-learning module could assist with the application of the new inspection strategy.

As described above, Manuscript 4 built upon Manuscript 3 and also investigated the validity of detection measures. To change response tendency, the share of images that contain threats – the target prevalence – was manipulated (compare e.g. Macmillan & Creelman, 2005; Wolfe et al., 2007)Manuscript 4 further investigated a research question that has become more prominent because of technological developments in the screening of passenger baggage: the effect

of time on task and breaks on screening performance. More and more airports pool X-ray images on servers and then redistribute them to screeners in order to increase capacity and reduce cost (Kuhn, 2017). Removing the one-to-one relationship between machine and screeners also created the option to conduct the X-ray image inspection remotely from the checkpoint in a separate office. This so called *remote cabin baggage screening* (RCBS) has the potential advantage of reduced distractions and noise, but isolates the screeners from their team members who work the other positions of the airport security checkpoint (instructing passengers on how to load trays, manually searching bags, etc.). This does not only limit communication, but also makes it more difficult to rotate from the screening position to other positions – which currently takes place every 20 min at most European airports because the screening duration is regulated to this limit (Commission Implementing Regulation [EU], 2015/1998). It is still unclear how the work should be optimally designed with RCBS. It is also unclear whether regulation should constrain the options of work design by limiting screening to 20 min. This limit is likely based on research into vigilance (personal communication with airport security expert, fall 2017), for which many studies have revealed decreases in vigilance within the first 15–30 min of the task (Mackworth, 1948; Nuechterlein et al., 1983). Manuscript 4 therefore investigated in an experiment how performance in X-ray image inspection develops over 60 min. Thereby, one group of participants had a 10-min break every 20-min according to current EU regulation (Commission Implementing Regulation [EU], 2015/1998) and each participant completed the task twice with different levels of target prevalence.

IEDs are one of the biggest concerns for aviation security (Baum, 2016; Novakoff, 1993; Singh & Singh, 2003). A fully functional IED consists of triggering device, a power source, and a detonator, which are typically connected by wires, plus explosive material (Wells & Bradley, 2012). Through training, screeners can learn to recognize these components and thereby achieve a high detection of IEDs (e.g. Halbherr et al., 2013; Koller et al., 2008). Bare explosives, however, lack these distinctive features and often look like harmless organic mass (Jones, 2003), which makes them more difficult to detect. To assist with the detection of IEDs

and explosives, so called explosive detection systems for cabin baggage (EDSCB) have been developed. EDSCB use high and low energy X-ray from different angles to estimate the effective atomic number and material density at each location within the bag and match it against these attributes from explosives (Singh & Singh, 2003). A difficulty with EDSCB is that it generates a high number of false alarms (15–20% according to personal communication with EDSCB experts, summer 2016) and it is unclear how to best resolve them. One option is that the screener is informed about an alarm and then has to decide whether a bag contains a threat or not. Thereby, a “cry wolf” effect might occur with operators ignoring system warnings due to many false alarms and only few hits, as it has been found in other domains (Breznitz, 1983; Bliss, 2003). EDSCB could alternatively be implemented with a higher level of automation by automatically diverting all bags that trigger an alarm to secondary search. Whereas this eliminates the potential problem of a cry wolf effect, a secondary search has to be conducted for each false alarm by the EDSCB, which usually consists of a manual search or explosives trace detection (ETD). Manuscript 5 therefore evaluated two different levels of automation for EDSCB: on-screen alarm resolution and automated decision. Familiarity with automation can affect how people interact with it (Parasuraman, Sheridan, & Wickens, 2000; Sauer, Chavaillaz, & Wastell, 2016; Strauch, 2017). Whereas participants of the first experiment were not familiar with automation aids, the second experiment was conducted with participants that were. Furthermore, the selection of an optimal level of automation depends on human performance (Parasuraman et al., 2000) and screener performance depends on job experience (Halbherr et al., 2013). The second experiment therefore also took job experience into account.

The question of how to implement EDSCB not only concerns detection performance, but also how the throughput of checkpoints is affected by the false alarms of the EDSCB. Manuscript 6 therefore used discrete event simulation to investigate how EDSCB with automated decision affects throughput depending on the false alarm rate of the EDSCB, the false alarm rate of the screener, and the duration of alarm resolution with explosives trace detection (ETD). It was also

evaluated how an additional security officer and the screener helping with secondary search would affect the impact of the EDSCB.

Summary of the Manuscripts

The following manuscripts constitute this thesis. Manuscript 1, 2, 3, 5 and 6 have already been published. Manuscript 4 has been submitted and it has received positive reviews. It is currently under revision for resubmission to the same journal.

1. 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, 1-17 https://doi.org/10.3389/fpsyg.2019.00525
2. Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (2019). Detection measures for visual inspection of X-ray images of passenger baggage. *Attention, Perception, & Psychophysics*, 1-15. doi:10.3758/s13414-018-01654-8
3. Sterchi, Y., Hättenschwiler, N., Michel, S. & Schwaninger, A. (2017). Relevance of visual inspection strategy and knowledge about everyday objects for X-ray baggage screening. *Proceedings of the 51st IEEE International Carnahan Conference on Security Technology*, 1-6. doi:10.1109/CCST.2017.8167812
4. Buser, D.[1], Sterchi, Y.[1] & Schwaninger A. (2019) *Why stop after 20 minutes? Breaks and target prevalence in a one hour X-ray image inspection task.* Manuscript under revision.
5. Hättenschwiler, N., Sterchi, Y., Mendes, M., & Schwaninger, A. (2018). Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection. *Applied Ergonomics, 72*, 58-68. doi:10.1016/j.apergo.2018.05.003
6. Sterchi, Y., & Schwaninger, A. (2015). A first simulation on optimizing EDS for cabin baggage screening regarding throughput. *Proceedings of the 49th IEEE International Carnahan Conference on Security Technology*, 55-60. doi:10.1109/CCST.2015.7389657

[1] Joint first authorship

The following conference proceedings are related to this thesis, but were omitted for brevity.

1. Hättenschwiler, N., Sterchi, Y., Michel, S., & Schwaninger, A. (2017). Relevanz von Wissen über Alltagsgegenstände und visueller Inspektionsstrategie für die Gepäckkontrolle mit Röntgengeräten. *63. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft (GfA)*, 1-5.
2. Riz à Porta, R., Sterchi, Y., & Schwaninger, A. (2018). Examining threat image projection artifacts and related issues: A rating study. *Proceedings of the 52th IEEE International Carnahan Conference on Security Technology*, 1-4.
3. Sterchi, Y., & Schwaninger, A. (2016). Eine erste Simulation zur Optimierung des Durchsatzes beim Einsatz automatischer Sprengstoffdetektion für die Handgepäckkontrolle an Flughäfen.*62. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft (GfA)*, 1-6. doi:10.13140/RG.2.1.4772.7604
4. Wyssenbach, T., Sterchi, Y., & Schwaninger, A. (2017). Simulationsunterstützte soziotechnische Optimierung von Luftsicherheitskontrollstellen. *63. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft (GfA)*, 1-6.

Manuscript 1: Traditional visual search versus X-ray image inspection in students and professionals: Are the same visual-cognitive abilities needed?

Motivation and aim of the study. Visual search, i.e. the act of looking for targets amongst an array of distractors is a cognitive task that has been studied extensively over several decades (for reviews see e.g. Carrasco, 2011, 2014, 2018; Chan & Hayward, 2013; Eckstein, 2011; Humphreys & Mavritsaki, 2012; Nakayama & Martini, 2011) and has many real-world applications. Research shows that specific visual-cognitive abilities are needed to efficiently and effectively locate a target among distractors. It is, however, not clear whether the results from such traditional, simplified visual search tasks extrapolate to real-world inspection tasks in which professionals search for targets that are more complex, ambiguous, and/or less salient (e.g. Biggs & Mitroff, 2015; Radvansky & Ashcraft, 2016).

A known example of a traditional visual search task that has been studied in many variations is the *L/T-letter search task*, in which participants are e.g. asked to identify the perfectly shaped letter *T* (target) surrounded by many distractor letters including *Ls* and symmetrical and asymmetrical *Ts* (a so called conjunction search task; Treisman & Gelade, 1980). In professional, real world search tasks like X-ray image inspection, searchers must use their prior knowledge in order to accurately and efficiently locate more ambiguous targets (Wolfe, Cain, & Aizenman, 2019) such as guns and knives or cancer cells and so forth among distractors with much more complex features compared to a traditional conjunction search task. As mentioned above, searching for familiar targets relies on object recognition (Wolfe, 1998). Here, top-down processing allows searchers to more efficiently identify targets with greater complexity (Zhaoping & Frith, 2011). X-ray image inspection is therefore best described as a search and decision task (Koller et al., 2009; Spitz & Drury, 1978) that relies more heavily on the decision component compared to traditional search tasks with unambiguous targets. Nonetheless, visual search with complex objects is assumed to rely on the same active scanning processes as conjunction search (e.g. L/T-letter search task) with less complex, contrived laboratory stimuli (Alexander &

Zelinsky, 2011, 2012). The partial overlap between professional and laboratory search tasks raises the question whether they rely on the same visual cognitive abilities and whether findings can be easily transferred from traditional visual search to X-ray image inspection.

Today, the Cattell–Horn–Carroll theory (CHC) is widely accepted as a comprehensive and empirically supported theory on the structure of human cognitive abilities, and it informs a substantial body of research (McGrew, 2005). CHC differentiates between three hierarchical strata of abilities. Visual processing (*Gv*), short-term memory (*Gsm*), and processing speed (*Gs*) are Stratum II abilities that are accepted components with a known influence on visual search performance. Visual processing (*Gv*) describes a broad ability to perceive, analyze, synthesize, and think in visual patterns, including the ability to store and recall visual representations. Short-term memory (*Gsm*) is characterized as the ability to apprehend and hold information in immediate awareness and then perform a set of cognitive operations on this information within a few seconds. Because analyzing, synthesizing, and thinking in visual patterns are also cognitive operations, *Gv* and *Gsm* are closely related, but can be distinguished by the limited capacity of short-term memory. Processing speed (*Gs*) describes the ability to quickly and accurately perceive visual details, similarities, and differences.

Several studies have confirmed the influence of higher scores in *Gv*, *Gsm*, and *Gs* on better performance in traditional visual search tasks (Alvarez & Cavanagh, 2004; Eriksen & Schultz, 1979). Comparable cognitive abilities have also been linked to X-ray image inspection of professionals (Bolfing & Schwaninger, 2009; Hardmeier, Hofer, & Schwaninger, 2005; Hardmeier & Schwaninger, 2008; Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002) or visual inspection in general (e.g. Lavie & De Fockert, 2005; Poole & Kane, 2009; Roper, Cosman, & Vecera, 2013). However, a direct comparison between the visual-cognitive abilities needed for traditional and professional visual search is difficult to draw from these findings, because they are based on different tests.

When adopting findings from traditional visual search tasks for professional X-ray image inspection, also the different populations have to be considered. University students are usually

the first choice as participants for traditional visual search research (Clark, Cain, Adamo, & Mitroff, 2012). On the other hand, professional searchers are selected and trained to perform well in the inspection tasks they perform (Commission Implementing Regulation [EU], 2015/1998; Halbherr et al., 2013).

The goal of the study was to compare influence of *Gv*, *Gsm*, and *Gs* on performance between traditional and professional visual search tasks and between students and professionals.

Method. 128 students (age: M = 25.7, SD = 6.4; 74% female) and 112 professional screeners (age: M = 43.7, SD = 11.9; 55% female) completed 10 standardized test scales from established intelligence tests based on the CHC theory of intelligence (Carroll, 1993, 2003; Cattell, 1941; Horn, 1965; a more detailed description of the scales can be found in the manuscript) to measure *Gv*, *Gsm*, and *Gs*. In addition, they also completed the Raven Standard Progressive Matrices Plus (SPM), a language-independent test of fluid intelligence (Raven, Raven, & Court, 2003). All tests were computer-based. Afterwards, participants completed the L/T-letter search task. In line with Biggs, Cain, Clark, Darling, and Mitroff (2013), the test had an increasing difficulty level and a search and decision component.

In a second session about two weeks later, participants completed an X-ray image inspection task that was designed to be solvable by people without domain-specific knowledge. The test included 256 black-and-white X-ray images, one-half containing a threat item. As threats, guns and knives with common shapes were used. In each trial, an X-ray image of a piece of luggage was presented for a maximum of 4 s.

Results. In a first step, a confirmatory factor analysis was conducted to confirm the CHC-model structure of the visual-cognitive abilities. The overall model fit was good with Chi^2 (32) = 56.56, p = .005, CFI = .961, TLI = .946 and RMSEA = .036. The fit remained good when tested for each population separately. The correlation between the factors *Gs* and *Gsm* (r = 0.65, p < .001), as well as between *Gs* and *Gv* (r = .53 p < .001) was moderate, whereas there was a strong correlation between *Gsm* and *Gv* (r = .83, p < .001).

In a next step, we calculated multiple linear regression analyses to predict detection performance on the L/T-letter search task and the X based on the *z*-standardized summarized scale scores of *Gv*, *Gsm*, and *Gs* and group (students vs. professionals; see Table 1). For the performance of the L/T-letter search task, *d′* was used, for the performance of the X-ray image inspection task, analyses are reported as d_a, which is more appropriate for X-ray image inspection of passenger bags (as will be shown in Manuscript 2 and 4, and has been suggested by other studies; Godwin, Menneer, Cave, & Donnelly, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010). However, all analyses were also repeated with *d′* without resulting in meaningful differences. The analyses revealed a significant effect of *Gv* on performance in the L/T-letter search task and the X-ray inspection task and of *Gsm* on the X-ray image inspection task (see Table 1). The analyses further revealed no moderation effect of group (students vs. professionals), $F(3, 232) = 1.83$, $p = .143$. Furthermore, we found strong evidence against the moderation model using the Bayes Factor ($BF_{10} = 90.9$). A mediation analysis revealed that only a small part of the effect of *Gsm* and *Gv* on the X-ray inspection task was mediated by the L/T-letter search task, which means that different aspects of *Gsm* seem to be relevant for each of the two tasks.

Table 1

Multiple Linear Regression Analyses and Mediation Model for Detection Performance

(a) Basic Model	LT Task (d')				X-ray Inspection Task (d_a)			
	β	*SE*(β)	*t*-value	*p*-value	β	*SE*(β)	*t*-value	*p*-value
zGs	-.013	.078	-0.164	.870	-.039	.044	-0.893	.373
zGsm	.119	.079	1.513	.132	.104	.044	2.348	.019*
zGv	.299	.078	3.83	.000	.195	.044	4.463	.000
zGroup	.029	.070	-0.416	.678	-.834	.039	-21.37	.000
adj. R^2	.126				.726			

(b) Moderation Model								
	β	*SE*(β)	*t*-value	*p*-value	β	*SE*(β)	*t*-value	*p*-value
zGs	-.018	.079	-0.223	.823	-.03	.044	-0.675	.501
zGsm	.127	.080	1.567	.119	.113	.045	2.533	.012*
zGv	.286	.082	3.504	.001	.190	.045	4.132	.000
zGroup	-.028	.070	-0.4	.700	-.835	.040	-21.458	.000
zGs*zGroup	-.030	.079	-0.378	.705	.064	.044	1.461	.145
zGsm*Group	.036	.080	0.451	.652	.064	.045	1.426	.155
zGv*Group	-.034	.080	-0.418	.676	-.054	.045	-1.206	.229
adj. R^2	.117				.730			

(c) Mediation Model	β	*SE*(β)	*t*-value	*p*-value
zLT da	.13	.04	3.5	.000
zGs	-.04	.044	-0.88	.382
zGsm	.09	.044	2.05	.042*
zGv	.16	.044	3.58	.000
zGroup	-.83	.044	-21.77	.000
adj. R^2	740		.	

Conclusion. We investigated whether the same visual cognitive abilities predict performance in students and professionals performing two tasks: a traditional visual search task—the L/T-letter search task—and an X-ray image inspection task. We tested students and professionals on three known facets of visual-cognitive abilities: visual processing (*Gv*), short-term memory (*Gsm*), and processing speed (*Gs*). Visual processing (*Gv*) was a predictor of performance for

both tasks. This result is in accordance with earlier studies showing a correlation between performance and visual processing for traditional visual search and an influence of mental rotation and figure-ground segregation on higher performance in X-ray screening (Bolfing & Schwaninger, 2009; Wolfe et al., 2002), which are narrow abilities of visual processing (*Gv*). However, our results showed that different aspects of visual processing explain variance in the traditional visual search task and the X-ray image inspection task. Possible reasons are that targets in the traditional visual search task (*Ls* and *Ts*) have salient shapes, whereas targets (guns and knives) and distractors in the X-ray image inspection task are not salient and may additionally produce clutter and superposition. Short term memory (*Gsm*) was a significant predictor of X-ray image inspection performance, but not for the traditional visual search task. However, the standardized coefficient for *Gsm* was not smaller for the L/T-letter search task, but it did not reach significance as a predictor for the L/T-letter search task (due to larger standard errors) and its relevance for that task is therefore unclear.

For both groups, visual-cognitive abilities were comparably relevant for their performance on the traditional visual search task and the X-ray image inspection task. However, professionals outperformed students on the X-ray image inspection task. Because the relevance of the visual-cognitive abilities tested in this study proved to be independent of the population and they had similar levels of visual-cognitive abilities, the higher detection performance of the professionals in the X-ray image inspection task cannot be explained by differences in visual-cognitive abilities. Such a difference could be due to the selection of the security personnel or, more likely, job experience and training. Objects possibly need fewer recognized features in order to be identified successfully (Koller et al., 2009), and features are known and recognized better and faster with repeated exposure (Halbherr et al., 2013; Koller et al., 2009, 2008; McCarley, Kramer, Wickens, Vidoni, & Boot, 2004).

In summary, the study implies that X-ray image inspection and traditional visual search tasks both depend on visual-cognitive abilities, but not on the same aspects of those. Future studies with more statistical power should investigate the difference on the level of narrow abilities.

Also, the amount of variance in visual search performance explained by the investigated visual cognitive abilities suggests the presence of other predictors of performance, which should be investigated in the future.

Manuscript 2: Detection Measures for Visual Inspection of X-ray Images of Passenger Baggage

Motivation and aim of the study. The most direct way to assess detection performance in X-ray image inspection is to calculate the hit rate and the false alarm rate. However, the hit and false alarm rate change unidirectionally when the response tendency changes (Green & Swets, 1966; Macmillan & Creelman, 2005). To analyze detection performance independently from response tendency, researchers often use detection measures like *d′* and *A′* that are calculated from the hit and false alarm rate (e.g., Brunstein & Gonzalez, 2011; Halbherr, Schwaninger, Budgell, & Wales, 2013; Ishibashi, Kita, & Wolfe, 2012; Madhavan, Gonzalez, & Lacson, 2007; Mendes, Schwaninger, & Michel, 2013; Menneer, Donnelly, Godwin, & Cave, 2010; Rusconi, Ferri, Viding, & Mitchener-Nissen, 2015; Schwaninger, Hardmeier, Riegelnig, & Martin, 2010; Yu & Wu, 2015). Each of these detection measures imply a specific so called receiver operating characteristic (ROC) curve: the pairs of hit rate and false alarm rate values that keep the detection measure constant. Previous studies that looked into the effect of target prevalence (the share of target images) on the performance in X-ray image inspection suggest that *d′* is actually not independent of response tendency (Godwin, Menneer, Cave, & Donnelly, 2010; Hofer & Schwaninger, 2004; Van Wert, Horowitz, & Wolfe, 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010), instead d_a (Simpson & Fitter, 1973) with a slope parameter of 0.6 seems more valid. In terms of signal detection theory, d_a assumes that the noise and the signal-plus-noise distribution are of unequal variance, whereas both *d′* and *A′* are symmetric—any point (HR_x, FAR_x) leads to the same value of *d′* and *A′* as $(1 - HR_x, 1 - FAR_x)$—which implies equal variance (Macmillan & Creelman, 2005).

The aim of our study was to investigate the validity of the detection measures *d′*, *A′*, and d_a and to derive recommendations on how to calculate detection performance in future studies on X-ray image inspection. We explored this using two experiments, in which professional X-ray

screeners completed a simulated X-ray baggage inspection task. In the first experiment, response tendency (criterion) was manipulated through instruction to test whether it affected the detection measures. The experiment included targets that were known from training and targets that were novel, which resulted in two levels of sensitivity. In the second experiment, the participants provided confidence ratings that were used to investigate whether the ROC curves are approximately linear in *z*ROC space (i.e. when the hit and false alarm rate are transformed with the inverse of the cumulative distribution function of the standard normal distribution), as assumed by both d' and d_a, and to estimate the *z*ROC slope.

Method of Experiment 1. A total of 31 professional screeners (20 females, between 26 and 61 years old, M = 45.4, SD = 8.9, between 2 and 26 years of work experience, M = 8.4, SD = 5.5) from an international airport participated in this experiment. The experiment used a 2 × 2 design with two instructions to manipulate response tendency (normal decision vs. liberal decision) and with two levels of task difficulty (targets known from training vs. novel target items) as within-subject factors. Dependent variables were HR, FAR, d', d_a with a slope parameter of 0.6, A', A_g (a detection measure derived from confidence ratings), response times, and eye-tracking data.

The participants completed an X-ray image interpretation test that consisted of 128 X-ray images of passenger bags, 64 of which contained one prohibited item: 16 X-ray images contained a gun, 16 images a knife, 16 images an IED, and 16 images contained other prohibited items. For each category, half the images were known to the screeners from their computer based training. For eye tracking, we used an SMI RED-m eye tracker with a gaze sample rate of 120 Hz, gaze position accuracy of 0.5°, and spatial resolution of 0.1°. For half the test, participants were instructed to inspect (i.e., search and decide) the image as if they were working at a checkpoint (*normal decision*). For the other half, they were instructed to visually analyze each object in the X-ray image and decide that the bag was harmless only if each object in the image could be recognized as harmless (*liberal decision*). Participants then had to inspect each image

and decide whether its content was harmless or not by pressing a key, and then had to give a confidence rating on a 10-point scale ranging from 1 (*very unconfident*) to 10 (*very confident*).

Results of Experiment 1. A manipulation check revealed that 10 of the participants did not even show a small increase in the rejection rate (i.e., increase smaller than a Cohen's *d* of 0.20). Because we were interested in whether the detection measures change when participants change their response tendency (and not how successfully we could induce such a change), we excluded participants that did not change their rejection rate from further analysis.

Table 2 shows the dependent variables and effect sizes for known and novel threats and the two decision conditions. Exact permutation tests revealed a significantly lower d' in the liberal decision condition for both known ($p = .041$) and novel ($p = .002$) targets. Moreover, A' was significantly lower for both known ($p = .034$) and novel ($p = .017$) targets. For both d_a (known targets: $p = .714$, novel targets: $p = .383$) and A_g (known targets: $p = .322$, novel targets: $p = .750$), differences did not attain significance. For both target-present and target-absent trials, permutation tests indicated a higher response times for *liberal decision* compared to *normal decision* and (target-present trials: $p = .004$, target-absent trials: $p < .001$). The proportion of target trials, where participants fixated the target – an indicator for search errors (McCarley, 2009; Rich et al., 2008) – did not differ significantly between liberal decision and normal decision.

Conclusion Experiment 1. In line with studies that manipulated target prevalence (Godwin, Menneer, Cave, & Donnelly, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010) our results showed a decrease in d' and A' when participants were instructed to decide more liberally. This decrease is unlikely to be an actual decrease in detection performance but rather in response tendency, as suggested by the increased response times (in line with the argumentation of Wolfe et al., 2007) and by the eye-tracking data, which did not indicate changes in search errors (in line with the argumentation of McCarley, 2009; Rich et al., 2008).

Table 2

Mean (SD) of the Normal and Liberal Decision Condition and the Effect Size (Standardized Difference) of the Decision Condition for Hit Rate (HR), False Alarm Rate (FAR), and Detection Measures d', A', d_a, and A_g

Decision condition	HR	FAR	d'	d_a	A'	A_g
	Known targets					
Normal decision	.79 (.10)	.09 (.08)	2.25 (0.61)	2.03 (0.57)	.916 (.044)	.894 (.072)
Liberal decision	.90 (.10)	.25 (.13)	2.01 (0.58)	2.08 (0.61)	.899 (.049)	.906 (.073)
Effect size			-0.40	-0.08	-0.42	0.23
	Novel targets					
Normal decision	.58 (0.14)	.09 (.08)	1.63 (0.41)	1.28 (0.38)	.851 (.040)	.799 (.082)
Liberal decision	.71 (0.13)	.25 (.13)	1.27 (0.44)	1.19 (0.43)	.817 (.074)	.793 (.076)
Effect size			-0.70	-0.19	-0.50	-0.07

Method Experiment 2. In Experiment 1, we calculated *d', A'*, and d_a, for which we set the slope to 0.6 based on previous findings (Godwin, Menneer, Cave, & Donnelly, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010). d_a was found to be a more valid detection measure than *d'* and *A'*. However, Experiment 1 did not allow for a precise estimation of the slope parameter. Further, 10 of the participants were excluded because they failed the manipulation check, which might have biased the sample. Experiment 2 was therefore intended to provide a more precise estimation of the slope parameter and to further investigate the validity of detection measures using another methodological approach: multiple ROC points were obtained by analyzing confidence ratings. Therefore, 124 professional screeners (68 female; between 22 and 64 years old, *M* = 44.3, *SD* = 11.2, one participant did not report his/her age; up to 29 years of work experience, *M* = 7.1, *SD* = 5.6, seven participants did not report their work experience) completed an X-ray image inspection task. The task consisted of 128 X-ray images of real passenger bags, half of which contained a prohibited item: 16 images contained a gun, 16 a knife, 16 an IED,

and 16 explosive material. For each image, participants had to press a key to decide whether the bag was harmless or not, and they then had to assign a confidence rating on a 5-point scale ranging from 1 (*very unconfident*) to 5 (*very confident*).

Results Experiment 2. The averaged *z*ROC curves displayed in Figure 1 seem to better fit the *z*ROC curve predicted by d_a than those predicted by *d′* or *A′*. Individual slope parameters (i.e. angles of incline) estimated using the LABROC3 algorithm (Metz, Herman, & Shen, 1998) show a mean of 0.54 (95%-BCa-CI [0.50, 0.60]) and median of 0.50 (95%-BCa-CI [0.46, 0.55]).

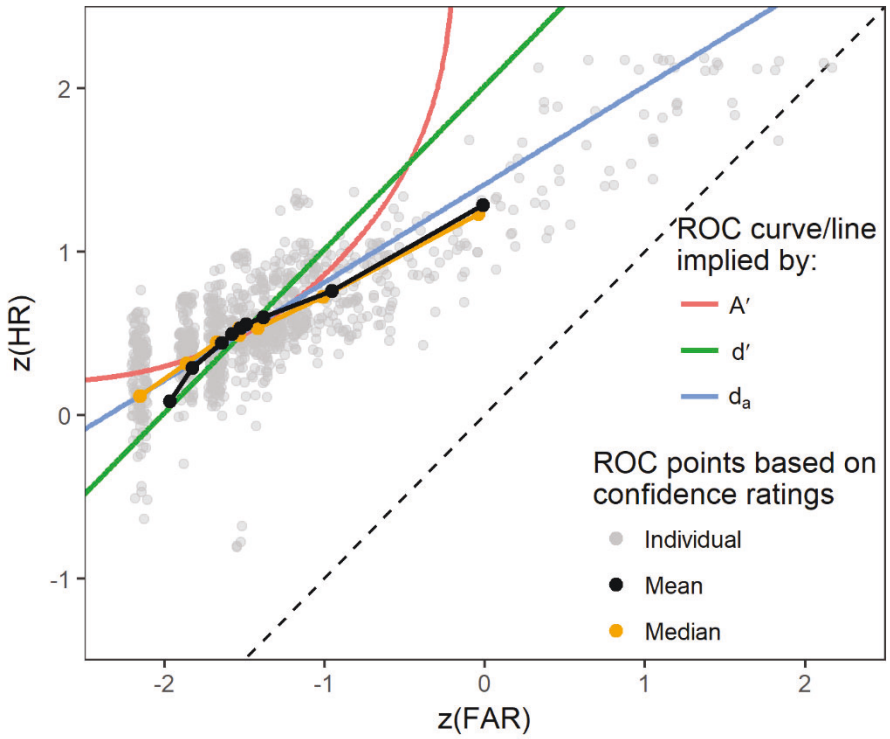

Figure 1. Individual (grey; jittered) and pooled (black) empirical *z*ROC curves, the lines corresponding to the mean *A′*, *d′*, and d_a with a slope of 0.6, and the chance line (dashed).

General conclusion. To investigate the validity of two detection measures commonly used in visual search and decision tasks such as X-ray image inspection, we conducted two studies

with different methodological approaches. Experiment 1 manipulated the criterion by direct instruction, whereas Experiment 2 used confidence ratings to generate multiple ROC points. For both studies, d' and A' were found to be invalid detection measures: they would have wrongly indicated lower sensitivity for a more liberal decision criterion.

In our experiments, the slope parameter was around 0.5–0.6, which corresponds well to the findings in other experiments that investigated the X-ray baggage inspection task (Godwin, Menneer, Cave, & Donnelly, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010). However, one should be cautious to always adopt d_a with a slope of 0.5–0.6 for any X-ray baggage inspection or other visual search task. For low sensitivity, a non-unit slope *z*ROC implies that the ROC curve falls meaningfully below chance performance (Macmillan & Creelman, 2005, p. 68). Therefore, the slope parameter is likely higher for low levels of sensitivity.

To better understand what factors influence the slope parameter, a better understanding of the inspection process is needed. From the perspective of Gaussian SDT, a *z*ROC slope smaller than one implies that the signal-plus-noise distribution has a higher standard deviation than the noise distribution. A possible explanation for this is that prohibited items can vary strongly in how well they can be recognized, for example, depending on item category (Halbherr et al., 2013; Koller et al., 2009) and the exemplar within categories (Bolfing, Halbherr, & Schwaninger, 2008; Schwaninger et al., 2007). The SDT framework might have to be extended to provide a better model of the visual inspection process, e.g. by assuming a sequence of decisions for single items within the image with both a decision and a quitting threshold (Koller et al., 2009; Wales, Anderson, et al., 2009; Wolfe & Van Wert, 2010).

In conclusion, X-ray image inspection research and related domains will have to be cautious when using one-point estimates of sensitivity such as d' and A'. We recommend always starting by performing an analysis and discussion of the directly accessible HR and FAR. For the use of detection measures, it should be considered that the *z*ROC slope can be expected to lie somewhere between 0.5 and 1 for X-ray baggage inspection tasks. With d_a, effects on sensitivity can

be estimated for these two slopes separately to test the two limits of the assumption of constant sensitivity (where the upper limit with a *z*ROC slope of one corresponds to *d'*).

Manuscript 3: Relevance of Visual Inspection Strategy and Knowledge about Everyday Objects for X-Ray Baggage Screening

Motivation and aim of the study. In order to effectively inspect X-ray images for prohibited items, security personnel need to know which items are prohibited and what they look like in X-ray images (Halbherr et al., 2013; Koller et al., 2009, 2008). Knowing what everyday objects look like in X-ray images could further facilitate the differentiation between harmless and prohibited items. (Hattenschwiler, Michel, Kuhn, Ritzmann, & Schwaninger, 2015) revealed a negative correlation between everyday object knowledge measured in an X-ray object categorization and naming test and false alarm rate in an X-ray baggage screening task. An intuitive explanation of this result could be that once an item is identified as harmless, it can no longer be mistaken for a threat item and thereby not result in a false alarm. This assumption implies that screeners search an X-ray image and decide for one object after another whether it is harmless or not (for a description of such a model see e.g. Spitz & Drury, 1978; Wales, Anderson, et al., 2009; Wolfe & Van Wert, 2010).

Knowledge about everyday objects could be especially relevant for the detection of prohibited items that the screeners have never seen before. Since they lack the knowledge about their appearance (knowledge based factors), such novel prohibited items are harder to detect, when they less resemble known prohibited items. It is possible that screeners with good knowledge about everyday objects can detect novel prohibited items by an exclusion principle: They could only declare a bag as harmless if all contained objects are identified as harmless everyday objects, which in terms of SDT means to shift the response tendency, i.e. to apply a liberal decision criterion. If screeners can successfully be instructed to apply such a liberal decision criterion, this could allow for interesting practical applications, e.g. for increased effectiveness when screening bags of high-risk passengers. Our study therefore investigated whether such a search strategy can be instructed and/or trained and how performance relates to the knowledge of everyday objects.

Method. The experiment used a mixed factorial design with two differently instructed inspection strategies (normal decision vs. liberal decision) as within-subjects factor and training of a new inspection instruction (short e-learning module vs. instruction only) as between-subjects factor. 31 professional screeners (64.5% female; between 26 and 61 years old, *M* = 45.4, *SD* = 8.9; between 2 and 26 years of work experience, *M* = 8.4, *SD* = 5.5) were each tested twice. At the first test date, all screeners completed three pre-tests: a test on the ability to recognize everyday objects, the X-Ray CAT (Koller & Schwaninger, 2006), and X-Ray ORT (Hardmeier, Hofer, & Schwaninger, 2006). The test on the ability to recognize everyday objects consisted of 32 X-Ray images of cabin baggage. In each image, three objects per bag were marked with a red frame and had to be named by typing their name into a textbox. At the second test date, participants were divided into two groups that were balance with regard to their performance in the pre-tests and work experience. One group first completed an e-learning module of about 10 min that consisted of a short definition of the new inspection strategy *liberal decision* followed by some examples with feedback. The other group only received a short instruction of the new inspection strategy. All participants then completed an X-ray image inspection task that consisted of 128 X-ray images of passenger bags with half the images containing one prohibited item of the categories guns, knives, IEDs and other prohibited items (16 images each). For each category, half of the prohibited items were known from training, the other half was novel. For one half of the test, participants were instructed to visually inspect the X-ray images like they were used to from their job (*normal decision*). For the other half, screeners were instructed to visually analyze each object in the X-ray image and only decide that the bag was harmless if each object in the image could be recognized as harmless (*liberal decision*). For each image, screeners had to decide whether it was harmless or not by pressing a key, followed by confidence ratings on a scale from 0 to 10. During this test, eye tracking was conducted using the SMI RED-m eye tracker with a gaze sample rate of 120 Hz, a gaze position accuracy of 0.5° and a spatial resolution of 0.1°.

Results. The hit rate (see Figure 2) was higher for known than for novel prohibited items, $W = 1934$, $p < .001$. In comparison to *normal decision*, *liberal decision* resulted in a higher hit rate for known prohibited items, $W = 85$, $p = .02$, and for novel prohibited items, $W = 95.5$, $p = .02$. In addition, the false alarm rate was significantly higher, $W = 60.5$, $p < .001$. Sensitivity (A_g, calculated from confidence ratings) did not differ between the two inspection strategies, neither for known, $W = 240$, $p = .88$ (normal decision: $M = .889$, $SD = .066$; liberal decision: $M = .890$, $SD = .068$) nor for novel items, $W = 262.5$, $p = .78$ (normal decision: $M = .794$, $SD = .079$; liberal decision: $M = .789$, $SD = .067$). In contradiction to our expectation, these effects were not significantly larger for the group who received the e-learning, neither for the hit rate of known items, $U = 145$, $p = .16$, novel items, $U = 141.5$, $p = .20$, the false alarm rate, $U = 141$, $p = .21$.

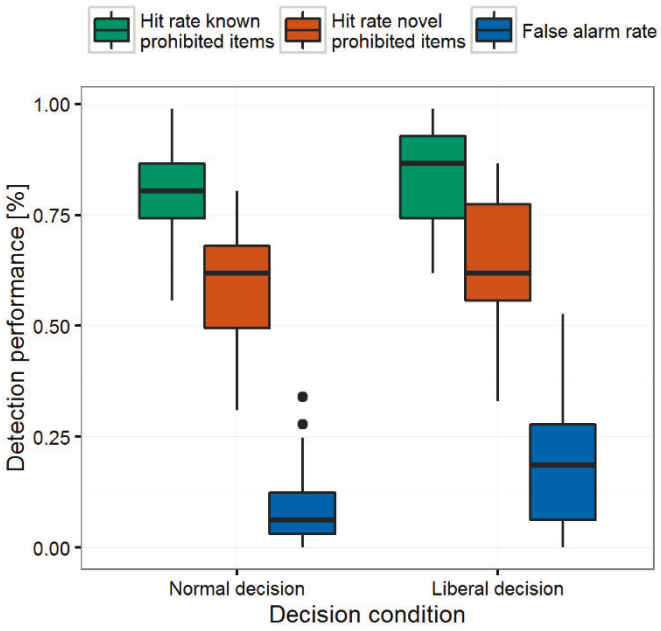

Figure 2. Box plots of hit and false alarm rates depending on decision condition and prohibited item class (known vs. novel). (*Note: Performance values are multiplied by an arbitrary constant for security purposes.*)

The eye tracking data showed that the overall increased response times in the liberal decision condition were associated with on average (mean) 22% longer scan paths (measured in

pixels) for target present trials, $W = 75$, $p = .009$, and 28% longer scan paths for target absent trials, $W = 49$, $p < .001$. However, this increase was disproportionate for target absent trials, leading to a shorter average scan path per response time, $W = 292$, $p = .002$, but not significantly so for target present trials, $W = 228$, $p = .23$. Also the total number of fixations increased, $W = 271$, $p = .014$, but also the average duration of these fixations increased, $W = 38$, $p < .01$.

When instructed for *normal decision,* screeners with a high ability to recognize everyday objects also detected more known prohibited items, had a marginally significant lower false alarm rate, but did not detect more novel items (see Table 3). Looking at the condition *liberal decision*, the pattern changes: the ability to recognize everyday objects was not associated with lower false alarm rates but with higher hit rates for novel prohibited items.

Table 3

Correlations between Everyday Object Test Score and Variables of the X-ray Inspection Task

Decision condition	Variable of X-ray image inspection task		
	HR[a] known prohibited items	HR[a] novel prohibited items	False alarm rate
Normal decision	$r_s = .430$ $p = .008$	$r_s = -.117$ $p = .735$	$r_s = -.298$ $p = .052$
Liberal decision	$r_s = .391$ $p = .015$	$r_s = .322$ $p = .038$	$r_s = -.018$ $p = .462$

[a]Hit rate

Conclusion. The experiment showed that instructing professional screeners to apply a more liberal decision led to increased hit and false alarm rates in an X-ray image inspection task. Sensitivity remained constant, implying that the observed change in hit and false alarm rates was due to a change in the decision criterion. We therefore conclude that screeners are generally capable to shift their criterion based on an instruction. Because this criterion shift leads to higher false alarm rates and response times, the application of the liberal decision strategy would decrease the efficiency of X-ray screening at security checkpoints. However, it could be useful for a targeted increase in hit rate, e.g. for high-risk flights.

In our experiment, we also investigated whether an e-learning module could assist with the application of the new inspection strategy, but found no improvement. For the liberal decision strategy, there was a correlation between the ability to recognize everyday objects and the hit rate for novel threat items. It should therefore be investigated whether specific training of everyday object recognition would assist with the criterion shift.

The eye tracking data shows that for images of harmless bags screeners have longer scan paths and more fixations in the liberal decision condition. Nevertheless, at the same time scanning was slower and fixations longer. This suggests that applying the liberal decision not only extended the search duration but also affected underlying cognitive processes, e.g. (Meghanathan, van Leeuwen, & Nikolaev, 2014).

Manuscript 4: Why stop after 20 minutes? Breaks and target prevalence in a one hour X-ray image inspection task

Aim and motivation: Current EU regulations restrict the duration of X-ray image inspection of passenger baggage at airport security checkpoints to 20 min as a precautionary measure to prevent performance decrements (Commission Implementing Regulation [EU], 2015/1998). However, this rule is not founded on solid empirical research. At the same time, the restriction of screening to 20 min limits the options of work organization at airport security checkpoints considerably. This is especially the case for *remote cabin baggage screening* (RCBS). With RCBS, security personnel visually inspect X-ray images in an office-like environment separate from the checkpoint and therefore need more time and coordination to rotate between positions compared to the conventional X-ray image inspection at the checkpoint (Kuhn, 2017). Manuscript 4 therefore investigated how performance changes over time (i.e., as a function of time on task).

Empirical research into how performance in X-ray image inspection develops over time is scarce. (Meuter & Lacherez, 2016) analyzed 4 months of threat image projection (TIP) data from an Australian airport. TIP is a technology that projects prerecorded threat items onto real X-ray images of passenger baggage during baggage screening at airport security checkpoints (Cutler & Paddock, 2009; Hofer & Schwaninger, 2005). The TIP hit rate (or percent detected) refers to the proportion of projected fictional threat items that screeners have detected. Meuter and Lacharez (2016) found a small decrease of approximately 2% in the hit rate with time on task, only when workload was high (operationalized as more than 5.4 images screened per min during one session of continuous screening). However, the study says little about the development of performance beyond 20 min and could not clear whether the decrease in hit rate was due to a criterion shift or a lower sensitivity. In a study by Ghylin, Drury, Batta, and Lin (2007), professional screeners completed an X-ray image inspection task over the course of four hours, showing a decline in hit rate and false alarm rate but not in the sensitivity measure A', indicating a criterion shift. The study does not provide any conclusions about the development of performance within the first hours.

In vigilance tasks, depending on task difficulty or task demands, performance decrements can already occur after 5 min (Arrabito et al., 2015). Most studies have revealed decreases in vigilance within the first 15–30 min of the task (Mackworth, 1948; Nuechterlein et al., 1983). However, besides the similarities, there are also considerable differences between X-ray image inspection and typical vigilance tasks (Wolfe et al., 2007).

Because time on task is likely to affect the decision criterion (Ghylin et al., 2007), it is important to use a valid sensitivity measure that is independent of the criterion. Based on findings regarding the target prevalence effect (Godwin et al., 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010) and Manuscript 2, we assumed d_a with a slope parameter of around 0.6 to be appropriate. In order to validate the appropriate detection measure, we also investigated the target prevalence effect in our experiment.

In the current study, we investigated the effect of time on task on screener performance when X-ray images were analyzed for 60 min. One group screened for 60 min continuously, whereas the other group took 10-min breaks between 20-min screening blocks. Because previous research found increased stress and less engagement after vigilance tasks (Helton, 2004; Matthews et al., 2002), we also monitored the perceived stress of the task by asking screeners to complete the Short Stress State Questionnaire (SSSQ; Helton, 2004).

Method. 75 professional screeners (46.47% female, between 20 and 67 years old, M = 32.01, SD = 12.78, 0.3 to 28 years of working experience, M = 2.26, SD = 3.13[2]) participated during their regular working hours. A 2 (break condition: with vs. without breaks) × 2 (prevalence condition: high vs. low prevalence) × 3 (time on task: 0–20 min, 20–40 min, 40–60 min) mixed factorial design was employed. The two break conditions *with breaks* and *without breaks* served as between-subject variable. All screeners completed the test twice, once in the low prevalence condition and once in the high prevalence condition.

[2] Two participants did not report their demographics

The test consisted of 864 X-ray images of passenger cabin (carry-on) baggage. In the high prevalence condition, one in eight images contained a threat item; in the low target prevalence condition, one in two images contained a threat item. Threat items belonged, in equal numbers, to the categories guns, knives, and improvised explosive devices (IEDs). To ensure that each participant's performance was assessed on the basis of the same images, the test was divided into 12 blocks of 72 images and after 5 min, the test jumped to the next block. The order of these blocks was counterbalanced. During the test, 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 image inspection task, screeners filled out the SSSQ and provided information on their shift schedule, work experience, age, and gender.

Results. For the analysis, 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. The high prevalence condition had a higher hit rate, $F(1, 69) = 37.92$, $p < .001$, $\eta_p^2 = .36$, and a lower false alarm rate, $F(1, 69) = 118.53$, $p < .001$, $\eta_p^2 = .63$ (see Figure 3). The individual differences in hit and false alarm rate between the two prevalence conditions was used to estimate the slope parameter. The resulting 0.65 (95% BCa-CI [0.41, 0.89]) was lower than the slope of 1 assumed by d', suggesting to use d_a (with the estimated slope of 0.65) as a sensitivity measure instead, in line with the argumentation of Wolfe et al. (2007). The results showed that the effect of time on task depended on the prevalence condition for the false alarm rate, $F(1.97, 136.18) = 17.9$, $p < .001$, $\eta_p^2 = .21$, (interaction of Prevalence × Time on task) and for the criterion c_a, $F(1.95, 134.28) = 11.82$, $p < .001$, $\eta_p^2 = .15$, but not quite significantly so for the hit rate, $F(1.96, 134.94) = 3.06$, $p = .051$, $\eta_p^2 = .04$, and not at all for d_a, $F(1.95, 134.72) = 0.11$, $p = .895$, $\eta_p^2 = .00$ (see Figure 4). More specifically, the criterion decreased from the first 20-min block (0–20 min) to the second 20-min block (20-40 min) for high prevalence, whereas the criterion increased for low prevalence. For d_a there was a small main effect of time on task, $F(1.97, 135.91) = 3.43$, $p = .036$, $\eta_p^2 = .05$, with

post-hoc tests revealing a small increase from the first 20-min block to the second 20-min block. There was no effect of breaks, neither for hit rate, false alarm rate, d_a, nor c_a.

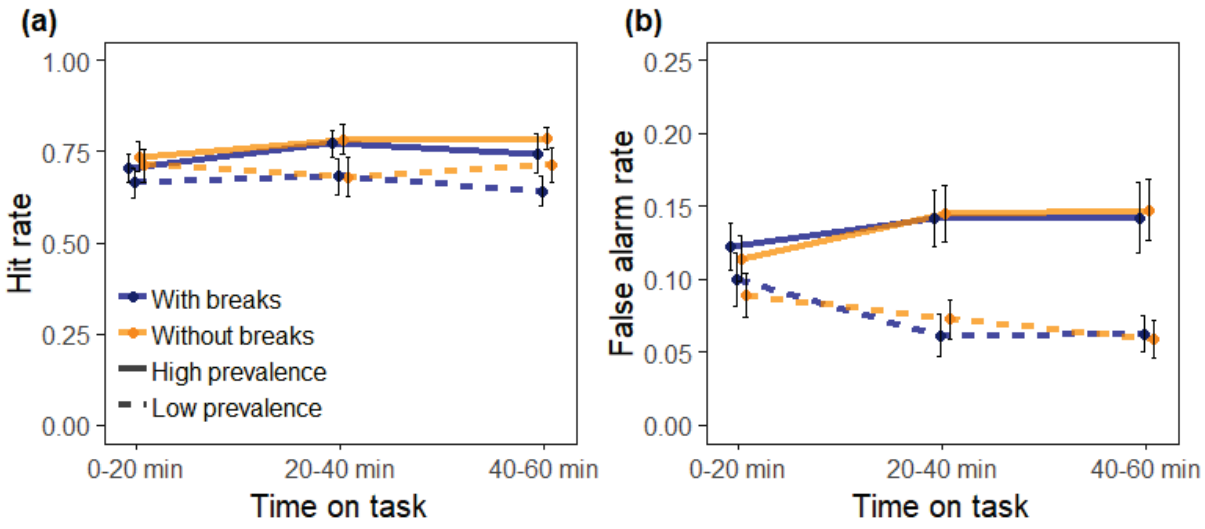

Figure 3. Hit rate (a) and false alarm rate (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task. Error bars represent standard errors.

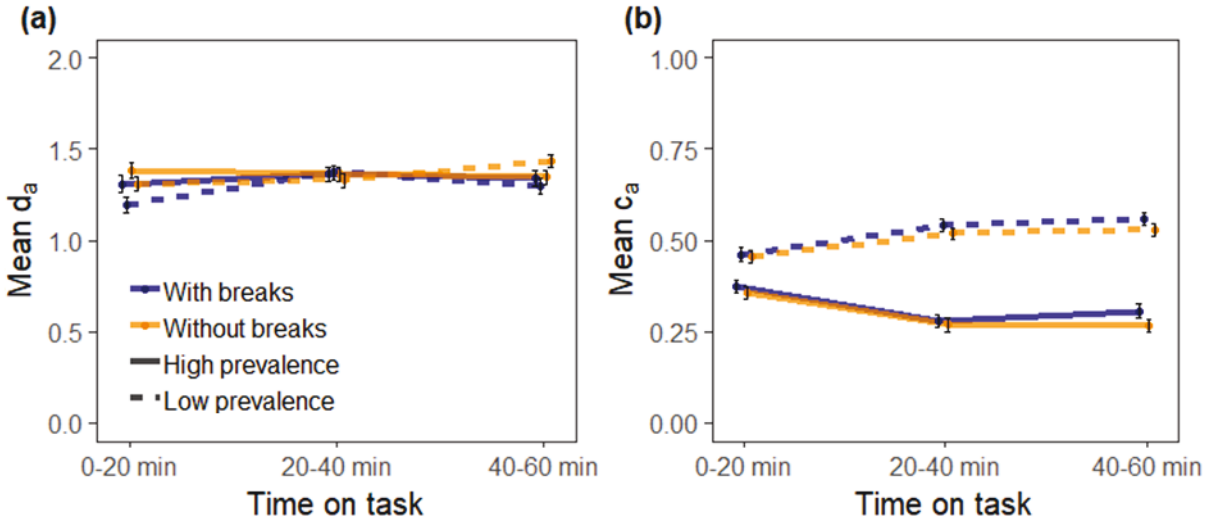

Figure 4. Sensitivity measure d_a (a) and criterion c_a (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task. Error bars represent standard errors.

For the subjective stress levels, we calculated 2 (with vs. without breaks) × 2 (high vs. low prevalence) ANOVAs with the three levels of stress *distress, worry*, and *engagement* as dependent variables. For *distress*, the ANOVA revealed a significant main effect of break, $F(1, 66) = 9.17$, $p = .004$, $\eta_p^2 = .12$. Because the data do not meet the assumptions of normal distribution or homoscedasticity, a Wilcoxon rank sum test was carried out, which also revealed a significant difference between the two break conditions ($W = 1616$, $p = .003$).

Conclusion. To examine time on task and the influence of breaks on screener performance, two groups of X-ray screeners performed an X-ray image inspection task for 60 min. Whereas one group took breaks in line with the 20-min rule in EU regulations, the other group worked for 60 min without breaks. Target prevalence was varied to determine the valid detection measure for this task.

In line with Wolfe et al. (2007), we would argue that it is implausible for screeners to become faster and better at detection when they expect fewer targets. It is more plausible that the equal variance assumption of d' 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. 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 Manuscript 2 and other previous studies have found for the task of X-ray image inspection (Godwin et al., 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010).

We found that a lower target prevalence caused a shift in response tendency resulting in a lower hit rate and false alarm rate. Previous studies have found that the target prevalence effect depends on implicit learning rather than on explicit instruction, and that it therefore takes some time until searchers adapt to the prevailing target prevalence by shifting their criterion accordingly (Ishibashi et al., 2012; Lau & Huang, 2010). Lau und Huang (2010) have found that the instructed target prevalence is not sufficient to create the target prevalence effect. In our study as well, although participants were instructed about the target prevalence, the target prevalence

effect evolved over time, supporting the notion that implicit learning plays an important part in evoking the target prevalence effect. In the high target prevalence condition, participants first shifted their criterion to a more liberal location, meaning that they increased their tendency to declare that an X-ray image contained a prohibited item. In the low target prevalence condition, they shifted their criterion to a more conservative location and thereby increased the ratio of images declared to be harmless. In both conditions, the criterion stayed stable after the initial 20 min. The effect of time on task on sensitivity did not depend on target prevalence. d_a increased in the beginning of the test and then remained constant. It is possible that there is a warm-up phase in X-ray image inspection, during which the cognitive processes necessary for this task are fully activated, as can be observed in other recognition tasks (e.g. Allport & Wylie, 1999; Monsell, 2003). It is however also possible that the observed ramp-up in performance was an accustomization to the specifics of the task employed in our experiment.

Whereas breaks have often had a positive effect on performance in previous studies (Arrabito et al., 2015; Balci & Aghazadeh, 2003; Colquhoun, 1959; Kopardekar & Mital, 1994; Steinborn & Huestegge, 2016), breaks are mainly thought to offer rest, recuperation, and prevention of fatigue (Tucker, 2003). Considering that participants who performed 60 min of continuous screening did not show a decrease in performance, there was no room for recuperation during breaks. Even though there was no effect of breaks on detection performance, there seems to have been an effect on well-being in terms of stress. 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, increased distress could have an effect on performance. It has, however, to be noted that there was considerable variance between screeners in the condition without breaks. Whereas the longer screening without breaks caused distress in some participants, it did not in others.

Whereas our study has shown that screeners can maintain detection performance over 60 min without breaks, it is still too early to conclude that the rule of a maximum of 20 min of

screening should be lifted. Our results show that 60 min of continuous screening caused distress for some screeners. Considering that participants only did 60 min of screening twice with 3 to 5 weeks in between, it is unclear how prolonged screening would affect performance and distress if repeated multiple times a day and over months. Further limitations to ecological validity are that poor performance did not have any consequences in our experiment whereas a miss can be disastrous in practice. This might make prolonged screening time more stressful in practice. Further, target prevalence is lower in practice and this might make it more difficult to sustain attention and performance. These limitations should be addressed by follow-up field studies. By showing to regulators that more than 20 min of screening can be possible without negatively affecting performance, our study paves the way for these field studies.

Manuscript 5: Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection

Motivation and aim of the study. To assist with the detection of improvised explosive devices (IEDs), explosives detection systems for cabin baggage screening (EDSCB) have been developed (Singh & Singh, 2003). EDSCB use high and low energy X-ray from different angles to estimate effective atomic number and density information at each location within bag (Singh & Singh, 2003). The EDSCB triggers an alarm if this information is similar to the effective atomic number and density of explosives. With the foreseeable spread of EDSCB in European countries, regulators and airport operators are currently discussing two implementation scenarios differing in their level of automation and human–machine function allocation: *on-screen alarm resolution* (OSAR) and *automated decision* (Sterchi and Schwaninger, 2015). EDSCB might be especially important for the detection of bare explosives, which lack a triggering device and can therefore be difficult to detect (Jones, 2003), but still pose a threat in cabin baggage screening.

In the OSAR scenario, automation is implemented as a diagnostic aid: EDSCB indicates potential explosive material by either marking an area on the X-ray image of a passenger bag with a colored rectangle or highlighting it in a special color (Nabiev and Palkina, 2017). EDSCB systems with high hit rates (close to 90%) have false alarm rates in the range of 15–20% (personal communication with EDSCB experts, summer 2016). Also in other domains, designers of automation aids often set low thresholds, because the consequences of automation misses are considered to be more costly than false alarms (Parasuraman and Wickens, 2008). However, if the base rate of dangerous events to be detected is low, the result will be many false alarms and only few hits (Parasuraman and Riley, 1997). This can produce a ‘cry wolf’ effect with operators ignoring system warnings (Breznitz, 1983; Bliss, 2003). Such an effect can drastically reduce or even eliminate the benefits of automation when it is implemented as a diagnostic aid.

The automated decision scenario uses a higher level of automation with a different human–machine function allocation. Bags on which the EDSCB raises an alarm are sent *automatically*

to secondary inspection using manual search and/or explosive trace detection (Sterchi and Schwaninger, 2015). Because secondary inspection is time-consuming, false alarm rates of 15–20% from EDSCB are not operationally acceptable in this scenario. Instead, the threshold of the EDSCB would have to be lower, resulting in a lower false alarm rate but also a lower hit rate. To compare the two scenarios that require different thresholds, the reliability of the EDSCB can be defined in terms of signal detection theory (Wickens and Dixon, 2007; Parasuraman and Wickens, 2008; Rice and McCarley, 2011).

The present study examined the benefits of automated explosive detection systems for cabin baggage screening (EDSCB) in two realistic implementation scenarios differing in the level of automation and human–machine function allocation (EDSCB with OSAR vs automated decision). It addressed the following three research questions: 1) Does EDSCB lead to higher human–machine system performance for detecting IEDs and explosives? 2) Does this depend on the level of automation (OSAR vs automated decision)? 3) Is this dependent on screener work experience? To address these research questions, two experiments using a simulated baggage screening were conducted at different European airports with screeners differing in work experience.

Method Experiment 1. 61 professional screeners with at least two years of work experience (M = 7.68, SD = 4.85) participated the experiment. These participants were not familiar with automation aids for cabin baggage screening, were in average 42.5 years (SD = 10.52, range 24–60 years) of age, and 57.37% of them were female. The participants were assigned to one of three conditions that were balanced with regard to age, work experience, and performance in an X-ray image competency test (Koller and Schwaninger, 2006): baseline, OSAR, and automated decision. The participants had to solve a test that consisted of 640 X-ray images, of which 80 contained a threat (target prevalence of 12.5%) from one of the following categories: IEDs, explosive materials, guns, gun parts, and knives (16 images for each threat category). In the baseline condition, screeners inspected each image on a laptop and reported whether the depicted

bag was harmless or not by clicking on a button on the screen. In the OSAR condition, screeners were instructed that they were supported by an EDSCB marking most of the IEDs and explosives and that this EDSCB can cause false alarms. In the OSAR condition, 14 of the 16 IEDs and 14 of the 16 explosives were marked with a read frame; and 94 of the 560 images without a threat displayed red frames as false alarms. This corresponds to an EDSCB with a reliability of $d' = 2.1$, a hit rate of 88%, and a false alarm rate of 17%. In the automated decision scenario, screeners were instructed that they were assisted by an EDSCB and if the EDSCB detected an IED or explosives, the X-ray image would not be displayed for analysis. They were further instructed that the EDSCB can miss some IEDs or explosives. In the automated decision condition, 10 of the 16 IEDs and 10 of the 16 explosives were predefined as detected by the EDSCB and not displayed to the participants for inspection. In addition, 20 images without a threat were not displayed for inspection, because they were predefined as triggering a false alarm of the EDSCB. This corresponds to an EDSCB with a reliability of $d' = 2.1$, a hit rate of 63%, and a false alarm rate of 4%. Participants needed up to 2 hr to complete the test, which included three breaks of 10 min.

Results Experiment 1. The results were analyzed with a series of ANOVAs and, when indicated, post-hoc tests with Holm-Bonferroni corrections were calculated (Holm, 1979). The results (see Figure 5) revealed that the EDSCB had no impact on the hit rate for guns, gun parts, or knives. The hit rate of the human–machine system for IEDs was higher in the automated decision condition than in the OSAR condition, whereas there was no significant difference between the baseline condition and any of the two condition with EDSCB. The hit rate for explosives was significantly higher in the automated decision condition compared to each of the other two conditions. However, the automated decision condition also had a higher false alarm rate compared to each of the other two conditions. When comparing the baseline and automated decision conditions only considering the human performance without hits and false alarms by the EDSCB, no significant difference was found.

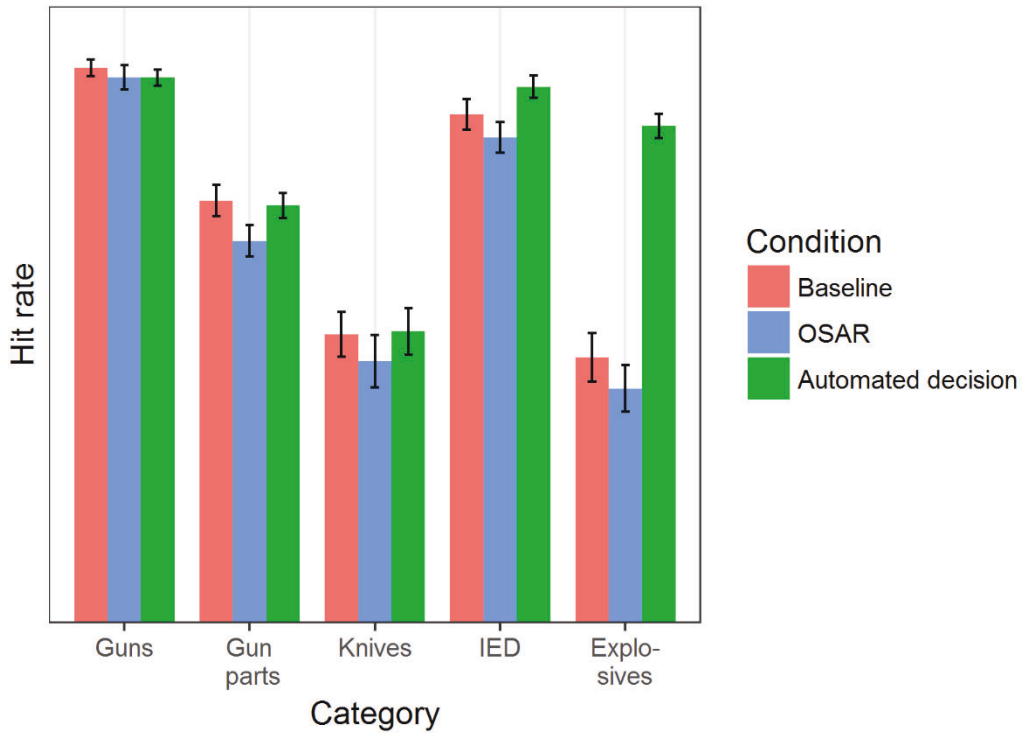

Figure 5. Mean human–machine hit rates by condition (baseline, OSAR, automated decision) and prohibited item categories (guns, gun parts, knives, IEDs, and explosives). Absolute values of hit rate are not shown due to security restrictions in this project. Error bars are ± one standard error.

Conclusion Experiment 1. The first experiment showed that screeners can detect IEDs fairly well even without EDSCB whereas they have difficulty detecting explosives on their own. However, EDSCB increased the number of detected IEDs and explosives only in the automated decision condition but not in the OSAR condition. A limitation of the first study was that participants had no prior experience in X-ray image inspection with decision aids, which might have affected how they interacted with it (Parasuraman and Manzey, 2010; Sauer et al., 2016; Strauch, 2016). Another limitation was that only screeners with at least two years of work experience were tested. Screeners with less work experience and training might benefit when it comes to detecting IEDs and explosives in the OSAR condition due to their lower baseline performance (Halbherr et al., 2013).

Methods Experiment 2. For Experiment 2, 77 professional screeners who were familiar with automation aids completed the same test already conducted in Experiment 1. Group 1 (44 screeners, 14 females) was as well-trained and experienced as the screeners in Experiment 1 (years of work experience: *M* = 8.45 years, *SD* = 5.66). Their average age was 36.55 years (*SD* = 8.46, range 21–53 years). Group 2 (33 screeners, 19 females) had less work experience and training (less than one year). Their average age was 30.81 years (*SD* = 10.93, range 18–53 years)[3].

Results Experiment 2. Again, EDSCB did not affect the hit rate for guns, gun parts, or knives (see Figure 6). For IEDs, the effect of the experimental condition on the hit rate differed between the inexperienced and experienced screeners: the inexperienced screeners achieved a higher hit rate in each of the conditions with EDSCB compared to the baseline condition, whereas experienced screeners did not differ significantly between the conditions. For explosives, inexperienced and experienced screeners both achieved a higher hit rate in the automated decision condition compared to the baseline and the OSAR conditions. For the false alarm rate, the analyses revealed a significant main effect of condition, but none of the post-hoc comparisons attained significance. A follow-up analysis on the benefit of OSAR for inexperienced screeners revealed that the OSAR condition had a higher sensitivity *d'* and a more conservative decision criterion *c* compared to the baseline condition.

[3] Two screeners did not report their age.

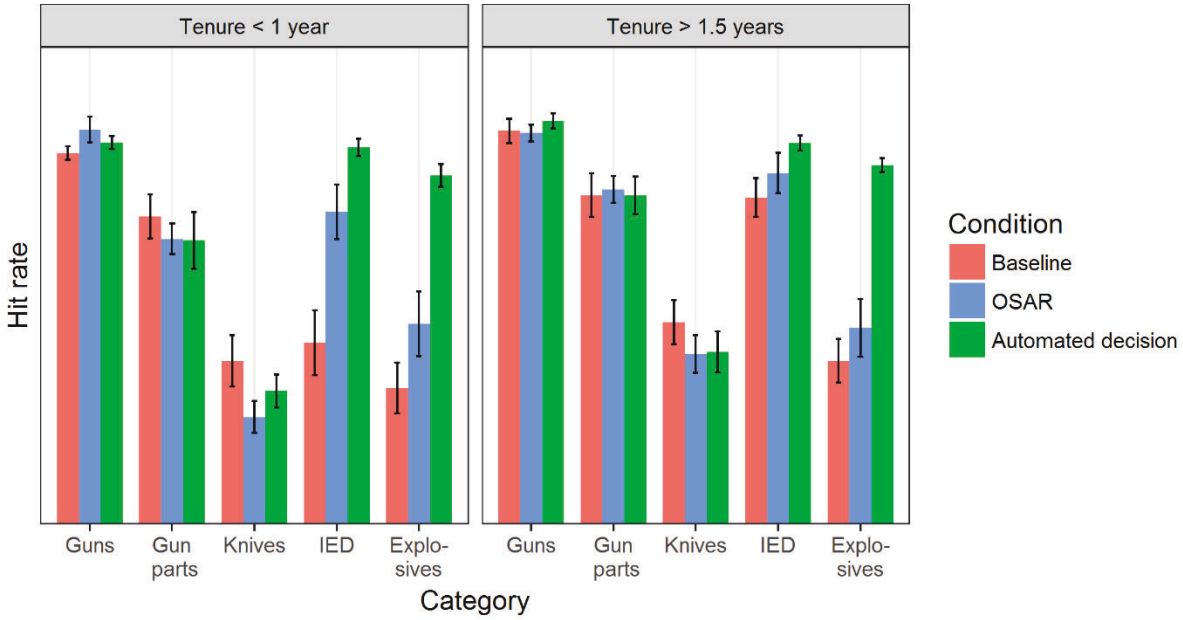

Figure 6. Mean human–machine hit rates by condition (baseline, OSAR, automated decision), threat categories (gun, gun parts, and knives), and variation of work experience (tenure < 1 year and tenure > 1.5 years). Absolute hit rate values are not shown due to security restrictions in this project. Error bars are ± one standard error.

General conclusion. This study examined the use of automation for the airport security screening of cabin baggage by testing two levels of automation that are currently being discussed by regulators and airport operators: on-screen alarm resolution (OSAR) and automated decision (Sterchi and Schwaninger, 2015). Two experiments were conducted with screeners working at two European airports and verying in their work experience.

The EDSCB in the OSAR condition provided a hit rate of 88%. For explosives, using EDSCB as a diagnostic aid lost its potential benefit, not increasing hit rate compared to no EDSCB. The OSAR scenario was beneficial for the detection of IEDs, but only for the less experienced screeners. We argue in line with (Cullen et al., 2013) that the automation system with OSAR assisted in the search component of X-ray image inspection by guiding attention to the relevant area – the first processing stage of sensory processing in the taxonomy proposed by Parasura-

man et al. (2000). OSAR did not improve the hit rate of experienced screeners for IEDs, possibly because they already achieved high hit rates for IEDs in the baseline condition without automation and therefore there was not much room for improvement through OSAR. In addition, experienced screeners may also have judged their own ability to detect prohibited items to be superior to the automation support – a reason for noncompliance also reported in other domains (e.g. Lee and Moray, 1992, 1994). Moreover, the low target prevalence in our study and, therefore, the low base rate led to many false alarms. This probably led to a 'cry wolf' effect with experienced screeners, meaning that they might simply have ignored the system warnings (Breznitz, 1983; Bliss, 2003). Future research could also explore whether specific training and familiarity with the automation aid (Sauer et al., 2016) might provide screeners with a mental model of its capabilities. Such mental models could be important for an effective use of an automation aid (Strauch, 2016).

In the automated decision condition, EDSCB did not affect human performance. Hence, the observed increase in detection performance was determined by the amount of explosives missed by screeners but detected by the EDSCB. Automated decision therefore increased the human-machine hit rate for explosives and for inexperienced screeners also for IEDs, but also increased the false alarm rate.

Manuscript 6: A first simulation on optimizing EDS for cabin baggage screening regarding throughput

Motivation and aim of the study. The introduction of EDS into cabin baggage screening is certainly an advantage security wise. But how does EDS affect throughput, i.e. the amount of items that can be screened within a certain time? Butler and Poole (2002) argued that EDS can reduce throughput, but since then EDS machines have become faster and more reliable. It would therefore be interesting to examine effects of EDS on throughput taking into account up-to-date information on technology, humans and processes. In this study this was explored for one specific process using discrete event simulation. In addition, two measures to cope with potential negative effects on throughput were tested for their effectiveness: The first measure is to assign a second security officer to the task of resolving alarms using manual search and/or ETD. This should double the rate at which alarms can be resolved (assuming there is sufficient room and equipment provided). The second evaluated measure was to instruct the X-ray screener to resolve one of the alarms when the backlog of bags queueing for secondary search causes the screening to stop.

Process description and assumptions. Once the EDS generates an alarm, there are at least two different approaches to resolve these alarms. One is on-screen alarm resolution: When the screener reviews the X-ray image of the bag that triggered the alarm by the EDS, a frame is displayed around the area of the X-ray image which might contain explosives. The screener then decides whether the bag needs further alarm resolution. Another approach is to increase the level of automation (for an overview of levels of automation see (Sheridan & Verplank, 1978)) and *automatically* redirect items that caused an alarm for alarm resolution by explosives trace detection (ETD) and/or manual search. In this study, we evaluated this second approach, which will be referred to as "automatic decision scenario".

In the evaluated automatic decision scenario, false alarms by the EDS are resolved with an ETD conducted by the same security officer that also resolves alarms by the X-ray screener.

The first and quite obviously relevant aspect of this process is the time needed for using an ETD to resolve the alarm by the EDS. It should be noted that modern ETD technology is fast, for example ("IONSCAN 500DT," n.d.) report 5-8 s for their IONSCAN 500DT ETD machine. However, the overall time needed for alarm resolution using ETD depends strongly on where and how many trace samples are taken (Butler & Poole, 2002).

Method and procedure. The simulation was implemented in FlexSim (see www.flexsim.com for more information), an off the shelf 3D modelling and discrete event simulation software. The basic layout, processes, and parameters of the model were set in accordance with a specific checkpoint design of a European airport. To take into account that different durations for alarm resolution with ETD are possible depending on swab sampling (Butler & Poole, 2002), three different scenarios were tested in the current study in the range reported by Butler and Poole (2002): One with a low average duration of 30 s, a second taking 60 s, and a third taking 120 s on average. Screening performance and duration was estimated based on the empirical data from Manuscript 5. For the response time it was thereby differentiated between hits, false alarms, and correct rejections and misses (see Table 4). For the false alarm rate, the three samples from Manuscript 5 were used as different simulation scenarios (see Table 5) to investigate the effect of the screener population. For the EDSCB, false alarm rates ranging from 1 to 15% were explored.

Separate simulations were run for each combination of the three reference groups, 15 false alarm rate levels of the EDS (1-15%) plus one level without EDS, and the three durations for alarm resolution using ETD (30 s, 60 s, 100 s). To test the effectiveness of the two measures described in the previous section, they were also run for each of the 15 false alarm rate levels of the EDS plus one level without EDS. To keep the results manageable, the measures were only tested in combination with the first reference group (which was recruited from screeners working at the checkpoint that served as reference for the model and therefore seemed most adequate) and only for the medium duration of alarm resolution using ETD (60 s). For each of the resulting 192 conditions, one hour of the baggage screening process was simulated 200 times.

Table 4

Model Parameters

Parameter	Distribution	Mean (SD)
Placing item on conveyor	Gamma	5 s (5 s)
Items per passenger	Poisson (translated)	3 (√2)
Evaluation time X-ray screener	Empirical	CR [a]/Miss: 3.90 s (1.32 s) FA[a]: 5.13 s (2.67 s) Hit: 4.05 s (1.67 s)
Duration of alarm resolution with manual search	Lognormal	116 s (132 s)
Duration of alarm resolution with ETD	Gamma (translated, shape = 1)	Condition 1: 30 s (5 s) Condition 2: 60 s (10 s) Condition 3: 120 s (20 s)

[a]CR: correct rejection; FA: false alarm

Table 5

Characteristics of Reference Groups, Mean (SD)

Reference Group	Airport	False alarm rate	Training hours	Tenure / work experience	Age
Reference group 1	Airport 1	.025 (.039)	101.40 (31.47)	7.68 (4.85)	42.50 (10.52)
Reference group 2	Airport 2	.040 (.046)	28.56 (12.41)	8.24 (5.78)	36.55 (8.46)
Reference group 3	Airport 2	.049 (.031)	2.58[a] (2.00)	< 1 year (-)	30.81 (10.93)

[a]Reference group 3 received initial training using another computer based training; the number of training hours could therefore not be determined exactly.

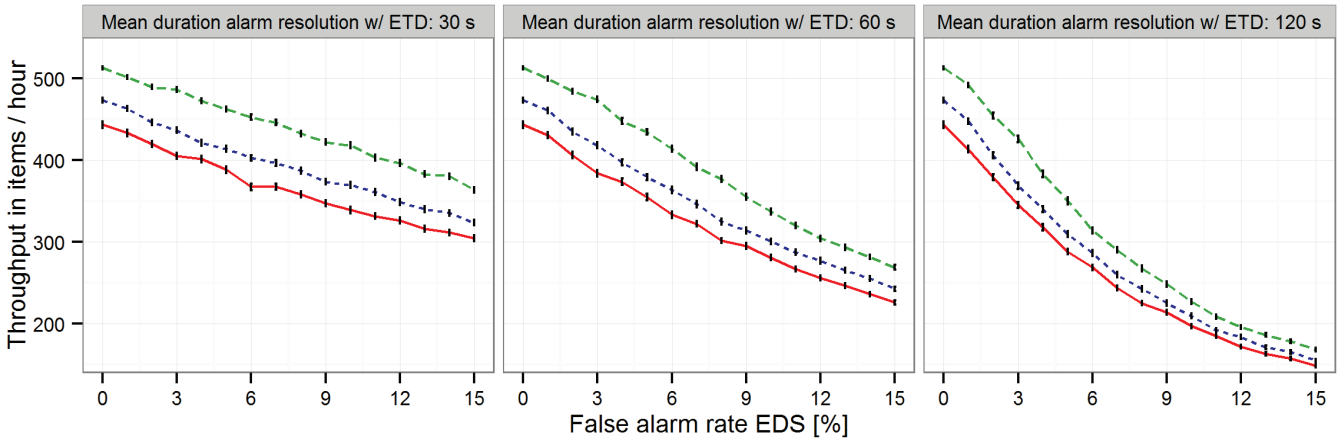

Figure 7. Mean human Mean and standard error (over 200 simulation runs) of throughput in items per hour, depending on false alarm rate of EDS (zero representing the baseline without EDS) and on reference group, green dashed: reference group 1 (airport 1, tenure > 2 years), blue dotted: reference group 2 (airport 2, tenure > 2 years), red solid: reference group 3 (airport 2, tenure < 1 year), and mean duration of alarm resolution using ETD of 30, 60 and 120s.

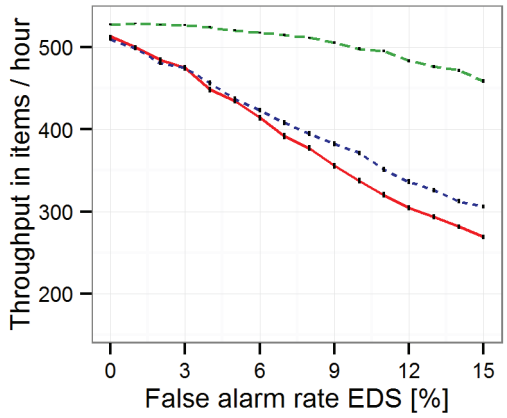

Figure 8. Mean and standard error of throughput in items per hour, depending on false alarm rate of EDS, either with: red solid: single security officer resolving alarms, blue dotted: X-ray screener assisting with alarm resolution in case screening process is interrupted, green dashed: second security officer assigned to resolution of alarms.

Results. Figure 7 shows the simulated throughput for a single checkpoint lane depending on reference group and false alarm rate of the EDS, whereby zero represents the absence of an EDS. As expected, capacity was negatively affected by the EDS's false alarm rate if no adaptions were made to cope with the increased workload due to additionally required alarm resolutions with ETD. This negative effect strongly depended on the average time required to resolve

the alarms by the EDS using ETD. The results were also largely dependent on whether X-ray screeners were well trained and experienced.

Figure 8 shows the relationship between capacity and the false alarm rate of the EDS for the standard security lane and the two measures that could be used to minimize negative effects on throughput as explained in the previous section. As could be expected, assigning a second security officer to the task of resolving alarms massively reduced the impact of the EDS's false alarm rate on throughput. Within the simulation, instructing the X-ray screener to resolve one alarm while the screening process is interrupted only started having a positive effect on throughput at higher levels of false alarm rate.

General Conclusion. The results of the discrete event simulation indicate that the baggage throughput of an airport security checkpoint can be strongly affected by EDS. This effect is mainly due to the time needed for alarm resolution using ETD, which highlights the importance of fast ETD alarm resolution procedures (e.g. efficient trace sampling) and a short analysis time of the equipment. Not only the false alarm rate of the EDS machine and alarm resolution time of ETD, but also the false alarm rates of the X-ray screeners were found to be very important. Training has been shown to reduce false alarm rates (Koller et al., 2009). Potential decreases in baggage throughput due to the introduction of an EDS could therefore be at least partially compensated by having well trained X-ray screeners.

Having a second security officer to expedite alarm resolution could reduce the negative impact of an EDS on throughput, while help by the X-ray screener with alarm resolution seems not to be a useful option based on the simulation results. A field study or a further work analysis combined with simulation could clarify if more coordinated assistance with alarm resolution by the X-ray screener (i.e. by only performing tasks that do not prolong the interruption of the X-ray screening process) has the potential to increase capacity.

General Discussion

This thesis addresses several research questions regarding the X-ray images inspection of passenger baggage, from a comparison with traditional visual search to the effects of automation on the X-ray image inspection performance and checkpoint capacity.

Manuscript 1 found visual processing and short-term memory, but not processing speed to be associated with higher performance in X-ray image inspection. Visual processing was also associated with performance in a variant of the L/T-letter search task. A mediation analysis however showed that other aspects of visual processing (a broad ability in terms of the Cattell–Horn–Carroll theory; Carroll, 1993; 2003; Cattell, 1941; Horn, 1965) are relevant for the two tasks. A comparison between students and professionals revealed that professionals outperformed students in the X-ray image inspection task, but not in the L/T-letter search task. Yet, for both tasks, the two samples did not differ significantly regarding the associations between the visual-cognitive abilities and performance.

In Manuscript 2, it was shown that d_a with a slope parameter of 0.5-0.6 is a more valid detection measure than d' and A'. Especially when there are large differences in response tendency, the use of an invalid detection measure can wrongly indicate a significant difference in detection performance (sensitivity) or lack thereof. In terms of signal detection theory (SDT), the slope parameter of 0.5-0.6 means that the signal-plus-noise distribution has a higher variance than the noise distribution.

Manuscript 3 investigated whether screeners can apply a decision strategy that focuses on the detection of novel threats. The study showed that the application of such a decision strategy resulted in a criterion shift, increasing both the hit rate and false alarm rate, whereas sensitivity remained constant. An e-learning module did not increase the criterion shift, screeners that scored higher on a test on the recognition of everyday objects in X-ray images found more novel items when they shifted their decision criterion.

Roughly in line with Manuscript 2, Manuscript 4 found d_a with slope of 0.65 to be a more valid detection measure for X-ray image inspection than d'. The experiment further showed that participants were able to screen for 60 min without a decline in detection performance. Performance thereby did not differ between participants that had a 10-min break every 20 min and participants that screened for 60 min straight. Participants without breaks however reported more distress.

Manuscript 5 evaluated two different scenarios for the introduction of explosive detection systems for cabin baggage (EDSCB) with two experiments, taking into account that the optimal use of automation might depend on the screeners' performance without automation and their familiarity with automation aids (Parasuraman & Manzey, 2010; Sauer et al., 2016; Strauch, 2017). The results show that, without automation, humans have difficulty with the detection of mere explosives. EDSCB with automated decision was shown to improve the detection of explosives and the detection of IEDs for inexperienced screeners. However, on-screen alarm resolution (OSAR) improved only the detection of IEDs for inexperienced screeners.

Manuscript 6 evaluated how EDSCB with automated decision, as investigated in Manuscript 5, would affect the baggage throughput of an airport security checkpoint by using discrete event simulation. The results suggest that without countermeasures, throughput declines with an increasing false alarm rate of the EDSCB, but also strongly depends on the false alarm rate of the screeners and the duration of alarm resolution with ETD.

All manuscripts include studies that were conducted in laboratory settings with experimental designs, providing high internal validity. However, as applied research projects, the studies also focused on providing high ecological validity. To increase the chance that the findings also hold in the everyday work life of airport security screening, all studies were conducted with professional participants and real X-ray images of passenger bags displayed on a user interface that is similar to practice – with the exception of Manuscript 1, which explicitly compared professionals with students and X-ray image inspection with traditional visual search. Conducting research with the specific population of professional screeners and in the specific context of X-ray image

inspection allows to test the robustness of theories developed in other contexts and with other participants (Brewer & Crano, 2014). This can provide theoretical contributions by revealing important moderators of the investigated effects (Petty & Cacioppo, 1996).

Theoretical contributions. Manuscript 1 found visual-cognitive abilities derived from CHC to be relevant for L/T-letter search tasks and for professional X-ray image inspection. However, the results also imply that other aspects of the visual-cognitive abilities are relevant for the two tasks. This suggests that the abstract search for certain deviations in letters has different underlying cognitive processes compared to searching for familiar objects, for which recognition is needed (Wolfe, 1998). When considering that recognition improves with familiarity of the target object, it is not surprising that professionals performed better in the X-ray image inspection task, even despite this task being designed to not require experience or training (by using black-and-white images and only guns and knives as threats). However, the visual-cognitive abilities were also relevant for the professional screeners. Future research should verify whether the association between the visual-cognitive abilities and performance in X-ray image inspection really is similar for students and professional screeners and thereby independent of expertise, or whether our study failed to find a moderation effect due to a lack of statistical power.

Manuscript 1 confirmed that X-ray image inspection is not a mere visual search task, but also strongly depends on object recognition and should therefore be seen as a search and decision task (Koller et al., 2009). Because the decision in X-ray image inspection is difficult and imperfect, the hit rate and false alarm rate depend on response tendency. This dependency can be well described with signal detection theory (for an introduction see Gescheider, 1997; Green & Swets, 1966; Macmillan & Creelman, 2005; T. D. Wickens, 2001). Our studies showed in line with research in other domains (Macmillan & Creelman, 2005) that response tendency (the criterion) can be manipulated with instruction, indirectly by using confidence ratings (Manuscript 2), or target prevalence (Manuscript 4).

The results further suggest a *z*ROC slope of smaller than one, which in terms of signal detection theory means that the target present distribution has a higher standard deviation than the

target absent distribution. Similar slope parameters have been found in other studies of passenger baggage screening and for the inspection of medical X-ray images (Kundel, 2000). However, we make the theoretical argument that the slope parameter cannot be fix and likely depends on sensitivity and the target set. It would be interesting and useful to identify the exact determinants of the slope parameter and the underlying cognitive process in the future. This would likely require a more detailed model of X-ray image inspection. Attempts to model X-ray image inspection as a series searches and decisions on the level of single objects in a bag (Wales, Halbherr, & Schwaninger, 2009; Wolfe & Van Wert, 2010) have offered some explanation on the interaction between the criterion and response times but do not yet provide an explanation for the slope parameter. The eye tracking data of Manuscript 2 strongly supports that X-ray image inspection heavily depends on a decision component in addition to a search component. Beyond replicating the target prevalence effect, Manuscript 4 supports the finding that the target prevalence effect is caused by implicitly learning the prevalence rather than by instruction (Ishibashi & Kita, 2014; Lau & Huang, 2010), which is in turn consistent with approaches to model visual search as Bayesian optimal foraging (Cain, Vul, Clark, & Mitroff, 2012).

Manuscript 2 and Manuscript 4 also provide methodological contributions. It was shown that the use of d' or A' can lead to wrong conclusions when criterion shifts are involved and d_a with a slope parameter of 0.5-0.6 is a more valid detection measure. However, as already mentioned above, the slope parameter is not necessarily constant and a conservative approach to test for a difference in sensitivity would be to use slope parameters of 1 (i.e. d') and of 0.5 as upper and lower bounds. Also A_g estimated with confidence ratings showed promising results and should be further investigated in the future. Besides the validation of detection measures, Manuscript 2 shows with a simulation (in the appendix of the manuscript) that pooling hit rates and false alarm rates across participants can severely distort the shape of an ROC curve and therefore pooling *z*-transformed hit rates and false alarm rates is recommended (under the assumption that Gaussian signal detection theory holds true).

Manuscript 5 has mainly contributed to research on automation. Sauer et al. (2018) found an automation aid to increase the performance of students inspecting X-ray images for guns and knives when system reliability was high. Whereas one might expect that these results translate to professionals and other threat categories, the two experiments of Manuscript 5 showed that EDSCB as an automation aid (i.e. with on-screen alarm resolution, OSAR) was very limited in increasing detection performance despite high system reliability in terms of sensitivity (as defined by signal detection theory). The experiments suggest that the evaluation of automation should differentiate between the search component and the decision component of X-ray image inspection. OSAR only increased the detection of IEDs for inexperienced screeners and did not increase the detection of explosives, neither for experienced nor inexperienced screeners. In contrast to IEDs, bare explosives lack distinctive features and look like harmless organic mass (Jones, 2003). In line with (Cullen et al., 2013), OSAR therefore likely helps with guiding attention. After the screeners' attention is guided to the critical object, they can make the correct decision in case of IEDs. However, they do not comply with the automation aid when the distinctive features of IEDs are lacking, possibly because of the high number of false alarms as found in other studies (Dixon, Wickens, & McCarley, 2007; Meyer, Wiczorek, & Günzler, 2014; C. D. Wickens & Dixon, 2007). This phenomenon is known as the "cry-wolf effect" (Bliss et al., 1995; Parasuraman et al., 2000). Future research should identify the preconditions that allow automation aids to assist with the decision component of X-ray image inspection. On a more general level, Manuscript 5 shows that it is difficult generalize findings in automation across different tasks but might actually even be specific to sub-tasks (e.g. separately for search and decision).

Practical implications. Beyond the theoretical implications mentioned above, this thesis provides several practical implications. When the X-ray image inspection performance of screeners is evaluated by security companies, airports, or regulators, our findings suggest that this should either be done on the basis of hit and false alarm rates or, when the evaluation should be independent of response tendency, based on d_a with a slope of around 0.6 (Manuscripts 2 and 4). When d' is used instead, screeners with the tendency to declare a bag as "not

ok" are likely disadvantaged. It was also shown that screeners can shift their criterion when instructed accordingly (Manuscript 3). This could be used to screen the baggage of high risk passengers more thoroughly. The results also suggest that improving the recognition of everyday objects might increase the screeners' ability to shift their criterion and assist with the detection of novel threats.

Manuscript 4 further showed that screeners can inspect X-ray images for 60 min without a decrease in performance and without a lower performance compared to screeners that received a 10-min break every 20 min. Because this result does not necessarily generalize to everyday work and because screeners without a break reported more distress, it would certainly be too early to recommend relaxing the regulatory restriction to 20 min of screening. However, based on the results regulators will likely allow to investigate longer screening durations with field studies, which then might result in more flexibility in designing the screening process.

Manuscript 5 showed that explosive detection systems for cabin baggage (EDSCB) are of limited use if implemented as automation aid without any further consideration. Future research should evaluate whether specific alarm resolution protocols or trainings allow screeners to incorporate the detection capability of the EDSCB and whether multi-view or 3D-CT images provide better results. Otherwise, EDSCB as automation aid seems to only provide some assistance with guiding attention of less experienced screeners. In that case, automated decision would be the better option – provided regulation allows to use EDSCB with a threshold that results in an acceptable false alarm rate. It was also shown with a discrete event simulation model that EDSCB with automated decision can be operationally feasible if fast explosive trace detection can be used to resolve the alarms of the EDSCB.

Conclusion

Inspecting X-ray images of passenger baggage is an important element of aviation security. Because mistakes in this task can have huge, even fatal consequences, research can provide a valuable contribution by reducing these errors. Existing research – e.g. on visual search, object recognition, automation, and vigilance – offers a good basis to investigate questions raised by new technologies, but the generalizability of findings across settings and populations is often unclear. By putting this generalizability to test, the experiments of this thesis provide important insights for practitioners in airport security and extend the theories and findings on which they are based by narrowing down their boundary conditions and identifying potential moderators. Manuscript 1 highlights that X-ray image inspection should not be seen as a mere visual search task by showing that the two tasks depend on different aspects of visual-cognitive abilities, whereas there was no difference found between students and professionals regarding the effect of these visual-cognitive abilities. Because X-ray image inspection often relies on decisions under uncertainty, the screeners' response is influenced by their response tendency. The experiments of this thesis show that the screeners' response tendency can be affected by instruction (Manuscript 2 and 3), but also depends on the screeners' confidence (Manuscript 2) and the target prevalence (Manuscript 4). It can therefore be crucial to evaluate performance in X-ray image inspection on valid detection measures that are independent of response tendency. When investigating the effect of time on task and breaks, Manuscript 4 showed that X-ray image inspection is not directly comparable to findings in vigilance research. Manuscript 5 suggests that EDSCB as an automation aid assists with attention allocation, but not necessarily with the decision. This distinction can explain why automation aids help students to find guns and knives and less experienced screeners to find IEDs, but do not increase the detection of explosives or for experienced screeners in general. EDSCB with automated decision however increases the detection of explosives (Manuscript 5) and can be operationally feasible when alarm resolution is fast and screeners produce few false alarms (Manuscript 6).

This thesis shows that a more integrative theoretical framework on search and recognition is needed that also covers recognition under high uncertainty. Also extending the existing frameworks on automation aids to more strongly differentiate between different subtasks seems promising for future research.

References

Alexander, R. G., & Zelinsky, G. J. (2011). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8), 1–15. https://doi.org/10.1167/11.8.9

Alexander, R. G., & Zelinsky, G. J. (2012). Effects of part-based similarity on visual search: The Frankenbear experiment. *Vision Research*, *54*, 20–30. https://doi.org/10.1016/J.VISRES.2011.12.004

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). New York, NY: Oxford University Press.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x

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

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

Baum, P. (2016). *Violence in the skies: A history of aircraft hijacking and bombing*. Chichester, UK: Summersdale Publishers.

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., & Mitroff, S. R. (2013). Assessing visual search performance differences between Transportation Security Administration Officers and nonprofessional visual searchers. *Visual Cognition*, *21*(3), 330–352. https://doi.org/10.1080/13506285.2013.790329

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

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. *Applied Cognitive Psychology*, *29*(1), 142–148. https://doi.org/10.1002/acp.3083

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38*(11), 2300–2312. https://doi.org/10.1080/00140139508925269

Bolfing, A., & Schwaninger, A. (2009). Selection and pre-employment assessment in aviation security x-ray screening. *Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology* (5–12). Zurich. https://doi.org/10.1109/CCST.2009.5335571

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 11–26). New York, NY.

Brunstein, A., & Gonzalez, C. (2011). Preparing for novelty with diverse training. *Applied Cognitive Psychology*, *25*(5), 682–691. https://doi.org/10.1002/acp.1739

Butler, V., & Poole, R. W. (2002). Rethinking Checked-Baggage Screening. *Reason Public Policy Institute, Policy Studies*, *297*, 1–25.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian Optimal Foraging Model of Human Visual Search. *Psychological Science*, *23*(9), 1047–1054. https://doi.org/10.1177/0956797612440460

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Carrasco, M. (2014). Spatial attention: Perceptual Modulation. In S. Kastner & A. C. Nobre (Eds.), *The Oxford handbook of attention* (pp. 183–230). Oxford University Press.

Carrasco, M. (2018). How visual spatial attention alters perception. *Cognitive Processing*, *19*(S1), 77–88. https://doi.org/10.1007/s10339-018-0883-4

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York,

NY: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence* (pp. 5–21). Pergamon. https://doi.org/10.1016/B978-008043793-4/50036-2

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 592.

Chan, L. K. H., & Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 415–429. https://doi.org/10.1002/wcs.1235

Clark, K., Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2012). Overcoming hurdles in translating visual search research between the lab and the field. In M. D. Dodd & J. H. Flowers (Eds.), *The Influence of Attention, Learning, and Motivation on Visual Search* (pp. 147–181). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-4794-8_7

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

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.

Cutler, V., & Paddock, S. (2009). Use of Threat Image Projection (TIP) to enhance security performance. *Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology*. https://doi.org/10.1109/CCST.2009.5335565

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: are automation false alarms worse than misses? *Human Factors*, *49*(4), 564–572. https://doi.org/10.1518/001872007X215656

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), 1–36. https://doi.org/10.1167/11.5.14

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, *25*(4), 249–263.

https://doi.org/10.3758/BF03198804

Gescheider, G. A. (1997). *Psychophysics : The Fundamentals*. Mahwah, NJ: L. Erlbaum Associates.

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, (93–97). SAGE PublicationsSage CA: Los Angeles, CA. 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*. New York: Wiley.

Halbherr, T., Schwaninger, A., Budgell, G. R., & Wales, A. W. J. (2013). Airport security screener competency: A cross-sectional and longitudinal analysis. *The International Journal of Aviation Psychology*, *23*(2), 113–129. https://doi.org/10.1080/10508414.2011.582455

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. *Proceedings of the 39th IEEE International Carnahan Conference on Security Technology*, 189–192. https://doi.org/10.1109/CCST.2005.1594876

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006). Increased detection performance in airport security screening using the X-Ray ORT as pre-employment assessment tool. *Proceedings of the 2nd International Conference on Research in Air Transportation*, 393–397.

Hardmeier, D., & Schwaninger, A. (2008). Visual cognition abilities in x-ray screening.

Proceedings of the 3rd International Conference on Research in Air Transportation, 311–316.

Hattenschwiler, N., Merks, S., & Schwaninger, A. (2018). Airport security X-ray screening of hold baggage: 2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol. *Proceedings of the 52nd IEEE International Carnahan Conference on Security Technology*, 1–5. https://doi.org/10.1109/CCST.2018.8585713

Hattenschwiler, N., Michel, S., Kuhn, M., Ritzmann, S., & Schwaninger, A. (2015). A first exploratory study on the relevance of everyday object knowledge and training for increasing efficiency in airport security X-ray screening. *Proceedings of the 49th IEEE International Carnahan Conference on Security Technology*. https://doi.org/10.1109/CCST.2015.7389652

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

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in x-ray screening. *Proceedings of the 38th IEEE International Carnahan Conference on Security Technology*, 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.

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic and developmental study of the structure among primary mental abilities*. University of Illinois.

Humphreys, G. W., & Mavritsaki, E. (2012). Models of visual search: From abstract function to biological constraint. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (Second edi, pp. 57–75). New York, NY: Guilford Press.

IONSCAN 500DT. (n.d.). Retrieved June 25, 2015, from http://www.smithsdetection.com/index.php/products-solutions/explosives-narcotics-

detection/61-explosives-narcotics-detection/ionscan-500dt.html#.VYvzAWP66sE

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, & Psychophysics*, *74*(1), 115–123. https://doi.org/10.3758/s13414-011-0225-4

Jones, T. L. (2003). *Court security: A guide for post 9-11 environments*. Springfield, IL: Charles C. Thomas.

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. *Journal of Transportation Security*, *1*(2), 81–106. https://doi.org/10.1007/s12198-007-0006-4

Koller, S. M., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd IEEE International Conference on Research in Air Transportation* (399–402).

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*. https://doi.org/10.1080/00140139408964946

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kuhn, M. (2017). Centralised image processing: The impact on security checkpoints. *Aviation Security International Magazine*.

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.), *Medical imaging 2000: Image perception and performance* (pp. 135–144). SPIE.

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

Lavie, N., & De Fockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin & Review*, *12*(4), 669–674. https://doi.org/10.3758/BF03196756

Mackworth, N. H. (1948). The breakdown of vigilance durning prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21. https://doi.org/10.1080/17470214808416738

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

Madhavan, P., Gonzalez, C., & Lacson, F. C. (2007). Differential base rate training influences detection of novel targets in a complex visual inspection task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 392–396. https://doi.org/10.1177/154193120705100451

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., … Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*. https://doi.org/10.1037/1528-3542.2.4.315

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

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

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–181). New York, NY: Guilford Press.

Meghanathan, R. N., van Leeuwen, C., & Nikolaev, A. R. (2014). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in Human Neuroscience*, *8*, 1063. https://doi.org/10.3389/fnhum.2014.01063

Mendes, M., Schwaninger, A., & Michel, S. (2013). Can laptops be left inside passenger bags if motion imaging is used in X-ray security screening? *Frontiers in Human Neuroscience*, *7*(October), 1–10. https://doi.org/10.3389/fnhum.2013.00654

Menneer, T., Donnelly, N., Godwin, H. J., & Cave, K. R. (2010). High or low target prevalence increases the dual-target cost in visual search. *Journal of Experimental Psychology: Applied*, *16*(2), 133–144. https://doi.org/10.1037/a0019569

Metz, C. E., Herman, B. A., & Shen, J. H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*(9), 1033–1053. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9612889

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

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors*, *56*(5), 840–849. https://doi.org/10.1177/0018720813512865

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

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research*, *51*(13), 1526–1537. https://doi.org/10.1016/j.visres.2010.09.003

Novakoff, A. K. (1993). FAA bulk technology overview for explosives detection. In J. M. Connelly & S. M. Cheung (Eds.) (Vol. 1824, pp. 2–12). SPIE. https://doi.org/10.1117/12.142893

Nuechterlein, K. H., Parasuraman, R., & Jiang, Q. (1983). Visual sustained attention: Image degradation produces rapid sensitivity decrement over time. *Science*.

https://doi.org/10.1126/science.6836276

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, *33*(1), 1–8. https://doi.org/10.1177/002224379603300101

Poole, B. J., & Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention. *Quarterly Journal of Experimental Psychology*, *62*(7), 1430–1454. https://doi.org/10.1080/17470210802479329

Radvansky, G. A., & Ashcraft, M. H. (2016). *Cognition* (6th ed.). Pearson.

Raven, J., Raven, J. C., & Court, J. H. (2003, updated 2004). *Manual for Raven's progressive matrices and vocabulary scales*. San Antonio, TX: Harcourt Assessment.

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision*, *8*(15), 15–15. https://doi.org/10.1167/8.15.15

Roper, Z. J. J., Cosman, J. D., & Vecera, S. P. (2013). Perceptual load corresponds with factors known to influence visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1340–1351. https://doi.org/10.1037/a0031616

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*, 1–18. https://doi.org/10.3389/fnhum.2015.00439

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training:

Effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. https://doi.org/10.1080/00140139.2015.1094577

Schwaninger, A. (2005). Increasing efficiency in airport security screening. *WIT Transactions on the Built Environment*, *82*, 407–416.

Schwaninger, A., Hardmeier, D., Riegelnig, J., & Martin, M. (2010). Use It and Still Lose It? *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *23*(3), 169–175. https://doi.org/10.1024/1662-9647/a000020

Sheridan, T. B., & Verplank, W. (1978). Human and Computer Control of Undersea Teleoperators. *Technical Report, MIT Man-Machine Systems Laboratory*. Cambridge, MA.

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

Singh, S., & Singh, M. (2003). Explosives detection systems (EDS) for aviation security. *Signal Processing*, *83*(1), 31–55.

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

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*. https://doi.org/10.1002/acp.3255

Strauch, B. (2017). The automation-by-expertise-by-training interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(2), 204–228. https://doi.org/10.1177/0018720816665459

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. *Attention, Perception, & Psychophysics*, *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

Wales, A. W. J., Halbherr, T., & Schwaninger, A. (2009). Using speed measures to predict performance in X-ray luggage screening tasks. *Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology*, 212–215. https://doi.org/10.1109/CCST.2009.5335536

Wells, K., & Bradley, D. A. (2012). A review of X-ray explosives detection techniques for checked baggage. *Applied Radiation and Isotopes*, *70*(8), 1729–1746. https://doi.org/10.1016/j.apradiso.2012.01.011

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*. New York: Oxford University Press.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33–39. https://doi.org/10.1111/1467-9280.00006

Wolfe, J. M., Cain, M. S., & Aizenman, A. M. (2019). Guidance and selection history in hybrid foraging visual search. *Attention, Perception, & Psychophysics*, *81*(3), 637–653. https://doi.org/10.3758/s13414-018-01649-5

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., Oliva, A., Horowitz, T. S., Butcher, S. J., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research*, *42*(28), 2985–3004.

https://doi.org/10.1016/S0042-6989(02)00388-7

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

Yu, R., & Wu, X. (2015). Working alone or in the presence of others: exploring social facilitation in baggage X-ray security screening tasks. *Ergonomics*, *58*(6), 857–865. https://doi.org/10.1080/00140139.2014.993429

Zhaoping, L., & Frith, U. (2011). A clash of bottom-up and top-down processes in visual search: The reversed letter effect revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 997–1006. https://doi.org/10.1037/a0023099

Acknowledgements

- Most of all, I would like to thank Adrian Schwaninger for providing the opportunity to conduct research in an interesting field, for the many valuable inputs, for teaching me how to transform complicated thoughts into comprehensible text and for motivating me by making the impact of our research visible.
- I am very grateful to Klaus Opwis, my thesis supervisor, for making this thesis possible, for offering reflection on my work, and for being a steady but friendly reminder to keep my pace in this endeavor.
- I am especially thankful to my research team for being my companions on this journey and for perseveringly offering support in busy times.
- I wish to thank Eleanor Nevill for the great emotional support and for providing balance between my life domains.
- I thank the doctoral committee for evaluating this work.
- I would also like to thank my parents for providing a childhood that left nothing to be desired and for laying the foundation of my academic journey.

Appendix

1. 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, 1-17 https://doi.org/10.3389/fpsyg.2019.00525
2. Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (2019). Detection measures for visual inspection of X-ray images of passenger baggage. *Attention, Perception, & Psychophysics*, 1-15. doi:10.3758/s13414-018-01654-8
3. Sterchi, Y., Hättenschwiler, N., Michel, S. & Schwaninger, A. (2017). Relevance of visual inspection strategy and knowledge about everyday objects for X-ray baggage screening. *Proceedings of the 51st IEEE International Carnahan Conference on Security Technology*, 1-6. doi:10.1109/CCST.2017.8167812
4. Buser, D.[4], Sterchi, Y.[4] & Schwaninger A. (2019) *Why stop after 20 minutes? Breaks and target prevalence in a one hour X-ray image inspection task.* Manuscript under revision.
5. Hättenschwiler, N., Sterchi, Y., Mendes, M., & Schwaninger, A. (2018). Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection. *Applied Ergonomics, 72*, 58-68. doi:10.1016/j.apergo.2018.05.003
6. Sterchi, Y., & Schwaninger, A. (2015). A first simulation on optimizing EDS for cabin baggage screening regarding throughput. *Proceedings of the 49th IEEE International Carnahan Conference on Security Technology*, 55-60. doi:10.1109/CCST.2015.7389657

[4] Joint first authorship

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 March 2019
doi: 10.3389/fpsyg.2019.00525

Traditional Visual Search vs. X-Ray Image Inspection in Students and Professionals: Are the Same Visual-Cognitive Abilities Needed?

Nicole Hättenschwiler†, Sarah Merks†, Yanik Sterchi and Adrian Schwaninger*

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

The act of looking for targets amongst an array of distractors is a cognitive task that has been studied extensively over many decades and has many real-world applications. Research shows that specific visual-cognitive abilities are needed to efficiently and effectively locate a target among distractors. It is, however, not always clear whether the results from traditional, simplified visual search tasks conducted by students will extrapolate to an applied inspection tasks in which professionals search for targets that are more complex, ambiguous, and less salient. More concretely, there are several potential challenges when interpreting traditional visual search results in terms of their implications for the X-ray image inspection task. In this study, we tested whether a theoretical intelligence model with known facets of visual-cognitive abilities (visual processing *Gv*, short-term memory *Gsm*, and processing speed *Gs*) can predict performance in both a traditional visual search task and an X-ray image inspection task in both students and professionals. Results showed that visual search ability as measured with a traditional visual search task is not comparable to an applied X-ray image inspection task. Even though both tasks require aspects of the same visual-cognitive abilities, the overlap between the tasks was small. We concluded that different aspects of visual-cognitive abilities predict performance on the measured tasks. Furthermore, although our tested populations were comparable in terms of performance predictors based on visual-cognitive abilities, professionals outperformed students on an applied X-ray image inspection task. Hence, inferences from our research questions have to be treated with caution, because the comparability of the two populations depends on the task.

Keywords: visual search, visual inspection, letter search task, X-ray image inspection, visual-cognitive abilities, students, professionals

OPEN ACCESS

Edited by:
Davood Gozli,
University of Macau, China

Reviewed by:
Jason Rajsic,
Vanderbilt University, United States
Britt Anderson,
University of Waterloo, Canada

***Correspondence:**
Sarah Merks
sarah.merks@fhnw.ch

†These authors share first authorship

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *06 October 2018*
Accepted: *22 February 2019*
Published: *21 March 2019*

Citation:
Hättenschwiler N, Merks S, Sterchi Y and Schwaninger A (2019) Traditional Visual Search vs. X-Ray Image Inspection in Students and Professionals: Are the Same Visual-Cognitive Abilities Needed? Front. Psychol. 10:525. doi: 10.3389/fpsyg.2019.00525

INTRODUCTION

Visual search, the act of looking for targets amongst an array of distractors, is a demanding cognitive task (e.g., Treisman and Gelade, 1980) that has many real-world applications. Some individuals conduct visual search tasks professionally, for example, airport security officers (screeners) who visually inspect X-ray images of passenger baggage to search for prohibited items or radiologists who are looking for cancer in mammograms. Because search errors can have huge, even fatal,

consequences in such professional applications, research can provide a valuable contribution by reducing these errors. The ability to locate a target amongst an array of distractors has been studied extensively over many decades (for reviews see e.g., Carrasco, 2011, 2014, 2018; Eckstein, 2011; Nakayama and Martini, 2011; Humphreys and Mavritsaki, 2012; Chan and Hayward, 2013). Research also shows that specific visual-cognitive abilities are needed to effectively and efficiently locate a target among distractors. However, many of the studies on visual search have been conducted using traditional, simplified tasks with salient stimuli and have been done with non-professional searchers (mostly students). These studies have provided vital insights into the cognitive mechanisms underlying visual search due to the high experimental control. It is, however, not clear whether the results from such traditional, simplified visual search tasks extrapolate to real-world inspection tasks in which professionals search for targets that are more complex, ambiguous, and/or less salient (e.g., Biggs and Mitroff, 2014; Radvansky and Ashcraft, 2016, p. 257). It is also unclear to what extent findings based on student samples can be transferred to professionals who often rely on extensive training and experience. To address these issues, we first introduce visual search in general before comparing insights on traditional visual search tasks vs. a real-world application, namely X-ray image inspection, and considering the populations conducting these search tasks.

Visual Search and Visual Search Tasks

Visual search typically involves an active scan of the visual environment for a particular target among many distractors. This is a demanding cognitive task requiring specific visual-cognitive abilities (Treisman and Gelade, 1980). Over the past several decades, psychological research has made tremendous headway in understanding the underlying cognitive processes when performing visual search tasks and the mechanisms that allow a successful identification of target items (Clark et al., 2012). Search thereby involves several processes such as perception (i.e., processing and interpreting visual features), attention (i.e., allocating resources to the relevant areas of a visual area), and memory (for reviews see e.g., Carrasco, 2011, 2014, 2018; Eckstein, 2011; Nakayama and Martini, 2011; Humphreys and Mavritsaki, 2012; Chan and Hayward, 2013; storing a representation of the target item or items). To conduct visual search and inspection, certain visual-cognitive abilities such as attention, memory, visual processing, or processing speed have been found to correlate with higher performance.

A known example of a traditional visual search task that has been studied in many variations is the *L/T-letter search task*. According to Treisman and Gelade (1980), this is called a conjunction search task. Conjunction search involves distractors (or a group of distractors) that may differ from each other but exhibit at least one common feature with the target and therefore require a combination of features to distinguish them (Shen et al., 2003). For example, the letters T and L share exactly the same features, differing only in their spatial arrangement (*L/T-letter* search task: Treisman and Gelade, 1980). In one variation of this task, participants are asked to identify the perfectly shaped letter *T* (target) surrounded by many distractor letters including *Ls* and symmetrical and asymmetrical *Ts*. The efficiency of such a conjunction search in terms of accuracy and reaction time depends on the distractor ratio and the number of distractors present (McElree and Carrasco, 1999), and the negative effect of limiting reaction time on accuracy is alleviated by training (Reavis et al., 2016).

In more complex real-world visual search applications, humans sometimes conduct visual search and inspection tasks professionally. For example, radiologists inspect mammograms for cancer (e.g., Nodine and Kundel, 1987; Krupinski, 1996; Horowitz, 2017) or screeners inspect X-ray images for prohibited items (Drury, 1975; Koller et al., 2009; Wales et al., 2009; Mitroff et al., 2015). In these scenarios, professionals search for targets that are less artificial and more familiar to them. They must use their prior knowledge in order to accurately and efficiently locate more ambiguous targets (Wolfe et al., 2019) such as guns and knives or cancer cells and so forth among distractors with much more complex features compared to a traditional conjunction search task. Searching for familiar stimuli relies on object recognition (Wolfe, 1998). Here, top-down processing allows searchers to more efficiently identify targets with greater complexity (Zhaoping and Frith, 2011). X-ray image inspection is therefore best described as a search and decision task (Spitz and Drury, 1978; Koller et al., 2009) that relies more heavily on the decision component compared to traditional search tasks with unambiguous stimuli. Nonetheless, visual search with complex objects is assumed to rely on the same active scanning processes as conjunction search (e.g., L/T-letter search task) with less complex, contrived laboratory stimuli (Alexander and Zelinsky, 2011, 2012).

When translating results from a traditional visual search task such as an L/T-letter search task to X-ray image inspection and vice versa, it is necessary to consider differences in the nature of stimuli and the characteristics of searchers. Differences in stimuli include target and distractor complexity as well as the requirement of domain-specific knowledge of the searcher in order to successfully recognize the target (e.g., Biggs and Mitroff, 2014). On the other hand, targets in a traditional visual search task are often commonly known to have salient shapes and colors, whereas targets in X-ray image inspection tasks are not well-specified, not salient, and not predictable through the context (Bravo and Farid, 2004). The large variety of potential threat items and distracting objects in passenger bags makes X-ray image inspection a difficult task (Hättenschwiler et al., 2015; Sterchi et al., 2017). This calls for domain-specific knowledge, because screeners must know which items are prohibited and what they look like in X-ray images (Schwaninger, 2004, 2005, 2006). Due to the differences between traditional visual search tasks and X-ray image inspection, it is unclear whether they require the same visual-cognitive abilities. We shall discuss this in the next section. Because research on traditional visual search tasks and X-ray image inspection differs in regard to not only the task but also the examined population, we shall discuss differences between students and professional screeners in section Populations Conducting Visual Search.

Cognitive Abilities for Visual Search

Both traditional visual search and X-ray image inspection can be characterized as a basic, core cognitive task. As defined by Carroll (1993), a cognitive task is any task in which correct processing of mental information is critical for successful performance. Therefore, specific cognitive abilities are needed to perform such a task successfully. These abilities can be assessed with specific correlated measures that can predict performance. With regard to visual search and inspection, certain visual-cognitive abilities such as attention, memory, visual processing, or processing speed have been found to correlate with higher performance (Wolfe et al., 2002; Bolfing and Schwaninger, 2009). If individual differences in performance are found on visual search or inspection tasks, these can be seen as the direct manifestation of differences in an underlying ability or latent trait (Carroll, 1993, 2003).

There is a large number of such abilities and many theories aiming to integrate cognitive abilities. Today, the Cattell–Horn–Carroll theory (CHC) is widely accepted as the most comprehensive and empirically supported theory on the structure of human cognitive abilities, and it informs a substantial body of research and the ongoing development of intelligence tests (McGrew, 2005). The CHC theory states that the relationships among these cognitive abilities can be derived by classifying them into three different strata: Stratum I, "narrow" abilities; Stratum II, "broad abilities"; and Stratum III, a single general ability also called *g* (Flanagan and Harrison, 2005). The factors describe stable and observable differences between individuals. However, the structure of the three strata is hierarchical, meaning that the abilities within one stratum (e.g., the narrow abilities of Stratum I) are positively intercorrelated, thereby allowing an estimation of Stratum II, the broad abilities. Likewise, the abilities of Stratum II have non-zero intercorrelations, thereby allowing an estimation of Stratum III. Hence, whereas the abilities within Strata I or II are related, a large amount of evidence shows that they are unique and reliably distinguishable (see e.g., Keith and Reynolds, 2012).

Visual processing (*Gv*), short-term memory (*Gsm*), and processing speed (*Gs*) are broad Stratum II abilities that are accepted components with a known influence on visual search and inspection performance. Therefore, they are included in most commonly used measures of intelligence (e.g., Stanford-Binet: Roid, 2003a,b; Wechsler Intelligence Scale: Wechsler, 1997). Visual processing (*Gv*) describes a broad ability to perceive, analyze, synthesize, and think in visual patterns, including the ability to store and recall visual representations. Short-term memory (*Gsm*) is characterized as the ability to apprehend and hold information in immediate awareness and then perform a set of cognitive operations on this information within a few seconds. Because analyzing, synthesizing, and thinking in visual patterns are also cognitive operations, *Gv* and *Gsm* are closely related, but can be distinguished by the limited capacity of short-term memory. Processing speed (*Gs*) describes the ability to quickly and accurately perceive visual details, similarities, and differences.

Several studies have confirmed the influence of higher scores in *Gv*, *Gsm*, and *Gs* on better performance in traditional visual search tasks (Eriksen and Schultz, 1979; Alvarez and Cavanagh, 2004). Cognitive abilities have also been linked to inspection performance in studies on X-ray image inspection with professionals (e.g., Schwaninger et al., 2004; Hardmeier et al., 2005; Hardmeier and Schwaninger, 2008). Detection performance decreases significantly if threat items are shown in close-packed bags, if threats are more superimposed by other items, and if they are shown in an unusual view. Studies linked the influence of mental rotation and figure–ground segregation, which are narrow abilities of visual processing (*Gv*), to higher X-ray image inspection performance (Wolfe et al., 2002; Bolfing and Schwaninger, 2009). Items presented from unusual or rotated viewpoints become more difficult to detect (effect of viewpoint; Palmer et al., 1981). Similarly, the position of a prohibited item in a bag and its superposition by other objects (effect of superposition), or the number and types of items in a bag that could attract attention (effect of bag complexity) also affect the difficulty in recognizing prohibited items. Bag complexity comprises the factors clutter (disarrangement, textural noise, chaos, etc.) and opacity (X-ray penetration of objects; see Schwaninger et al., 2008). Memory capacity, which can be classified as short-term memory (*Gsm*), is strongly associated with visual inspection in general (e.g., Lavie and DeFockert, 2005; Poole and Kane, 2009; Roper et al., 2013). In addition, processing speed (*Gs*) might be relevant for the efficiency of the visual inspection task (Salthouse, 1996). Based on the reviewed literature, the question arises whether the same visual-cognitive abilities can predict performance in a traditional visual search task and an X-ray image inspection task.

Populations Conducting Visual Search

As a positive correlation was found between certain visual-cognitive abilities and performance in X-ray screening, many European airports conduct preemployment assessments that test for these visual abilities and aptitudes when recruiting new personnel (e.g., X-Ray Object Recognition Test; see Hardmeier et al., 2005; Hardmeier and Schwaninger, 2008). Professional screeners conducting X-ray image inspection have therefore been selected accordingly, and they usually have a lot of experience on this specific task through many hours of training and years of job experience. In comparison, university students are the first choice as participants for traditional visual search research because they are an easily accessible population. Therefore, differences between professional screeners and students could be due either to characteristics of the searchers as a result of self and pre-employment selection or to training and job experience as professionals (Clark et al., 2012).

Training for threat detection has the goal of creating internal visual representations of objects and storing them in memory. To identify whether an object in an X-ray image is a threat or not, a searcher must successfully match the visual information of this object to representations stored in visual memory (Kosslyn, 1975, 1980). Depending on the similarity of objects and its features presented in an X-ray image to those stored in visual memory, the screener will then decide whether the respective

object is harmless or not. More familiar objects therefore need fewer recognized features in order to be identified successfully (Koller et al., 2009). Detection of objects—known and especially unknown—should therefore improve with training because features become familiar and are recognized better through repeated exposure. For example, features of guns and knives are known from everyday life and can therefore also be detected by novices without specific experience or training. However, screeners have been exposed to these objects more often and have therefore more detailed and specific target templates and are more familiar with them (Koller et al., 2009). However, other prohibited items that are rather uncommon or have never been seen before (e.g., improvised explosive devices, IEDs) become very difficult to recognize for novices if they have not been trained to recognize certain features of these threats (Schwaninger, 2004, 2005).

Current Study

Over the past several decades, psychological research has made tremendous headway in understanding the underlying cognitive processes when performing visual search tasks and the mechanisms that allow for the successful identification of target items (Clark et al., 2012).

However, most of the research on this theoretical basis was conducted with students using tasks applying artificial stimuli to allow for maximum experimental control (for reviews, see e.g., Duncan and Humphreys, 1989; Wolfe, 1994, 1998; Eckstein, 2011). It is therefore unclear to what extent professional X-ray image inspection relies on the same cognitive processes. Because the tasks in traditional visual search and X-ray image inspection are often conducted by different populations, it is also necessary to ask whether the two populations rely on the same cognitive processes. To date, no study has examined the influence of visual-cognitive abilities on visual search performance by comparing a traditional visual search task and an X-ray image inspection task.

Based on the literature on visual-cognitive abilities, we postulate a theoretical model in which several known facets (visual processing *Gv*, short-term memory *Gsm*, and processing speed *Gs*) can predict performance in a traditional visual search task and an X-ray image inspection task. We shall test this model on two populations (students and professionals) using the same experimental stimuli. This will provide an indication on whether the two populations require the same visual-cognitive abilities or whether visual-cognitive abilities can be compensated by experience and training in X-ray image inspection. To have a fair comparison, we created a traditional visual search task with *Ls* and *Ts* on a high difficulty level and an X-ray image inspection task with no need for domain-specific knowledge that included only black and white images as well as familiar target items such as guns and knives. Features of guns and knives as well as letters such as L or T, are known from everyday life experience and can therefore be recognized without specific experience and training. We used this comparison to address the following research questions: (1) Do different visual-cognitive abilities predict performance in a traditional visual search task and an X-ray image inspection task? (2) Do the results differ between students and professionals? Answers to these questions could provide important information on how well studies conducted with students and traditional visual search tasks can be generalized to professional X-ray image inspection.

TABLE 1 | Description of participants.

	N	Age	Gender	SPM
Students	128	$M = 25.7$	74% female	$M = 30.8$
		$SD = 6.4$		$SD = 3.0$
Professionals	112	$M = 43.7$	55% female	$M = 28.3$
		$SD = 11.9$		$SD = 4.2$

255 participants gave informed consent to be part of this experiment. 15 participants had to be excluded from statistical analyses (5.9% of the sample) due to a malfunction of a simulator (n = 4) or performance below chance (n = 11). Therefore, the final sample included 240 participants. SPM, Standard Progressive Matrices raw scores as a baseline measure of fluid intelligence.

METHODS

Participants

Table 1 reports the participants' descriptives. 128 participants were *students* from the University of Applied Sciences and Arts Northwestern Switzerland. 112 participants were *professionals* (airport security screeners employed at an international airport) 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 EU regulation (European Commission, 2015). The current research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board of the University of Applied Sciences and Arts Northwestern Switzerland.

Apparatus

We used six HP ProBooks 4730s and 4720s with Intel Core i5 2410M and 520M processors and 19″ TFT monitors. The six testing stations were separated, and the room was dimly lit for testing. Participants sat approximately 50 cm away from the monitor. Non-professional searchers were tested in the laboratory at the University of Applied Sciences and Arts. Professional searchers were tested at the test facilities of the Center for Adaptive Security Research and Applications (CASRA) using the same computers and monitors.

Stimuli

Visual Cognitive Test Battery

A visual-cognitive test battery (VCTB) was developed to measure a broad spectrum of visual-cognitive abilities assessing a wide variety of narrow abilities underlying *visual processing* (*Gv*), *short-term memory* (*Gsm*), and *processing speed* (*Gs*) in order to make predictions on visual search performance. The VCTB consists of 10 standardized tests scales taken mostly from well-established intelligence tests based on the CHC theory of intelligence (Cattell, 1941; Horn, 1965; Carroll, 1993, 2003). Four scales came from a major German intelligence test, the Leistungsprüfsystem 2 (LPS-2; Kreuzpointner et al., 2013). Three tests were taken from a cognitive development test, that assesses

TABLE 2 | Psychometric criteria of the VCTB test scales (objectivity, reliability, validity).

Test	Scale	Objectivity	Reliability	Validity
LPS	LPS 6: Mental rotation (Gs) LPS 7: Number of surfaces (Gs) LPS 8: Shape Comparison (Gs) LPS 10: Row comparison (Gs)	Standardized	Cronbach's α: 0.86–0.94 Split-half: 0.81–0.96	Factor analyses Correlations with g
WSI	WSI Slices (Gsm) WSI Mental rotation (Gsm) WSI Unfold (Gsm)	–	–	–
TVPS	TVPS Visual Memory (Gv) TVPS Form Constancy (Gv) TVPS Figure Ground (Gv)	Standardized	Cronbach's α: 0.74 Test–Retest: 0.71	–
SPM	SPM: Speed-Test	Standardized	Cronbach's α: 0.97–1.00 Split-half: > 0.90 Test–Retest: 0.80–0.90	Correlations with nonverbal IQ

Psychometric criteria are retrieved as follows: LPS from Kreuzpointner et al. (2013); TVPS from Brown et al. (2010); SPM from Horn (2009).

visual perceptual weaknesses and strengths—the Test of Visual Perceptual Skills (TVPS-3; Martin, 2006). Another three scales were used from a Swiss online assessment test for students (WSI; Hell et al., 2009; Päßler and Hell, 2012) In addition, we included Raven's standardized progressive matrices (SPM; Horn, 2009) as a general measure of fluid intelligence. Because most scales were originally in paper-and-pencil format, we created computer-based versions. **Table 2** reports the psychometrical criteria of the test scales.

Visual processing (Gv)

We assessed visual processing with three scales from the TVPS-3 (visual memory, form constancy and figure-ground segregation; see **Figure 1**). For visual memory, participants have to memorize a design for 5 s and then recognize this pattern from four alternatives presented on the next slide. The scale consists of 16 tasks and the score is the sum of correct responses. To measure form constancy, participants are instructed to find a target shape within five alternative, more complex patterns that can be rotated, increased, or decreased in size. There are 16 trials and the score is the number of correct responses. Figure-ground segregation is defined as the ability to recognize a target shape within a very cluttered, busy background. Participants have to choose one out of four complex patterns that include the target shape. There are 16 trials, and the score is the number of correct responses.

Short-term memory (Gsm)

Short-term memory was measured using three scales from the WSI (slicing, spatial rotation, and unfold; **Figure 2**). Slicing can be referred to as another form of three-dimensional visualization. During the task, participants see a full three-dimensional object and next to this a cube with two or three dividers. The task is to visualize how the presented dividers slice the full objects and then choose all these pieces from a series of alternatives. Each correctly chosen piece is scored. We used spatial rotation to have another measure of the ability to mentally rotate objects. Participants see different three-dimensional objects. Besides one original figure, six additional figures are shown and the participant's task is to choose which of the figures represents the original figure when rotated or moved. The score is the number of correct responses. Unfold is another measure of visualization in which participants see a three-dimensional object and a series of folding templates. They then have to visualize the template that forms the original three-dimensional object. The score is the number of correct responses.

Processing speed (Gs)

Processing speed was measured with Subtests 6, 7, 8, and 10 of the LPS-2 (spatial relation, visualization, perceptual speed, and scan/search; see **Figure 3**). All scales measure the ability to quickly and accurately perceive visual details, similarities, and differences. Spatial relation was measured with Subtest 6 in which participants have to search for the one mirror-inverted number or letter in a list. Several signs can be rotated, but only one sign is mirrored and has to be marked. The scale consists of 40 trials. Scored are the correct responses reached within 2 min. We measured visualization, the ability to visualize a three-dimensional object, with Subtest 7. The participants' task is to determine the number of surfaces of a given geometrical figure. To do this, they need to visualize the figure in a three-dimensional space by counting the number of sides of the given object and indicating the number of sides by clicking on the corresponding number. There are 40 trials. The score is determined by counting the number of correct responses reached within 3 min. In subtest 8, perceptual speed, the participants' task is to recognize one out of five shapes embedded in a more complex pattern. The scale contains 40 patterns of increasing complexity. The score is the number of correct responses reached within 2 min. In subtest 10, scan and search, participants have to compare two lists of characters shown next to each other and mark characters that are different in the second list. Whereas, some rows are identical, others can differ in more than one character. The score is the number of correct markings within 2 min.

Fluid intelligence

The Raven Standard Progressive Matrices Plus (SPM) is a language-independent test of fluid intelligence. Participants see

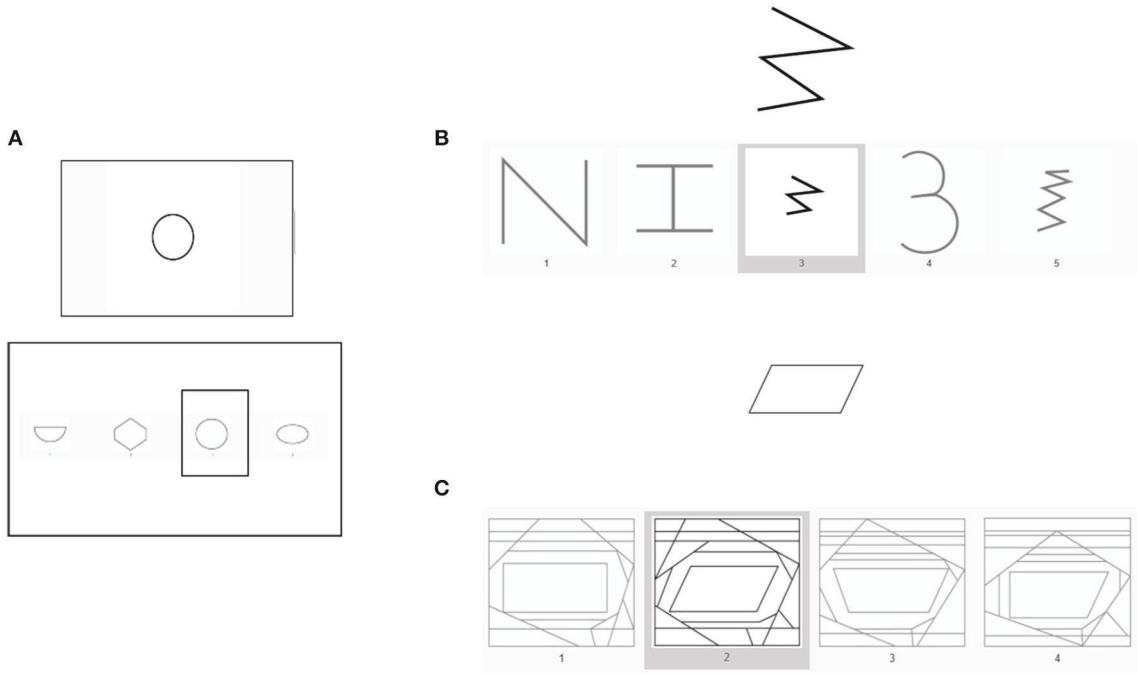

FIGURE 1 | Image example of the three scales of TVPS-3: **(A)** visual memory, **(B)** form constancy, and **(C)** figure–ground segregation.

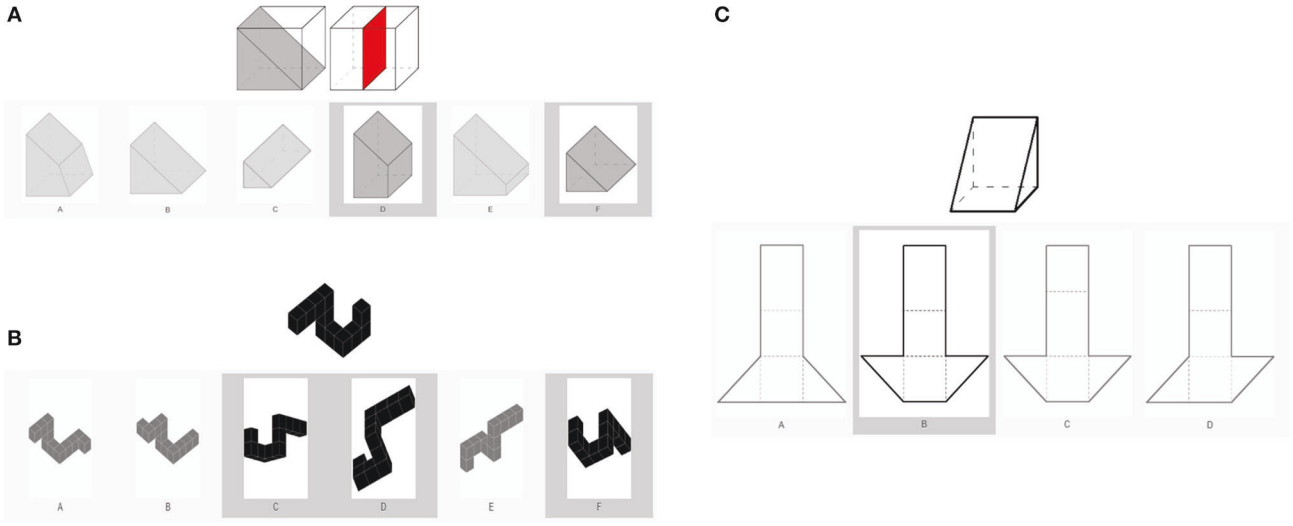

FIGURE 2 | Image example of the three scales from the WSI: **(A)** slicing, **(B)** spatial rotation, and **(C)** unfold.

a matrix of logical patterns and have to choose the missing piece out of six to eight abstract figures (Raven et al., 2003). The tests consists of 48 items of increasing complexity. The score is the number of correct responses reached within 10 min.

Simulated Baggage Screening Task

The simulated baggage screening task (SBST) was created based on the X-Ray Object Recognition Test (X-Ray ORT, Schwaninger et al., 2005; Hardmeier et al., 2006). The original ORT was designed to measure how well professional and non-professional searchers can cope with image-based factors that impact on the detection of prohibited items (viewpoint, superposition, and bag complexity) rather than measuring knowledge-based determinants of threat detection performance (which is largely dependent on training). To this end, guns and knives are used in the ORT, that is, object shapes that can be assumed to be known by most people. All X-ray images are in black and white, because colors mainly diagnose the material of the objects in the bag, and thus, could primarily help experts. In addition, all guns and knives are shown for 10 s before the test starts,

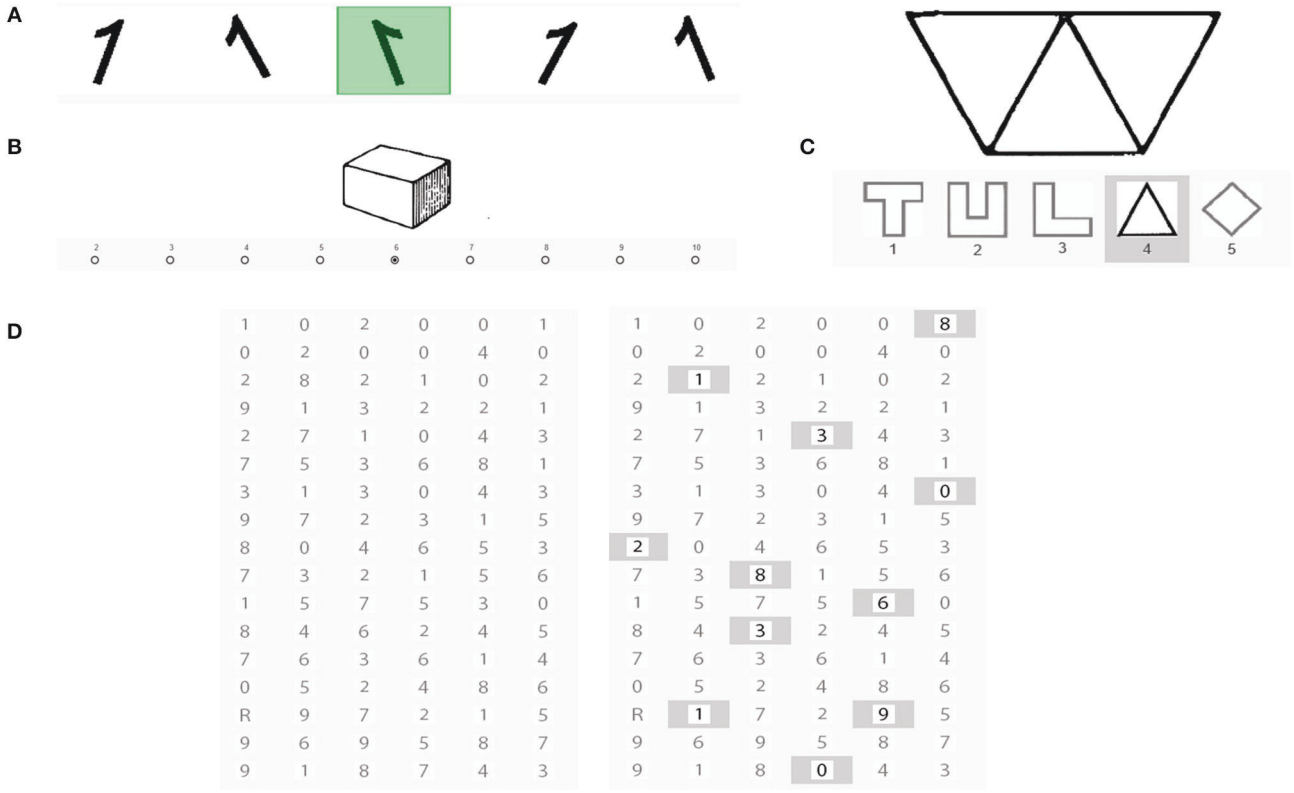

FIGURE 3 | Image example of Subtests 6, 7, 8, and 10 of the LPS-2: **(A)** spatial relation, **(B)** visualization, **(C)** perceptual speed, and **(D)** scan/search.

thereby further reducing the role of knowledge-based factors in this test.

The SBST created for this experiment included 256 X-ray images, with one half of the images containing threat item. As threats, eight guns and eight knives with common shapes were used. The X-ray images used in the SBST vary systematically in image difficulty by varying the degree of view difficulty, bag complexity, and superposition, both independently, and in combination (see **Figure 4** for examples). Therefore, each gun and each knife was displayed in an easy view and a rotated view to measure the effect of viewpoint. Each view was combined with two bags of low complexity: once with low superposition, and once with high superposition. These combinations were also generated using two close-packed bags with a higher degree of bag complexity. In addition, each bag was presented once with and once without a threat item. Thus, there were a total of 256 trials: 2 weapons (guns, knives) × 8 (exemplars) × 2 (views) × 2 (bag complexities) × 2 (superpositions) × 2 (harmless vs. threat images). The test was divided into four blocks of 64 trials each. The order of blocks was counterbalanced across four groups of participants using a Latin square. Within each block, the order of trials was random.

L/T-Letter Search Task

Comparable to previous research using laboratory visual search tasks, we created an L/T-letter search task to evaluate visual search abilities that are independent of a specific domain. In line with Biggs et al. (2013), we created a test with an increasing difficulty level and a search and decision component. The test consisted of 96 trials. Each image comprised 25 pseudo-Ls as distractors, and one-half of the images contained one target T against a gray background (see **Figure 5** as an example). Items were randomly located in a 8 × 7 grid. Each item comprised two perpendicular black lines that varied on six levels of transparency (70, 67, 65, 40, 35, and 30%) and four levels of rotation. Target Ts had a crossbar directly in the middle, whereas distractor Ls had a crossbar sliding to variable distances away from the center. The distractor stimuli varied in shape with some being very similar to the target Ts. This increased task difficulty in line with a complex conjunction search task. All items were distractors for the target-absent condition, and in the target-present condition, all items were distractors except for one target T.

Procedure

All participants were first tested with the visual-cognitive test battery (VCTB). In addition, the participants conducted a basic visual L/T- letter search task. In a second session, all participants were invited to conduct a simulated baggage screening task (SBST) using single-view X-ray images.

For the VCTB, all tests were computer-based and not conducted in the original paper-and-pencil format. Each of the 10 subtests started with general instructions followed by

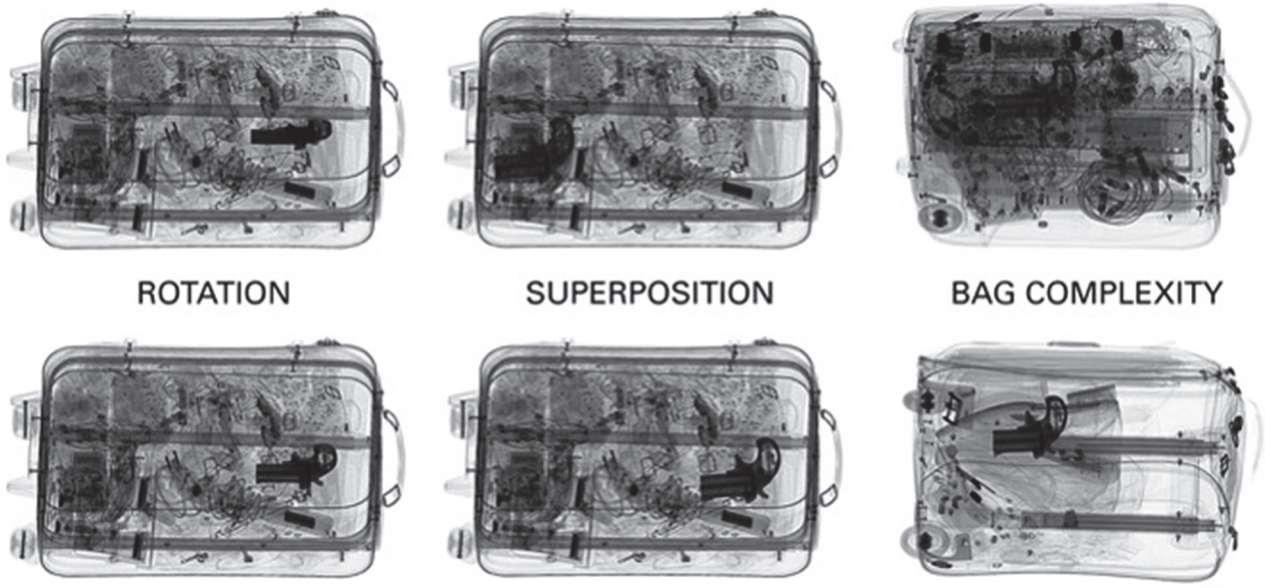

FIGURE 4 | Examples of X-ray images varying in rotation of the threat object (viewpoint effect), superposition, and bag complexity.

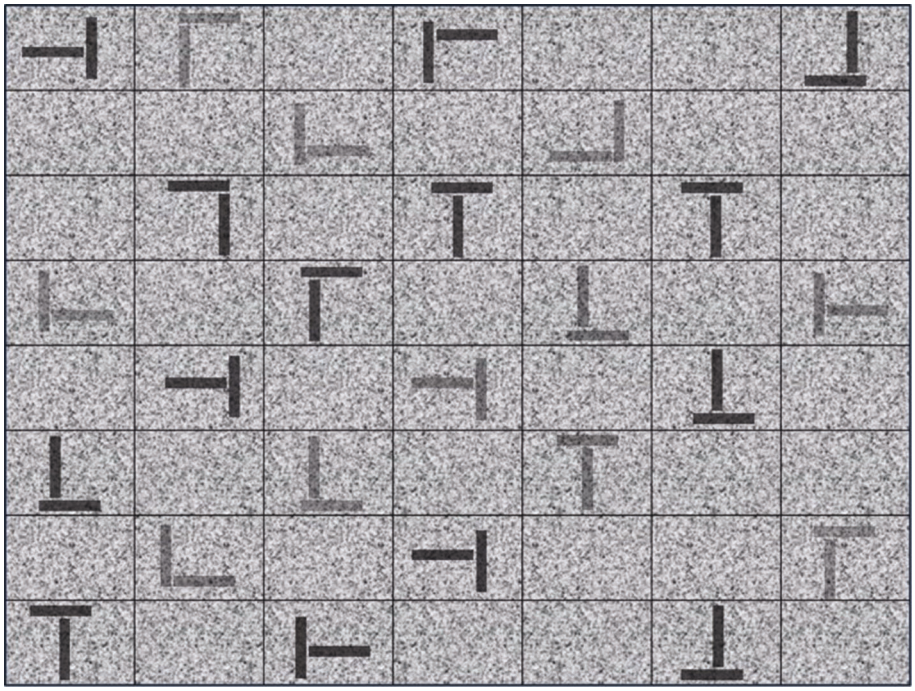

FIGURE 5 | Example of an image from the L/T-letter search task. Image containing several pseudo-Ls as distractors and one target T against a gray background.

an example. The same procedure was applied to the SPM following the VCTB scales. The test was divided into three blocks and participants were asked to take a break of 10–15 min between blocks. For the SBST, participants came to the testing facilities again, approximately 2 weeks later. Each participant sat approximately 50 cm away from the monitor. The X-ray images covered about two-thirds of the screen. After task instructions, an introductory session followed using two guns and two knives not displayed in the test phase. In each trial, an X-ray image of a piece of luggage was presented for a maximum of 4 s. We chose this duration to match the demands of high passenger flow in which average X-ray image inspection time at checkpoints is in the range of 3–5 s. The participants' task was to decide as accurately and as quickly as possible whether the bag was OK

TABLE 3 | Definition of hit, false alarm, miss, and correct rejection according to SDT (Green and Swets, 1966).

Stimulus	Target-present response	Target-absent response
Target-present stimulus	Hit	Miss
Target-absent stimulus	False alarm	Correct rejection

(no threat item) or NOT OK (a gun or knife present) by clicking on the respective button. Prior to the actual test phase, the eight guns and eight knives used in the test were each presented for 10 s. Feedback was provided after each trial, but only in the introductory phase. For the L/T- letter search task, the same computers and monitors were used as for the SBST. Again, participants sat approximately 50 cm away from the monitor and the images covered about two-thirds of the screen. Each trial started with a fixation cross in the middle of the screen. After 0.5 s, a grid with 25 stimuli was presented for a maximum of 15 s. Each grid had 0 or 1 T's. If participants recognized a target T, they had to press "Y" on the keyboard and then mark the target T with the mouse. If they did not see a target T, they had to press "space" on the keyboard. As soon as participants marked the target T with the mouse or pressed the spacebar, the next trial started. If there was no decision after 15 s, the next trial started.

Analyses

Both tasks used in this experiment can be described as a visual inspection consisting of visual search and decision (Spitz and Drury, 1978; Koller et al., 2009; Wales et al., 2009). The outcome of this task is based on the searchers decisions on whether a target is present or absent. According to signal detection theory (SDT) (Green and Swets, 1966), there are four possible outcomes depending on stimuli and participant responses (**Table 3**). Because individuals with identical detection ability can have different levels of hit rate and false alarm rate due to different response tendencies, it is often more appropriate to express detection performance in terms of a sensitivity measure (Green and Swets, 1966; Macmillan and Creelman, 2005). We therefore used d' as detection measure for the L/T-letter search task based on the following formula in which z refers to the inverse of the cumulative distribution function of the standard normal distribution (Green and Swets, 1966; Macmillan and Creelman, 2005):

$$d' = z\,(\mathrm{HR}) - z(\mathrm{FAR}) \tag{1}$$

d' is based on the equal variance Gaussian model, a common model of SDT (Pastore et al., 2003). SDT can also assume other underlying evidence distributions. One example is a SDT model that assumes the two evidence distributions to be normal but with unequal variance. For a given ratio s between the standard deviation of the target-present and target-absent distribution, the resulting zROC has slope s. For this SDT model, Macmillan and Creelman (2005) propose using Simpson and Fitter's (1973) detection measure:

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

Concerning the task of X-ray screening, several studies have raised doubts about the equal variance Gaussian model. Wolfe et al. (2007) proposes a zROC slope of 0.6, which indicates that the noise (target-absent) distribution has a smaller standard deviation than the signal-plus-noise (target-present) distribution. Further publications (Van Wert et al., 2009; Godwin et al., 2010) have reported zROC slopes similar to those reported by Wolfe et al. (2007) while a study reported by Wolfe and Van Wert (2010) found a slope of 0.56 and a study by Sterchi et al. (2019) a slope of 0.5 to fit the data more accurately. In our study, data from the basic visual search task (L/T-letter search task) were analyzed under the assumption of an equal variance model using d', whereas data from the X-ray image inspection task SBST were analyzed under the assumption of an unequal variance model with a zROC slope of 0.5 using d_a[1].

In a first step, we examined descriptive statistics (means and standard deviations) as well as correlations (Spearman correlations; Spearman, 1927) with basic functions of R Statistics version 3.4.4 (R Core Team, 2018). We then performed confirmatory factor analysis (CFA) using maximum likelihood methods of estimation with the package "lavaan" (Rosseel, 2012) in R Statistics version 3.4.4 (R Core Team, 2018). We report factor loadings of CFA, which should be minimally 0.50 and optimally higher than 0.70. To estimate the goodness of fit for the models, we report Chi^2 values, the comparative fit index (CFI), the Tucker–Lewis index (TLI), and the root-mean-square error of approximation (RMSEA). CFI and TLI values close to 0.95 or higher (Hu and Bentler, 1999) and RMSEA values up to 0.07 (Steiger, 2007) indicate a good fit between the data and the proposed model. For the multiple regression analyses, predictors were entered into the regression using the "enter method" (forced entry). For results, we report R^2, F, and p to evaluate the overall model fit. Furthermore, we report β, SE, t, and p for each predictor. In order to compare regression models, we used Wald's test and the Bayes factor. Bayes factor was calculated with the package "BayesFactor" (Morey et al., 2018) in R Statistics version 3.4.4 (R Core Team, 2018). The interpretation of the Bayes factor as evidence for the alternative hypothesis was reported in line with Raftery (1995).

RESULTS

We first report descriptive statistics and Spearman correlations. In accordance with the CHC model of intelligence (e.g., Flanagan and Dixon, 2013), we then computed a CFA over the VCTB scales with three latent factors: *visual processing* (*Gv*), *short-term memory* (*Gsm*), and *perceptual speed* (*Gs*) in order to confirm the construct validity of the used VCTB. Further, we performed

[1]The choice between d_a and d' would be a concern if there was systematic variance in the criterion. Although we did not expect this in our study, we recalculated the data using d' and found no relevant differences in the results; that is, all significant effects remained significant.

TABLE 4 | Means and standard deviations.

		Students				Professionals			
	Max. score	*n*	*M*	*SD*	Cronbach's α	*n*	*M*	*SD*	Cronbach's α
d_a SBST	3.5	128	1.6	0.3	0.83	112	2.6	0.4	0.80
RT SBST	4.0	128	3.2	1.1		112	2.6	0.7	
d' L/T	3.5	128	1.0	0.5	0.71	112	1.0	0.5	0.70
RT L/T	15.0	128	8.1	1.3		112	8.2	11.4	
Gs	116	128	80.9	13.7	0.86–0.95	112	64.2	16.6	0.89–0.94
Gv	48	128	37.4	5.1	0.20–0.65	112	36.3	6.2	0.45–0.77
Gsm	31	128	21.7	5.5	0.56–0.75	112	19.8	5.7	0.62–0.68

n, number of participants; M, mean; SD, standard deviation; Cronbach's α, internal consistency of scale; SBST, simulated baggage screening task; L/T-letter search task; Gs, processing speed; Gv, visual processing; Gsm, visual memory; statistical abbreviations: d_a and d′, detection performance measures; RT, reaction time in seconds.

TABLE 5 | Correlational analyses.

	d_a SBST	RT SBST	d′ L/T	RT L/T	SPM	Gs	Gsm	Gv
STUDENTS								
d_a SBST	–							
RT SBST	0.20*	–						
d' L/T	0.34***	0.08	–					
RT L/T	0.23**	0.26**	0.45***	–				
SPM	0.28**	0.03	0.24**	0.20*	–			
Gs	0.22*	0.07	0.16	−0.03	0.57***	–		
Gsm	0.46***	0.23**	0.32***	0.25**	0.47***	0.33***	–	
Gv	0.40***	0.25**	0.35***	0.30**	0.37***	0.30**	0.64***	–
Age	0.19*	0.06	0.14	0.16	−0.03	−0.14	0.13	0.11
PROFESSIONALS								
d_a SBST	–							
RT SBST	0.18	–						
d' L/T	0.35***	0.02	–					
RT L/T	0.23*	0.09	0.39***	–				
SPM	0.25**	−0.02	0.33***	0.21*	–			
Gs	0.11	−0.17	0.26**	0.02	0.61***	–		
Gsm	0.24*	0.07	0.28**	0.16	0.60***	0.43***	–	
Gv	0.39***	0.16	0.38***	0.34***	0.62***	0.43***	0.58***	–
Age	−0.05	0.48***	−0.03	−0.05	−0.19*	−0.36***	−0.15	−0.11

*Spearman Correlations. *$p < 0.05$. **$p < 0.01$. and ***$p < 0.001$.*

multiple regression analyses to test whether the *z*-standardized summarized scale scores of *Gv*, *Gms*, and *Gs* could predict performance in the traditional L/T-letter search task and the X-ray image inspection task (SBST). Last, we tested whether the performance of the L/T-letter search task could mediate the effects of *Gv*, *Gms*, and *Gs* on the performance of the SBST.

Descriptive Statistics and Correlations

Table 4 shows means and standard deviations of all independent (*Gs*, *Gv*, *Gsm*) and dependent variables (d_a SBST, RT SBST, d' L/T, RT L/T) for students and professionals. **Table 5** reports the Spearman correlations between all variables separately for students and professionals. Correlations with SPM scores served as a control and showed high significance with all the VCTB scales and a significant relationship with performance in both tasks. Correlations among the detection performance of the L/T-letter search task and SBST with the VCTB measures *Gv* and *Gsm* were all statistically significant within both populations. *Gs* correlated with detection performance of the L/T-letter search task for professionals and with the X-ray image inspection task for students. The intercorrelations of the VCTB scales were mostly in a medium range. We also correlated age as a control variable with both tasks as well as the VCTB scales. Within the population of professionals, we found negative correlations between age and SPM and between age and *Gs* as well as a positive correlation between age and detection performance in the SBST. These are expected results, because fluid intelligence, processing speed, and performance in SBST are known to decrease with

age. In the student population, we did not find these relations. This could be due to the lower mean and range of age in this population.

Measuring Model–Confirmatory Factor Analysis

In order to confirm the CHC-model structure of the VCTB scales, we constructed three latent factors: visual processing (*Gv*), short-term memory (*Gsm*), and perceptual speed (*Gs*). CFA showed that the theoretical model fitted the data well. All factor loadings reached statistical significance ($p < 0.001$), even though the factor loading of LPS10 was minimally under the recommended quality criterion of 0.50 (Hair et al., 2014) and the factor loading of LPS6 was clearly under 0.50. The overall model fit was good with Chi^2 (32) = 56.56, $p = 0.005$, CFI = 0.961, TLI = 0.946 and RMSEA = 0.0359. As postulated by the CHC-model, the broad abilities of Stratum II were related, but distinct constructs. The correlation between the factors *Gs* and *Gsm* ($r = 0.65$, $p < 0.001$) as well as between *Gs* and *Gv* ($r = 0.53$, $p < 0.001$) was moderate, whereas there was a strong correlation between *Gsm* and *Gv* ($r = 0.83$, $p < 0.001$). The CHC-model structure was further tested for both populations separately and showed a good fit. This was taken as confirming the construct validity of the VCTB. For further analyses, we used the summarized and standardized scale scores of *Gv*, *Gsm*, and *Gs* in order to investigate those three abilities as more heterogeneous constructs.

Multiple Linear Regression Analyses

In a next step, we calculated multiple linear regression analyses to predict detection performance on the L/T-letter search task and the SBST based on the *z*-standardized summarized scale scores of *Gv*, *Gsm*, and *Gs* and group (students vs. professionals). For predicting detection performance d' on the L/T-letter search task, we found a significant regression equation $F_{(4,235)} = 9.64$, $p < 0.001$, with an adjusted R^2 of 0.13. *zGv* was the only significant predictor of detection performance (**Table 6A**). The same analysis was calculated again with group as moderator variable. However, the moderation did not improve the model fit (adjusted $R^2 = 0.12$, see **Table 6B**) and the comparison of the two models using Wald's test did not reach statistical significance $F_{(3,232)} = 0.14$, $p = 0.939$. Using the Bayes Factor to compare the two models revealed strong evidence against the moderation model ($\text{BF}_{10} = 40.4$).

For predicting detection performance d_a on the SBST, we found a significant regression equation $F_{(4,235)} = 159.3$, $p < 0.001$, with an adjusted R^2 of 0.73. Group, *zGsm*, and *zGv* were significant predictors of detection performance (**Table 6A**). The same analysis was calculated again with group as moderator variable. However, the moderation did not improve the model fit (adjusted $R^2 = 0.73$, see **Table 6B**) and the comparison of the two models using Wald's test did not reach statistical significance $F_{(3,232)} = 1.83$, $p = 0.143$. Furthermore, we found strong evidence against the moderation model using the Bayes Factor ($\text{BF}_{10} = 90.9$). Because the explained variance was much higher in the SBST compared to the L/T-letter search task, we wanted to test whether this was due to the effect of group, which was only found for the SBST. When partialing out the group variable, the R^2 decreased to 0.23. To further explore the effect of group, we tested whether work experience of professionals (years: $M = 6.83$, $SD = 5.82$) could explain some variance. However, there was no significant correlation between performance in the SBST and the log-transformed work experience ($p = 0.09$) and the model fit did not improve when including work experience as an additional variable (adjusted $R^2 = 0.72$).

Up to this point, we found indication that both populations require the same visual-cognitive abilities to predict performance in both measured tasks. The regression models showed that performance on both visual search tasks was predicted by *zGv* and also *zGsm* (although only significantly for performance on SBST). Based on this result, it could be concluded that performance on, the L/T-letter search task and the SBST are predicted by the same visual-cognitive abilities. If this was the case, performance on the L/T-letter search task should fully mediate the effect of *zGv* and *zGsm* on performance in the SBST. This mediation effect would provide important information on whether results from traditional visual search tasks can be directly applied to professional X-ray image inspection. We investigated this hypothesis by conducting a mediation analysis using performance on the L/T-letter search task as mediator between the visual-cognitive abilities and performance on the SBST. We found a significant regression equation for the mediation model $F_{(5,234)} = 135.9$, $p < 0.001$, with an adjusted R^2 of 0.74. **Table 6C** shows that even though performance on the L/T-letter search task significantly predicted performance on the SBST, the direct effects of *Gv*, *Gsm*, and group still attained significance. The mediation model therefore showed that the effect of *Gv* and *Gsm* on performance of SBST was only partially mediated by performance on the L/T-letter search task. This means that L/T-letter search task performance by itself explains only part, but not all of the direct effects of *Gv* and *Gsm* on performance on the SBST, while *Gv* and *Gsm* explain an additional part of variance in performance on the SBST. To explore this result in more detail, we tested the size of the indirect effect of the visual-cognitive abilities on performance on the SBST through performance on the L/T-letter search task using bootstrapping procedures. These calculations give indication on how much variance of the total effect on performance on SBST can be explained by the effect of visual-cognitive abilities on performance on the L/T-letter search task, which in turn has an effect on performance on the SBST task. Indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5 and 97.5th percentiles. The bootstrapped indirect effects were 0.00 for *Gs* ($SD = 0.01$, 95% CI [−0.02, 0.02]); 0.01 for *Gsm* ($SD = 0.01$, 95% CI [−0.01, 0.04]); 0.04 for *Gv* ($SD = 0.02$, 95% CI [0.01, 0.08]); and −0.01 for group ($SD = 0.02$, 95% CI [−0.05, 0.03]). Thus, the indirect effects were small and not statistically significant, revealing that only a small part of the effect of *Gv* and *Gsm* on performance of the SBST was mediated by performance on the L/T-letter search task.

Since *Gs* did not show any effect on performance on the visual search tasks, we calculated the same analyses using response times (RT) as dependent variables (**Table 7**). For the L/T-letter search task, we found a significant regression equation

TABLE 6 | Multiple linear regression analyses and mediation model for detection performance.

	L/T-letter search task (d′)				SBST (d_a)			
	β	*SE*(β)	*t*-value	*p*-value	β	*SE*(β)	*t*-value	*p*-value
(A) BASIC MODEL								
zGs	−0.013	0.078	−0.164	0.870	−0.039	0.044	−0.893	0.373
zGsm	0.119	0.079	1.513	0.132	0.104	0.044	2.348	0.019*
zGv	0.299	0.078	3.830	0.000***	0.195	0.044	4.463	0.000***
zGroup	0.029	0.070	−0.416	0.678	−0.834	0.039	−21.370	0.000***
adj. R^2			0.126***				0.726***	
(B) MODERATION MODEL								
zGs	−0.018	0.079	−0.223	0.823	−0.03	0.044	−0.675	0.501
zGsm	0.127	0.080	1.567	0.119	0.113	0.045	2.533	0.012*
zGv	0.286	0.082	3.504	0.001***	0.190	0.045	4.132	0.000***
zGroup	−0.028	0.070	−0.400	0.700	−0.835	0.040	−21.458	0.000***
zGs*zGroup	−0.030	0.079	−0.378	0.705	0.064	0.044	1.461	0.145
zGsm*Group	0.036	0.080	0.451	0.652	0.064	0.045	1.426	0.155
zGv*Group	−0.034	0.080	−0.418	0.676	−0.054	0.045	−1.206	0.229
adj. R^2			0.117***				0.730***	
					β	*SE*(β)	*t*–value	*p*-value
(C) MEDIATION MODEL								
zL/T d_a					0.13	0.04	3.5	0.000***
zGs					−0.04	0.044	−0.88	0.382
zGsm					0.09	0.044	2.05	0.042*
zGv					0.16	0.044	3.58	0.000***
zGroup					−0.83	0.044	−21.77	0.000***
adj. R^2							0.740***	

*$p < 0.05$; and ***$p < 0.001$.

TABLE 7 | Multiple linear regression analyses for response times (RT).

	β	*SE β*	*t-value*	*p*-value
L/T-LETTER SEARCH TASK				
zGs	−0.209	0.077	−2.721	0.007**
zGsm	0.078	0.078	1.002	0.317
zGv	0.383	0.077	4.953	0.000***
Group	0.048	0.138	0.350	0.727
X-RAY IMAGE INSPECTION TASK SBST				
zGs	−0.149	0.076	−1.963	0.051
zGsm	0.114	0.077	1.484	0.139
zGv	0.176	0.076	2.307	0.022*
Group	−0.777	0.136	−5.699	0.000***

*$p < 0.05$; **$p < 0.01$, ***$p < 0.001$.

$F_{(2,235)} = 10.95$, $p < 0.001$, with an adjusted R^2 of 0.14. *zGs* and *zGv* were significant predictors of response times (**Table 7**). We recalculated the same analysis including group as moderator variable. However, the moderation did not improve the model fit (adjusted $R^2 = 0.14$) and the comparison of the two models using Wald's test did not reach statistical significance $F_{(3,232)} = 0.26$, $p = 0.85$. Using the Bayes Factor for model comparison, results suggested strong evidence against the moderation model ($BF_{10} = 37.46$). For the SBST, the regression equation was also significant $F_{(4,235)} = 12.74$, $p < 0.001$, with an adjusted R^2 of 0.16. Group and *zGv* were significant predictors of response times (**Table 7**). Using group as moderator variable slightly improved the model fit (adjusted $R^2 = 0.18$), however, the comparison of the two models using Wald's test did not reach statistical significance $F_{(3,232)} = 2.37$, $p = 0.07$. Using the Bayes factor for model comparison, results suggested only weak evidence against the moderation model ($BF_{10} = 2.40$). Again, to further explore the effect of group, we entered work experience as an additional variable, but this did not improve the model fit (adjusted $R^2 = 0.18$).

DISCUSSION

Many studies on the topic of visual search have been conducted with students using traditional, simplified visual search tasks and salient stimuli. Although such research is vital to explore the underlying cognitive mechanisms in a controlled environment, it is not always clear whether the results extrapolate to real-world inspection in which professionals search their visual

fields for targets that are more complex, ambiguous, and less salient (e.g., Radvansky and Ashcraft, 2016, p. 257). Furthermore, visual search research is often conducted with students, who differ systematically from professional searchers. We investigated whether the same visual cognitive abilities predict performance in students and professionals performing two tasks: a traditional visual search task—the L/T-letter search task—and an X-ray image inspection task. We tested students and professionals on three known facets of visual-cognitive abilities: visual processing (*Gv*), short-term memory (*Gsm*), and processing speed (*Gs*). We shall now use our results to answer the following research questions: (1) Do different visual-cognitive abilities predict performance and response times in a traditional visual search task and an X-ray image inspection task? (2) Do the results differ between students and professionals?

Our results show that visual search ability as measured with a traditional visual search task involves different underlying visual-cognitive processes compared to an applied X-ray image inspection task. Whereas, visual search ability as measured with the L/T-letter search task was significantly predicted by visual processing (*Gv*), performance on the SBST was significantly predicted by visual processing (*Gv*) and short-term memory (*Gsm*). However, the mediation model revealed that only a small part of the effect of *Gv* and *Gsm* on performance of the SBST was mediated by performance on the L/T-letter search task. This leads to the conclusion that different aspects of *Gv* and *Gsm* predict performance in the measured tasks. Furthermore, the influence of the measured visual-cognitive abilities on performance did not differ between students and professional screeners. However, professionals outperformed students in the X-ray image inspection task.

Traditional Visual Search vs. X-Ray Image Inspection

Multiple linear regression analyses were calculated for both visual search tasks in order to predict performance based on three visual-cognitive abilities (*Gv*, *Gsm*, *Gs*) and group (students vs. professionals). We further added the L/T-letter search task as a mediator of the effects of the visual-cognitive abilities to the model. The L/T-letter search task should reduce the direct effect of the visual-cognitive abilities on the X-ray image interpretation test if the two tasks depend on the same aspects of these abilities. However, the mediation model showed that only a small amount of the effects from the visual-cognitive abilities on X-ray image interpretation performance was mediated through the L/T-letter search performance. That different visual-cognitive abilities are relevant for the two tasks, is therefore indicated by the different underlying cognitive processes.

In the regression model, visual processing (*Gv*) was a predictor of performance for both tasks. This result is in accordance with earlier studies showing a correlation between performance and visual processing for traditional visual search (Wolfe et al., 2002; Bolfing and Schwaninger, 2009) and an influence of mental rotation and figure-ground segregation on higher performance in X-ray screening (Wolfe et al., 2002; Bolfing and Schwaninger, 2009), which are narrow abilities of visual processing (*Gv*). However, our results showed that different aspects of visual processing explain variance in the traditional visual search task and the X-ray image inspection task. According to the CHC theory, visual processing describes a broader ability to perceive, analyze, synthesize, and think with visual patterns, including the ability to store and recall visual representations. Both, the L/T-letter search task and the X-ray image inspection task require visual processing abilities, that is, the ability to mentally rotate objects and see them in their spatial relation and the ability to visualize and recognize patterns (e.g., visual memory, figure–ground segregation, or form constancy). However, visual processing includes a broad spectrum of abilities. Even though the traditional visual search task and X-ray image inspection task in this study were created to make them comparable, the tasks differed in regard to stimuli and distractor complexity. Targets in the traditional visual search task (*Ls* and *Ts*) have salient shapes, whereas targets (guns and knives) and distractors in the X-ray image inspection task are not salient and may additionally produce clutter and superposition. These are all potential reasons for our finding that different aspects of *Gv* are needed to perform faster and better in the measured tasks.

Short term memory (*Gsm*) was a significant predictor of X-ray image inspection performance, but not for the traditional visual search task. However, even though the standardized coefficient for *Gsm* was not smaller for the L/T-letter search task, it did not reach significance as a predictor for the L/T-letter search task (due to larger standard errors) and its relevance for that task is therefore unclear. *Gsm* is characterized as the ability to apprehend and hold information in immediate awareness and then use it within a few seconds. When comparing the stimulus complexity of the L/T-letter search task and the X-ray image inspection task, one would assume that *Gsm* might be especially important for a real-world task such as the SBST, which uses more complex and realistic stimuli and needs more top-down processing and the use of memory capacity, whereas simple letters are easy to remember. It can be further assumed that short-term memory becomes even more important when predicting performance in tasks with increasing complexity and unknown features that need previous knowledge. Regarding the X-ray image inspection task, the differentiation of targets from distractors needs memory capacity, because distractors appear in the form of everyday objects that can look similar to target items (Hättenschwiler et al., 2015; Sterchi et al., 2017), and prior object knowledge is needed to differentiate targets from non-targets.

Processing speed, the ability to quickly and accurately perceive visual details, similarities, and differences, did not predict detection performance in the measured tasks. We therefore additionally calculated a model for response times, in which processing speed predicted performance in the L/T-letter search task but fell short of significance for the X-ray image inspection task (significance in the SBST: $p = 0.051$). Participants with higher *Gs* scores therefore performed faster. This result is consistent with previous research that found processing speed to be relevant in terms of efficiency (Salthouse, 1996).

Comparison of Students and Professionals

For both groups, visual-cognitive abilities were comparably relevant for their performance on the traditional visual search task and the X-ray image inspection task. However, professionals outperformed students on the X-ray image inspection task. Because the relevance of the visual-cognitive abilities tested in this study proved to be independent of the population and they had similar levels of visual-cognitive abilities, the higher detection performance of the professionals in the SBST cannot be explained by differences in visual-cognitive abilities. Consistent with this interpretation, after removing the group variable from the analyses in the X-ray image inspection task, a similar amount of variance could be explained as in the L/T-letter search task (especially when considering that the SBST was more reliable). This leaves mainly two possible explanations for this difference: Students and professionals might differ in other cognitive abilities than the ones measured, and these other abilities account for the improved detection performance only on the SBST but not the L/T-letter search task. Such a difference could be due to the selection of the security personnel. Or more likely, the group effect could be due to differences related to training and job experience of the professionals.

Halbherr et al. (2013) found that the biggest increase in performance is seen incrementally up to 40 h of training. The professionals participating in this study all had more than 2 years of training and work experience. Additional training hours might therefore not result in a large performance increase. This is consistent with our finding that partialling out age and work experience did not improve the model fit. McCarley et al. (2004) found detection performance improvements to be based on improvements in object recognition rather than the visual search task *per se*. Based on that, more familiar objects possibly need fewer recognized features in order to be identified successfully (Koller et al., 2009), and features are known and recognized better and faster with repeated exposure (McCarley et al., 2004; Schwaninger and Hofer, 2004; Koller et al., 2008, 2009; Halbherr et al., 2013). In our study, we created a traditional visual search task with a higher difficulty level and an X-ray image inspection task containing targets with no need of domain-specific knowledge. Features of guns and knives as well as letters such as *L* or *T* are known from everyday life and can therefore be detected without specific experience and training. However, the X-ray screening task requires the ability to resolve object occlusion, whereas the L/T-letter search task does not. Therefore, inferring the full shape of occluded objects may be superior in professionals due to higher object familiarity. It can further be assumed that work experience leads to richer object templates or representations of everyday objects in X-ray images (Hättenschwiler et al., 2015). As discussed above, distractors in an X-ray image inspection task are merely everyday objects that can look like threat items, especially if no target representation is stored. In comparison to a traditional L/T-letter search task in which distractors are salient and known, many everyday object distractors cannot be recognized easily in X-ray images without prior knowledge. This lack of knowledge can be a disadvantage for students who are not used to X-ray images and might lead them to incorrectly judge a bag to be harmful (Sterchi et al., 2017).

Regarding response times, the visual-cognitive abilities were comparably relevant for both groups in the traditional visual search task and the X-ray image inspection task. Using group as moderator variable only resulted in a small and not quite significant increase of the model fit. We, however, believe that this difference in R^2 is too small to indicate a relevant moderation. Also the Bayes factor provides weak evidence against the moderation model. Therefore, differences between groups as discussed above only seem to be relevant for detection performance and not response times.

Taken together, the influence of the measured visual-cognitive abilities on performance did not differ between students and professional screeners. However, professionals outperformed students in the X-ray image inspection task, which we assume to be due to training and job experience of the professionals. The presence of a group difference, but apparent absence of a moderation suggests that experience (or any alternative reason for the group difference) does not interact with the relevance of the visual-cognitive abilities for the X-ray image inspection task. However, we would caution against assuming that this pattern can be generalized to other visual-cognitive abilities or other implementations of the X-ray image inspection task. The X-ray image inspection task as used in this study is not the same task as the one screeners conduct at checkpoints—particularly regarding target prevalence, coloring of images, and target categories. Prohibited items that are rather uncommon or have not been seen before (e.g., improvised explosive devices, IEDs) become very difficult to detect without training in the recognition of certain features of these threats (Schwaninger, 2004, 2005). Assuming that the performance in detecting such threats is still dependent on certain visual-cognitive abilities and that only professionals can detect them, these visual-cognitive abilities would only be relevant for the performance of professionals. We therefore expect that results would look different if a task was used that requires domain-specific knowledge.

LIMITATIONS AND FUTURE DIRECTIONS

One limitation of this study is the representativeness of the tested populations. Our samples of students and professionals showed similar means and standard deviations on the measured visual-cognitive abilities. Professionals participating in this study all passed a preemployment test for these visual abilities (e.g., X-Ray Object Recognition Test; see Hardmeier et al., 2005; Hardmeier and Schwaninger, 2008). It could therefore be possible that they have high levels of certain other relevant visual-cognitive abilites that were not included in this study. Future studies could investigate applicants for the screening job and investigate how far preemployment assessment limits variation in visual-cognitive abilities. It would further be interesting to observe whether the influence of the visual-cognitive abilities really remains stable when the screeners' performance increases through training and job experience. Further, the students tested in our study proved to be a very heterogeneous sample, especially

with a high variance in age, which is not directly comparable to a typical student sample (students from universities of applied sciences tend to be more heterogeneous than students at other universities). This raises the question whether regression results would be affected if the tested sample were more homogeneous on some variables.

Our results suggest that different aspects of *Gv* and *Gsm* are relevant for performance on the L/T-letter search task and X-ray image inspection. Future studies should investigate the influence of narrow (Stratum I) abilities on these tasks. Implications based on current results could be that either a simple and short version of the visual-cognitive test battery (*Gv* scales) could be used to measure abilities and predict performance in students and professionals. Or in an applied setting, the SBST could be used as a criterion for abilities. Because there are major individual differences in visual-cognitive abilities, it should be tested whether someone is suited to perform well in a visual search and inspection task. Especially with regard to X-ray screening, airports could conduct preemployment assessments that test for certain visual abilities and aptitudes when recruiting new personnel. However, visual-cognitive abilities might become less important as performance predictors for tasks in which domain-specific knowledge is not only helpful but necessary. For example, when radiologists search for cancer in mammograms or screeners search for improvised explosive devices that include unknown features, training for these features should have a stronger influence on performance than visual-cognitive abilities. Future studies could also investigate whether visual-cognitive abilities change over time, and whether these abilities could be trained through repeated exposure to visual search tasks.

CONCLUSION

With this study, we tried to determine how far results on a traditional visual search task can be translated to an X-ray image inspection and vice versa, and whether populations of students and professionals are comparable. Comparing visual-cognitive abilities and their influence on performance revealed that the different visual-cognitive abilities were able to predict performance on the measured tasks. The CHC proved to be a good model for mapping the visual-cognitive abilities needed to conduct a visual search task. Our mediation analyses revealed that the used tasks are not comparable *per se* as there was only a partial overlap between the required aspects of visual-cognitive abilities. Furthermore, although our tested populations were comparable in terms of performance predictors based on visual-cognitive abilities, professionals outperformed students on an applied X-ray image inspection task, suggesting that the performance is not solely predictable by visual-cognitive abilities. The implications of our second research question therefore have to be treated with caution, because the comparability of the two populations is dependent on the task. One should therefore be cautious about translating results from the L/T-letter search task to X-ray image inspection.

AUTHOR CONTRIBUTIONS

All authors substantially contributed to the conceptualization of the manuscript as well as to the acquisition, analysis, and interpretation of data. All authors critically revised the content of the manuscript repeatedly and approved the final version to be published. All authors agreed to be accountable for all aspects of the work. NH and SM as the leading authors contributed to the development of the tests, the acquisition, analysis, and interpretation of data. NH was responsible for the conceptualization and the writing of the manuscript. YS predominantly contributed to the acquisition, analyses, and interpretation of data. NH, SM, and YS repeatedly revised and refined the content of the manuscript critically. AS predominantly contributed to the development of the tests, the analyses and interpretation of data. AS repeatedly revised and refined the content of the manuscript critically.

ACKNOWLEDGMENTS

The authors thank Myrta Isenschmid and Vivienne Kunz for their valuable help during the recruitment of participants and the data collection.

REFERENCES

Alexander, R., and Zelinsky, G. (2011). Visual similarity effects in categorical search. *J. Vis.* 11:9. doi: 10.1167/11.8.9

Alexander, R., and Zelinsky, G. (2012). Effects of part-based similarity on visual search: the frankenbear experiment. *Vis. Res.* 54, 20–30. doi: 10.1016/j.visres.2011.12.004

Alvarez, G. A., and Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol. Sci.* 15, 106–111. doi: 10.1111/j.0963-7214.2004.01502006.x

Biggs, A. T., Cain, M. S., Clark, K., Darling, E. F., and Mitroff, S. R. (2013). Assessing visual search performance differences between transportation security administration officers and nonprofessional visual searchers. *Vis. Cogn.* 21, 330–352. doi: 10.1080/13506285.2013.790329

Biggs, A. T., and 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. Cogn. Psychol.* 29, 142–148. doi: 10.1002/acp.3083

Bolfing, A., and Schwaninger, A. (2009). "Selection and pre-employment assessment in aviation security x-ray screening," in *Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology* (Zurich). doi: 10.1109/CCST.2009.5335571

Bravo, M. J., and Farid, H. (2004). Recognizing and segmenting objects in clutter. *Vis. Res.* 4, 385–396. doi: 10.1016/j.visres.2003.09.031

Brown, T., Sutton, E., Burgess, D., Elliott, S., Bourne, R., Wigg, S., et al. (2010). The reliability of three visual perception tests used to assess adults. *Percept. Motor Skills* 111, 45–59. doi: 10.2466/03.24.27.PMS.111.4.45-59

Carrasco, M. (2011). Visual attention: the past 25 years. *Vis. Res.* 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Carrasco, M. (2014). "Spatial attention: perceptual modulation," in *The Oxford Handbook of Attention*, eds S. Kastner and A. C. Nobre (Oxford: Oxford University Press), 183–230.

Carrasco, M. (2018). How visual spatial attention alters perception. *Cogn. Proc.* 19(Suppl. 1), 77–88. doi: 10.1007/s10339-018-0883-4

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511571312

Carroll, J. B. (2003). "The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors," in *The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen*, ed. H. Nyborg (San Diego, CA: Pergamon), 5–22.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychol. Bull.* 38:592.

Chan, L. K. H., and Hayward, W. G. (2013). Visual search. *WIREs Interdiscipl. Rev.* 4, 415–429. doi: 10.1002/wcs.1235

Clark, K., Cain, M. S., Adamo, S. H., and Mitroff, S. R. (2012). "Overcoming hurdles in translating visual search research between the lab and the field," in *The Influence of Attention, Learning, and Motivation on Visual Search*, eds. M. D. Dodd and J. H. Flowers (New York, NY: Springer), 147–181.

Drury, C. G. (1975). Inspection of sheet materials–model and data. *Hum. Fact.* 17, 257–265.

Duncan, J., and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol. Rev.* 96, 433–458. doi: 10.1037/0033-295X.96.3.433

Eckstein, M. (2011). Visual search: a retrospective. *J. Vis.* 11, 1–36. doi: 10.1167/11.5.14

Eriksen, C. W., and Schultz, D. W. (1979). Information processing in visual search: a continuous flow conception and experimental results. *Percep. Psychophys.* 25, 249–263. doi: 10.3758/BF03198804

European Commission (2015). *Commission Implementing Regulation (EU) 2015/1998*. Available online at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R1998andfrom=DE

Flanagan, D. P., and Dixon, S. G. (2013). "The Cattell–Horn–Carroll theory of cognitive abilities," in *Encyclopedia of Special Education*, eds C. R. Reynolds, K. J. Vannest, and E. Fletcher-Janzen (Hoboken, NJ: John Wiley & Sons), 368–382.

Flanagan, D. P., and Harrison, P. L. (2005). *Contemporary Intellectual Assessment: Theories, Tests, and Issues, 2nd Edn.* New York, NY: The Guilford Press.

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

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

Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 1 Edn*. Thousand Oaks, CA: Sage.

Halbherr, T., Schwaninger, A., Budgell, G., and Wales, A. (2013). Airport security screener competency: a cross-sectional and longitudinal analysis. *Int. J. Aviat. Psychol.* 23, 113–129. doi: 10.1080/10508414.2011.582455

Hardmeier, D., Hofer, F., and 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," *Proceedings of the 39th IEEE International Carnahan Conference on Security Technology* (Las Palmas), 189–192.

Hardmeier, D., Hofer, F., and Schwaninger, A. (2006). Increased detection performance in airport security screening using the X-Ray ORT as pre-employment assessment tool," in *Proceedings of the 2nd International Conference on Research in Air Transportation. ICRAT 2006* (Belgrade), 393–397.

Hardmeier, D., and Schwaninger, A. (2008). "Visual cognition abilities in x-ray screening," in *Proceedings of the 3rd International Conference on Research in Air Transportation*. ICRAT 2008 (Virginia: Fairfax), 311–316.

Hättenschwiler, N., Michel, S., Ritzmann, S., and Schwaninger, A. (2015). "A first exploratory study on the relevance of everyday object knowledge and training for increasing efficiency in airport security X-ray screening," in *Proceedings of the 49th IEEE International Carnahan Conference on Security Technology* (Taipei), 25–30.

Hell, B., Päßler, K., and Schuler, H. (2009). Was-studiere-ich. de: Konzept, Nutzen und Anwendungsmöglichkeiten. *Zeitsc. Stud. Berat.* 4, 9–14.

Horn, J. L. (1965). *Fluid and Crystallized Intelligence: A Factor Analytic and Developmental Study of the Structure Among Primary Mental Abilities*. Unpublished doctoral dissertation, University of Illinois, Champaign.

Horn, R. (2009). "Standard progressive matrices (SPM)," in *Deutsche Bearbeitung und Normierung nach, 2nd Edn.*, ed. J. C. Raven (Frankfurt: Pearson Assessment).

Horowitz, T. S. (2017). Prevalence in visual search: from the clinic to the lab and back again. *Japanese Psychol. Res.* 59, 65–108. doi: 10.1111/jpr.12153

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Humphreys, G. W., and Mavritsaki, E. (2012). "Models of visual search: from abstract function to biological constraint," in *Cognitive Neuroscience of Attention, 2nd Edn.,* ed. M. I. Posner (New York, NY: Guilford Press), 420–456.

Keith, T. Z., and Reynolds, M. R. (2012). "Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests," in *Contemporary Intellectual Assessment: Theories, Tests, and Issues, 3rd ed*, eds. D. P. Flanagan and P. L. Harrison (New York, NY: Guilford Press), 758–799.

Koller, S., Hardmeier, D., Michel, S., and Schwaninger, A. (2008). Investigating training, transfer and viewpoint effects resulting from recurrent CBT of x-ray image interpretation. *J. Transport. Sec.* 1, 81–106. doi: 10.1007/s12198-007-0006-4

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

Kosslyn, S. M. (1975). Information representation in visual images. *Cogn. Psychol.* 7, 341–370. doi: 10.1016/0010-0285(75)90015-8

Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.

Kreuzpointner, L., Lukesch, H., and Horn, W. (2013). *Leistungsprüfsystem* 2. *LPS*-2. Göttingen: Hogrefe.

Krupinski, E. (1996). Visual scanning patterns of radiologists searching mammograms. *Acad. Radiol.* 3, 137–144. doi: 10.1016/S1076-6332(05)80381-2

Lavie, N., and DeFockert, J. (2005). The role of working memory in attentional capture. *Psychon. Bull. Rev.* 12, 669–674. doi: 10.3758/BF03196756

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide, 2nd Edn*. Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, N. A. (2006). *Test of Visual Perceptual Skills (TVPS-3), 3rd Edn*. Novato, CA: Academy Publishers.

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

McElree, B., and Carrasco, M. (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *J. Exp. Psychol.* 25, 1517–1539. doi: 10.1037/0096-1523.25.6.1517

McGrew, K. S. (2005). "The Cattell-Horn-Carroll theory of cognitive abilities: past, present, and futur," in *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, eds. D. P. Flanagan, J. L. Genshaft, and P. L. Harrison (New York, NY: Guilford), 136–182.

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

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., and Ly, A. (2018). *Package "Bayesfactor"*. Available online at: ftp://alvarestech.com/pub/plan/R/web/packages/BayesFactor/BayesFactor.pdf

Nakayama, K., and Martini, P. (2011). Situating visual search. *Vis. Res.* 51, 1526–1537. doi: 10.1016/j.visres.2010.09.003

Nodine, C. F., and Kundel, H. L. (1987). "The cognitive side of visual search in radiology," in *Eye Movements: From Physiology to Cognition*, eds. J. K. O'Regan and A. Levy-Schoen (North-Holland: Elsevier Science), 573–582.

Palmer, S., Rosch, E., and Chase, P. (1981). "Canonical perspective and the perception of 40 objects," in *Attention and Performance IX*, eds J. Long and A. Baddeley (Hillsdale, NJ: Lawrence Erlbaum), 135–151.

Päßler, K., and Hell, B. (2012). Do interests and cognitive abilities help explain college major choice equally well for women and men? *J. Career Assess.* 20, 479–496. doi: 10.1177/1069072712450009

Pastore, R. E., Crawley, E. J., Berens, M. S., and Skelly, M., A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychon. Bull. Rev.* 10, 556–569. doi: 10.3758/BF03196517

Poole, B. J., and Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: the influences

of sustained and selective attention. *Q. J. Exp. Psychol.* 62, 1430–1454. doi: 10.1080/17470210802479329

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Computer software. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Radvansky, G. A., and Ashcraft, M. H. (2016). *Cognition, 6 Edn*. New Jersey, NJ: Pearson Education, Inc.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163. doi: 10.2307/271063

Raven, J., Raven, J. C., and Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Reavis, E. A., Frank, S. M., Greenlee, M. W., and Tse, P. U. (2016). Neural correlates of context-dependent feature conjunction learning in visual search tasks. *Hum. Brain Mapp.* 37, 2319–2330. doi: 10.1002/hbm.23176

Roid, G. H. (2003a). *Stanford-Binet Intelligence Scales, 5th Edn*. Itasca, IL: Riverside Publishing.

Roid, G. H. (2003b). *Stanford-Binet Intelligence Scales, 5th Edn. Technical Manual*. Itasca, IL: Riverside Publishing.

Roper, Z. J., Cosman, J. D., and Vecera, S. P. (2013). Perceptual load corresponds with factors known to influence visual search. *J. Exp. Psychol.* 39, 1340–1351. doi: 10.1037/a0031616

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* 103, 403–428. doi: 10.1037/0033-295X.103.3.403

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviat. Sec. Int.* 2004, 31–36.

Schwaninger, A. (2005). Increasing efficiency in airport security screening. *WIT Trans. Built Environ.* 407–416. doi: 10.2495/SAFE050401

Schwaninger, A. (2006). "Airport security human factors: From the weakest to the strongest link in airport security screening," in *Proceedings of the 4th International Aviation Security Technology Symposium* (Washington, DC), 265–270.

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., and Hay, L. (2008). "The impact of image based factors and training on threat detection performance in X-ray screening," in *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008* (Fairfax, VA), 317–324.

Schwaninger, A., Hardmeier, D., and Hofer, F. (2004). "Measuring visual abilities and visual knowledge of aviation security screeners," in *Proceedings of the 38th IEEE International Carnahan Conference on Security Technology* (Albuquerque, NM: IEEE ICCST Proceedings) Vol. 38, 258–264.

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

Schwaninger, A., and Hofer, F. (2004). "Evaluation of CBT for increasing threat detection performance in X-ray screening," in *The Internet Society 2004, Advances in Learning, Commerce and Security*, eds K. Morgan and M. J. Spector (Wessex: WIT Press), 147–156.

Shen, J., Reingold, E. M., and Pomplun, M. (2003). Guidance of eye movements during conjunctive visual search: the distractor-ratio effect. *Can. J. Exp. Psychol.* 57, 76–96. doi: 10.1037/h0087415

Simpson, A. J., and Fitter, M. J. (1973). What is the best index of detectability? *Psychol. Bull.* 80, 481–488.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. New York, NY: Macmillan.

Spitz, G., and Drury, C. G. (1978). Inspection of sheet materials – test of model predictions. *Hum. Factors* 20, 521–528. doi: 10.1177/001872087802000502

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Pers. Individ. Dif.* 42, 893–898. doi: 10.1016/j.paid.2006.09.017

Sterchi, Y., Hättenschwiler, N., Michel, S., and Schwaninger, A. (2017). "Relevance of visual inspection strategy and knowledge about everyday objects for X-Ray baggage screening," in *Proceedings of the 51th IEEE International Carnahan Conference on Security Technology* (Madrid), 23–26.

Sterchi, Y., Hättenschwiler, N., and Schwaninger, A. (2019). Detection measures for visual inspection of X-ray images of passenger baggage. *Attent. Percept. Psychophys*. 1–15. doi: 10.3758/s13414-018-01654-8

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

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

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, 3rd Edn. (WAIS-3®)*. San Antonio, TX: Harcourt Assessment.

Wolfe, J. M. (1994). Guided Search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238. doi: 10.3758/BF03200774

Wolfe, J. M. (1998). What do 1,000,000 trials tell us about visual search? *Psychol. Sci.* 9, 33–39.

Wolfe, J. M., Cain, M. S., and Aizenman, A. M. (2019). Guidance and selection history in hybrid foraging visual search. *Attent. Percep. Psychophys*. doi: 10.3758/s13414-018-01649-5

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., and Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *J. Exp. Psychol.* 136, 623–638. doi: 10.1037/0096-3445.136.4.623

Wolfe, J. M., Olivia, A., Horowitz, T. S., Butcher, S. J., and Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vis. Res.* 42, 2985–3004. doi: 10.1016/S0042-6989(02)00388-7

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

Zhaoping, L., and Frith, U. (2011). A clash of bottom-up and top-down processes in visual search: the reversed letter effect revisited. *J. Exp. Psychol.* 37, 997–1006. doi: 10.1037/a0023099

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-018-01654-8

Detection measures for visual inspection of X-ray images of passenger baggage

Yanik Sterchi[1] · Nicole Hättenschwiler[1] · Adrian Schwaninger[1]

Abstract
In visual inspection tasks, such as airport security and medical screening, researchers often use the detection measures d' or A' to analyze detection performance independent of response tendency. However, recent studies that manipulated the frequency of targets (target prevalence) indicate that d_a with a slope parameter of 0.6 is more valid for such tasks than d' or A'. We investigated the validity of detection measures (d', A', and d_a) using two experiments. In the first experiment, 31 security officers completed a simulated X-ray baggage inspection task while response tendency was manipulated directly through instruction. The participants knew half of the prohibited items used in the study from training, whereas the other half were novel, thereby establishing two levels of task difficulty. The results demonstrated that for both levels, d' and A' decreased when the criterion became more liberal, whereas d_a with a slope parameter of 0.6 remained constant. Eye-tracking data indicated that manipulating response tendency affected the decision component of the inspection task rather than search errors. In the second experiment, 124 security officers completed another simulated X-ray baggage inspection task. Receiver operating characteristic (ROC) curves based on confidence ratings provided further support for d_a, and the estimated slope parameter was 0.5. Consistent with previous findings, our results imply that d' and A' are not valid measures of detection performance in X-ray image inspection. We recommend always calculating d_a with a slope parameter of 0.5 in addition to d' to avoid potentially wrong conclusions if ROC curves are not available.

Keywords X-ray image inspection · Visual search · Signal detection theory · Detection measures

Introduction

X-ray baggage screening at airports is an essential component for securing air transportation. To prevent passengers from bringing potential threats onto an aircraft, airport security officers visually search X-ray images of passenger bags and decide within seconds whether a bag contains a prohibited item or is harmless. This task can be described as visual inspection consisting of visual search and decision making (Koller, Drury, & Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, & Horne, 2009) in line with the two-component model of Spitz and Drury (1978). An airport security officer's (screener's) decision on whether a bag is harmless (*target absent*) or might contain a prohibited item (*target present*) determines whether a secondary search must be conducted at airport security checkpoints (typically using explosive trace detection and a manual search of passenger bags; Sterchi & Schwaninger, 2015). Table 1 presents the four possible decision outcomes and associated terminology from visual search studies (e.g., Biggs & Mitroff, 2015; Eckstein, 2011; Wolfe, 2007, p. 99), signal detection theory (SDT; e.g., Gescheider, 1997, p. 106; Green & Swets, 1966, p. 34), and X-ray baggage screening (e.g., Cooke & Winner, 2007; Schwaninger, Hardmeier, & Hofer, 2005).

In detection theory (Macmillan & Creelman, 2005), the percentage of bags that contain a prohibited item that are correctly classified as such is called the *hit rate* (*HR*), whereas the percentage of harmless bags that are falsely considered to contain a prohibited item is the *false alarm rate* (*FAR*). There is a trade-off between the HR and the FAR: If, for example, someone's tendency to respond with *target present* increases, both the HR and FAR will increase. At its extremes, someone could decide to always respond with *target present*, thereby resulting

✉ Yanik Sterchi
yanik.sterchi@fhnw.ch

[1] University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Riggenbachstrasse 16, CH-4600 Olten, Switzerland

Published online: 25 January 2019

Table 1 Outcome of decisions depending on stimulus using the terminology of visual search, signal detection theory, and X-ray baggage inspection

Stimulus	Decision	
	Target absent No signal Bag is harmless	Target present Signal Bag requires secondary search
Target absent Noise No prohibited item present	Correct rejection	False alarm
Target present Signal plus noise Prohibited item present	Miss	Hit

Note. Target present and *target absent* are terms used in visual search studies (Biggs & Mitroff, 2015; Eckstein, 2011; Wolfe, 2007, p. 99). *Noise, no signal, signal plus noise, signal, hit, miss, false alarm,* and *correct rejection* are terms used in signal detection theory (Gescheider, 1997, p. 106; Green & Swets, 1966, p. 34). The other terms have been used in X-ray baggage inspection studies (Cooke & Winner, 2007; Schwaninger, Hardmeier, & Hofer, 2004)

in a HR and FAR of 100%. Individuals with the same ability to detect prohibited items can have different HRs and FARs because of differences in their response tendency (also referred to as *response bias*; Macmillan & Creelman, 2005). SDT provides measures (such as d' and A') for assessing detection performance. These can be calculated from HR and FAR and are assumed to be (relatively) independent of the observer's response tendency (Macmillan & Creelman, 2005, p. 39). Since 9/11, a growing body of research on X-ray image inspection of passenger bags has led to an increasing use of d' and A' in this domain (e.g., Brunstein & Gonzalez, 2011; Halbherr, Schwaninger, Budgell, & Wales, 2013; Ishibashi, Kita, & Wolfe, 2012; Madhavan, Gonzalez, & Lacson, 2007; Mendes, Schwaninger, & Michel, 2013; Menneer, Donnelly, Godwin, & Cave, 2010; Rusconi, Ferri, Viding, & Mitchener-Nissen, 2015; Schwaninger, Hardmeier, Riegelnig, & Martin, 2010; Yu & Wu, 2015). Moreover, d' and A' are also frequently used in related domains, such as the inspection of medical X-ray images (e.g., Chen & Howe, 2016; Evans, Tambouret, Evered, Wilbur, & Wolfe, 2011; Evered, Walker, Watt, & Perham, 2014; Nakashima et al., 2015) and visual search tasks with artificial stimuli (e.g., Appelbaum, Cain, Darling, & Mitroff, 2013; Huang & Pashler, 2005; Ishibashi & Kita, 2014; Miyazaki, 2015; Russell & Kunar, 2012).

However, as will be discussed in more detail below, the results of several studies in recent years cast doubt on the validity of using d' or A' for X-ray image inspection tasks (i.e., visual search and decision tasks). Before discussing these findings, we shall briefly summarize the theory behind d' and A', and the methods used to evaluate their validity.

First, d' is based on SDT, which, in turn, has its roots in statistical decision theory. For a detailed introduction to SDT, we recommend Green and Swets (1966), Macmillan and Creelman (2005), Wickens (2002), and Gescheider (1997, pp. 105–124). The basic idea of SDT is that when confronted with a binary detection or decision task, cognitive information processing will ultimately result in some type of one-dimensional subjective evidence variable for or against one of the two alternatives (Wickens, 2001, p. 150). This subjective evidence variable is also called the *decision variable* (Macmillan & Creelman, 2005, p. 16). Figure 1a and b show this evidence/decision variable on the *x*-axis. Because the process leading to the evidence is noisy, target-absent (noise) and target-present (signal plus noise) trials both produce a distribution of the decision variable. Whereas the expected value is higher for the target-present trials than for the target-absent trials, the two distributions overlap and do not allow a perfect distinction between the two alternatives. SDT further assumes that individuals derive their decisions by setting a threshold, called the *criterion*, to the decision variable. If the evidence falls short of the criterion, subjects decide that a target is absent (noise); if it exceeds the decision criterion, then they decide that a target is present (signal plus noise). The HR and FAR then each correspond to the cumulative density of one of the two evidence distributions with the criterion as the lower bound (colored areas in Fig. 1a and d). SDT assumes that the criterion can be shifted, with a *liberal* criterion resulting in a higher HR and FAR, and a *conservative* criterion, resulting in a lower HR and FAR. Figure 1a presents an example based on the assumption that the evidence distributions of the two alternatives are normal with equal variance. This equal-variance Gaussian model is the most common model of SDT (Pastore, Crawley, Berens, & Skelly, 2003) and the basis for the detection measure d'. In the equal-variance Gaussian model, d' is the distance between the means of the two distributions in units of their standard deviation and it fully defines the detection performance, called *sensitivity*. The detection measure d' can be calculated as

$$d' = z(HR) - z(FAR) \tag{1}$$

where z is the inverse of the cumulative distribution function of the standard normal distribution (Green & Swets, 1966). The

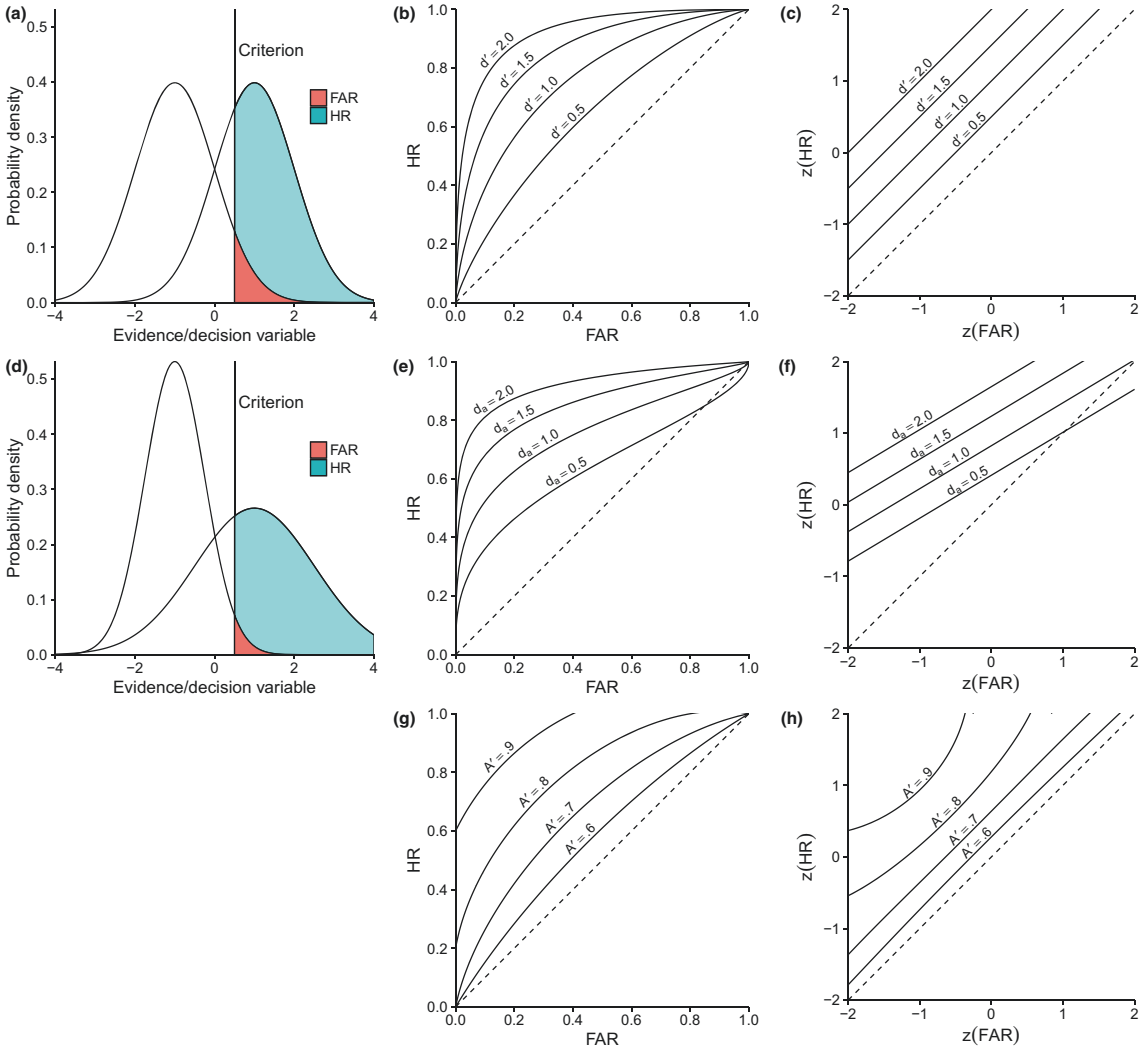

Fig. 1 Illustration of noise and signal-plus-noise distribution (first column), receiver operating characteristic (ROC) curves (second column), and ROC curves in *z*-transformed space (*z*ROC; third column) corresponding to d' (first row), d_a (second row), and A' (third row)

receiver operating characteristic (ROC) curve (Fig. 1a) describes pairs of HR and FAR values for constant levels of d'. If these ROC curves are illustrated in *z* units with *z*(FAR) as the abscissa and *z*(HR) as the ordinate (hereafter, *z*ROC), they form lines with slope 1 and d' as their intercept (Fig. 1b).

Whereas SDT is often interpreted as implying the equal variance Gaussian model (Pastore et al., 2003), SDT can also assume other underlying evidence distributions. One example is an SDT model that assumes the two evidence distributions to be normal, but with unequal variance. For a given ratio *s* between the standard deviation of the target-absent (noise) and target-present (signal-plus-noise) distribution, the resulting *z*ROC has slope *s*. For this SDT model, Macmillan and Creelman (2005) proposed using Simpson and Fitter's (1973) detection measure:

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

If the ROC curve is known empirically, there are also detection measures that can be estimated without any model assumptions. The most popular of these measures is the area under the curve (AUC; Pepe, Longton, & Janes, 2009). When only one point of the ROC curve is known, Pollack and Norman (1964) provide a *one-point estimation* of the AUC:

$$A' = 0.5 + \frac{(HR - FAR)(1 + HR - FAR)}{4HR(1 - FAR)} \;\Big|\; HR \geq FAR. \tag{3}$$

By estimating the AUC with one ROC point, A' should not be considered assumption-free (Macmillan & Creelman, 2005, p. 103; Wickens, 2001, p. 71). Whereas SDT models make explicit assumptions about the decision process that define the shape of the ROC curves, A' also implicitly defines very specific ROC curves as specified by the formula for its calculation. This results in the ROC curves shown in Fig. 1g.

To summarize, each one-point detection measure (detection measure based on only one ROC point, i.e., one value for HR and one for FAR), such as d' or A', implies a specific ROC curve; that is, a specific assumption about how HR and FAR change when response tendency (i.e., the decision criterion) changes. Whether the implied ROC curve is approximately correct determines whether the detection measure is a valid measure of detection performance. Most importantly, because different detection measures imply different ROC curves, they can lead to different conclusions when, for example, interpreting results of X-ray image inspection tasks.

The shape of the ROC curve for a specific task can be investigated by empirically measuring multiple points of the ROC curve. Macmillan and Creelman (2005) describe four methods with which to gather ROC data from study participants. The first is based on confidence ratings. Instead of providing only a binary decision, the participants provide a rating on a k-point Likert scale – for example, ranging from *target certainly absent* to *target certainly present*. Alternatively, the participants deliver the binary response (e.g., *target present* or *target absent*) and then rate their confidence regarding that decision. Each change in level of confidence is then considered as a possible decision criterion (Macmillan & Creelman, 2005, pp. 51–54). With this approach, k - 1 ROC points can be derived for k response categories.

The other three methods for deriving multiple points of the ROC curve are based on manipulating response tendency (i.e., criterion; Macmillan & Creelman, 2005, p. 71). One method is to manipulate the rewards and costs of a decision (e.g., study participants can be paid according to the amount of hits and false alarms, and the reward of a hit and cost of a false alarm can be manipulated). A second method is to instruct the participants directly to change their criterion by, for example, being conservative in responding *target present* on one set of trials and being more liberal on another set. The third method for gathering ROC points is to manipulate the presentation probability of the signal (Macmillan & Creelman, 2005, p. 72) – the so-called *target prevalence* (Wolfe, Horowitz, & Kenner, 2005). If, for example, most trials contain a prohibited item, subjects will shift their response tendency toward *target present* and therefore achieve a higher HR and FAR. Manipulating the criterion means that each point of the ROC curve requires a separate condition (payoff, instruction, or target prevalence).

Of these four methods, gathering confidence ratings can be applied relatively easily and rapidly, but it is heavily based on the concept of SDT. It is assumed that the subject's decision process is based on a decision variable and that a subject derives a confidence rating from that variable. The other three methods do not require such assumptions because they measure actual decisions under different conditions.

When multiple ROC points are gathered, they can be interpolated to calculate A_g – an estimate of the AUC – without relying on assumptions about the shape of the ROC curve (Pollack & Hsieh, 1969). Hofer and Schwaninger (2004) compared different measures of detection performance and investigated ROC curves derived from confidence ratings in an X-ray image inspection task. They derived ROC curves from pooled confidence ratings and found deviances from symmetrical ROC curves that would be more consistent with the two-state low-threshold theory (Luce, 1963) or non-equal variance Gaussian SDT. However, they also found that d', A', and Δm (a measure for non-equal variance SDT; Wickens, 2001) were highly correlated.

Several other studies using target prevalence manipulations have cast further doubt on the validity of d' and A' for X-ray baggage inspection. Wolfe et al. (2007) conducted a series of experiments in which subjects performed an X-ray baggage inspection task under varying target prevalence conditions. They found a reduced HR and FAR in low target prevalence conditions with averaged results seeming to lie on a zROC line with a slope of 0.6. Two further publications (Godwin, Menneer, Cave, & Donnelly, 2010a; Van Wert, Horowitz, & Wolfe, 2009) reported zROC slopes similar to those reported by Wolfe et al. (2007), and another study reported a slope of 0.56 (Wolfe & Van Wert, 2010), which is also close to 0.6.

Under Gaussian SDT assumptions, a zROC slope of 0.6 indicates that the target-absent (noise) distribution has a smaller standard deviation than the target-present (signal-plus-noise) distribution. A possible explanation for this is that prohibited items vary in difficulty and this brings additional variation into the target-present distribution.

The aim of our study was to investigate the validity of the detection measures d', A', and d_a and to derive recommendations on how to calculate detection performance in future studies on X-ray image inspection, visual search, and decision tasks. We explored this using two experiments, in which professional X-ray screeners completed a simulated X-ray baggage inspection task. In the first experiment, response tendency (criterion) was manipulated through instruction to test whether it affected the detection measures. The experiment included targets that were known from training and targets that were novel, which resulted in two levels of sensitivity. Valid detection measures should be independent of response tendencies; however, they should differentiate well between different levels of sensitivity. We therefore calculated the effect size of the difference in the detection measures between known and novel targets as an indicator of how well they differentiate between the two levels of sensitivity. In the second experiment, the participants provided confidence ratings that were used to investigate whether the ROC curves are approximately linear in zROC space, as assumed by both d' and d_a, and to estimate the zROC slope.

Experiment 1

For this study, we reanalyzed data from Sterchi, Hättenschwiler, Michel, and Schwaninger (2017). The original study evaluated how the rejection rate of screeners can be manipulated, and how performance was related to knowledge about everyday objects. In the experiment, 31 professional screeners completed a simulated X-ray baggage screening task in which the criterion was manipulated directly through instructions. Half of the prohibited items used in the study were known to the screeners from training, whereas the other half were novel. This corresponds to two levels of task difficulty. This experiment allowed us to observe a criterion shift with two levels of sensitivity induced by other means than the previously applied manipulations of target prevalence.

For a detection measure to be valid, it should not be affected by a shift in the decision criterion. In line with the results of the previous studies mentioned above (Godwin, Menneer, Cave, & Donnelly, 2010a; Hofer & Schwaninger, 2004; Van Wert et al., 2009; Wolfe et al., 2007; Wolfe & Van Wert, 2010), we expected the *z*ROC slope to be around 0.6, and therefore for d' to decrease when the criterion was shifted to a more liberal level (more target-present responses) in Experiment 1. Both d' and A' are symmetric – any point (HR_x, FAR_x) leads to the same value of d' and A' as ($1 - HR_x$, $1 - FAR_x$) – and this implies equal variance in terms of SDT (Macmillan & Creelman, 2005, p. 103). We therefore also expected A' to decrease when the criterion decreased. As a result of the expected *z*ROC slope of 0.6, a criterion shift should not affect d_a based on that slope. We also aimed at validating A_g. As already described in the introduction, A_g is an estimate of the AUC that does not assume a specific shape of the ROC curve but requires multiple ROC points (e.g., derived from confidence ratings) and is therefore not a one-point detection measure like d', d_a, or A'. Because A_g should not depend on the shape of the ROC curve, it was expected to remain constant. A detection measure should not change when the decision criterion changes; however, it should differentiate well between different levels of ability to detect targets. We therefore analyzed effect sizes of the detection measures when comparing detection performance for the two levels of task difficulty resulting from known and novel prohibited items.

Method

Participants

A total of 31 screeners (20 females) from an international airport participated in this experiment. They were all certified screeners, which means that they were qualified, trained, and certified according to the standards set by the appropriate national authority (civil aviation administration) in accordance with the European Regulation (European Commission, 2015). The participating screeners were between 26 and 61 years old ($M = 45.4$, $SD = 8.9$) and had between 2 and 26 years of work experience ($M = 8.4$, $SD = 5.5$). The research complied with the American Psychological Association 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 each participant.

Design

The experiment used a 2 × 2 design with two instructions to manipulate response tendency (normal decision vs. liberal decision) and with two levels of task difficulty (targets known from training vs. novel target items) as within-subject factors. Dependent variables were HR, FAR, d', d_a, A', A_g, response times, and eye-tracking data.

Stimuli and materials

The simulated X-ray baggage inspection task contained 128 X-ray images of passenger bags. Of these, 64 images contained one prohibited item (target-present images). They were merged into X-ray images of passenger bags using a validated X-ray image merging algorithm (Mendes, Schwaninger, & Michel, 2011). Four categories of prohibited items were used to create these target-present images: 16 X-ray images contained a gun, 16 images a knife, 16 images an IED, and 16 images contained other prohibited items. To create these 16 X-ray images per threat category, eight threat items per category were each used twice, once in an easy view (as defined by the two X-ray screening experts and the authors) and once rotated (by 85° around the horizontal or vertical axis).

Further, for each threat category, half of the prohibited items were part of the training system (Koller, Hardmeier, Michel, & Schwaninger, 2008; Schwaninger, 2004) used at the particular airport (known targets). The other half of the prohibited items were newly recorded (novel targets). Visual comparisons were used to ensure that they were different from the prohibited items contained in the training system (see Fig. 2 for an example).

All 128 X-ray images were equally divided into four test blocks such that each block contained the same number of known and novel targets per category and viewpoint. X-ray images were presented in a random order within each of the four blocks. The order of the blocks was counterbalanced across the participants.

For eye tracking, we used an SMI RED-m eye tracker with a gaze sample rate of 120 Hz, gaze position accuracy of 0.5°, and spatial resolution of 0.1°. This noninvasive, video-based eye tracker was attached to a 22-in. TFT LCD screen with a

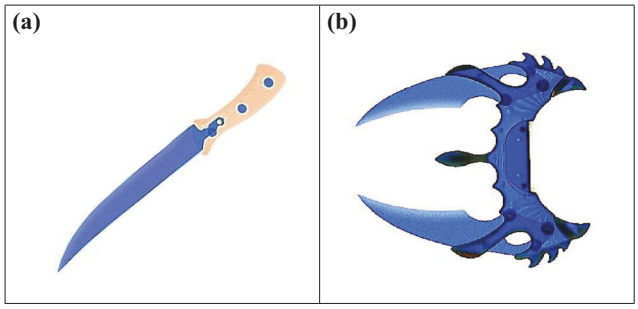

Fig. 2 Two examples of the prohibited item category *knife*: (**a**) example of a known target item and (**b**) example of a novel target item (Asian combat knife)

resolution of 1,280 × 1,024 pixels placed 50–75 cm from the participant. The stimuli (X-ray images) covered about two-thirds of the screen. Eye tracking was used to examine the users' eye movements using a post hoc analysis of visual fixations falling within a certain area of interest (AOI). Therefore, in each target-present image, a screening expert manually drew the AOI around the target item (BEGAZE Software; SensoMotoric).

Procedure

The screeners were tested individually. Each session began with a 9-point calibration of the eye-tracking apparatus. The participants had to follow a moving black dot with their eyes. Then, the task was introduced with on-screen instructions. The screeners were instructed to visually inspect X-ray images of passenger bags by searching for prohibited items and deciding whether each bag was harmless (*target absent*) or might contain a prohibited item (*target present*) and would therefore require a secondary search. The screeners were further instructed that the test contained four blocks. For two blocks, they should inspect (i.e., search and decide) the image as if they were working at a checkpoint (referred to in this article as a *normal decision*). For the other two blocks, they were instructed to visually analyze each object in the X-ray image and decide that the bag was harmless only if each object in the image could be recognized as harmless (*liberal decision*). After the instructions, ten practice trials followed to familiarize the screeners with the task itself and the user-interface of the simulator. The practice trial consisted of five target-absent and five target-present images presented in random order without any feedback on the correctness of the response.

For the test, each trial started with a fixation cross displayed at the center of the screen. After this had been fixated continuously for 1.5 s, it was replaced by an X-ray image. Screeners had to decide whether the content of this image was harmless or not by pressing a key, and then had to give a confidence rating on a 10-point scale ranging from 1 (*very unconfident*) to 10 (*very confident*). There was no feedback on the correctness of responses, and the participants took about 30 min to complete the test.

Data analysis

A HR of one or FAR of zero leads to an infinite value of d' and d_a. For the calculation of d' and d_a, HR and FAR values were therefore transformed using the log-linear rule to correct for extreme proportions (Hautus, 1995), which is one of the two common adjustments to avoid infinite values (Macmillan & Creelman, 2005, p. 8). All within-subject contrasts were tested with exact permutation tests that are appropriate for skewed data and smaller sample sizes. For the estimation of d_a, the slope parameter was set to 0.6 in accordance with previous findings from studies that manipulated target prevalence (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). For *z*ROC slopes and effect sizes, we report bootstrapped BCa-CIs (Efron, 1987) based on 20,000 resamples.

In a review of ROC curves in recognition memory, Yonelinas and Parks (2007) raised the concern that the manipulation of the criterion (i.e., pay-off, instruction, or target prevalence) might also influence sensitivity. In our experiment, we analyzed eye-tracking data to control whether our manipulation also affected search performance and not just decision making. It can be assumed that failure to detect a target can arise from a *scanning error* (Cain, Adamo, & Mitroff, 2013; Kundel, Nodine, & Carmody, 1978; Nodine & Kundel, 1987), where the target is never fixated. If the target is fixated, inspection can still fail because of *recognition* or *decision errors*, and it is unclear whether a distinction between recognition and decision errors is possible and useful (Cain et al., 2013).

In accordance with McCarley's (2009) study, we tested the effect of our manipulation by calculating the proportion of target-present trials with one or more fixations within the AOI (i.e., the location of the target). Rich et al. (2008) also distinguished fixated and non-fixated targets to analyze search errors. They noted that if a target is not fixated, this does not necessarily mean that it was missed during the visual search. However, a target missed during the visual search is more likely to not have been fixated. If the proportion of target-present trials on which the target was fixated is not affected by the manipulation of the criterion, this indicates that the changes in HR and FAR are not caused by search errors in which the study participants simply failed to look at the relevant part of the image (Rich et al., 2008).

Results

The instructions for the liberal decision condition were designed to change response tendency, that is, to increase the

participants' relative frequency of responding with *target present* (rejection rate). A manipulation check revealed an effect of the instruction on the rejection rate with a Cohen's *d* of 0.58. However, ten of the participants did not even show a small increase in the rejection rate (i.e., increase smaller than a Cohen's *d* of 0.20). Because we were interested in whether the detection measures change when participants change their response tendency (and not how successfully we could induce such a change), we excluded participants who did not change their rejection rate from further analysis. The excluded participants did not differ significantly in their HR for known targets (excluded: $M = .78$, included: $M = .79$, $p = .636$), HR for novel targets (excluded: $M = .63$, included: $M = .58$, $p = .298$), or FAR (excluded: $M = .11$, included: $M = .09$, $p = .570$). Table 2 shows the means and standard deviations of the normal decision and liberal decision condition for HR, FAR, d', d_a, A', and A_g. Exact permutation tests revealed a significantly lower d' in the liberal decision condition for both known ($p = .041$) and novel ($p = .002$) targets. Moreover, A' was significantly lower for both known ($p = .034$) and novel ($p = .017$) targets. For both d_a (known targets: $p = .714$, novel targets: $p = .383$) and A_g (known targets: $p = .322$, novel targets: $p = .750$), differences did not attain significance. Table 2 also shows the standardized average difference of the detection measures between the two decision conditions as an indicator for the within-subject effect.

The HR and FAR of the two decision conditions were used to calculate individual *z*ROC slopes for known and novel targets separately. The estimated slope had a median of 0.53 (95% BCa-CI [0.24, 0.75]) and a mean of 0.62 (95% BCa-CI [0.34, 1.04]) for known target items, and a median of 0.56 (95% BCa-CI [0.00, 0.83]) and mean of 0.49 (95% BCa-CI [0.27, 0.78]) for novel target items (slopes were first converted into angles of incline and converted back after averaging because steep slopes would otherwise disproportionately influence the mean).

Table 3 summarizes the response time (time from the onset of image display until the submission of the decision by the participant) for correct responses by image type (target-present trials vs. target-absent trials) and decision condition (normal decision vs. liberal decision). For both target-present and target-absent trials, permutation tests indicated a significant difference in response time between normal and liberal decision (target-present trials: $p = .004$, target-absent trials: $p < .001$).

To control whether the criterion manipulation affected search errors, we calculated the proportion of target-present trials with at least one fixation within the AOI (i.e., the location of the target; see McCarley, 2009). Three participants had to be excluded from the analysis of eye-tracking data because they had either no fixations or no saccades recorded in 73%, 52%, or 24% of their trials, which indicated difficulty with eye tracking for these participants. The remaining 18 participants had a total of 1,151 target-present trials. Twelve (1%) of these had to be excluded because either no fixations or no saccades were recorded. One further trial was excluded because the fixation was in the AOI at the time of stimulus onset. Then, for each participant, the proportion of target images on which the participant fixated the target was calculated separately for the two decision conditions (normal and liberal decision) and the two target types (known and novel targets). Table 4 shows the means and standard deviations of these proportions. The difference between the two decision conditions did not attain significance for either known targets ($p = .459$) or novel targets ($p = .675$), which suggests that the instruction to decide with a more liberal criterion did not affect search errors.

To investigate the statistical power of the detection measures in terms of reflecting differences in task difficulty (known vs. novel targets) for each detection measure and each of the two decision conditions, we calculated standardized differences (i.e., differences divided by the standard deviation of the differences) as effect sizes of the detection measures between known and novel targets (Table 5). Because d_a is a linear transformation of d' when the false alarm rate is constant, the effect sizes of d' and d_a were identical.

Figure 3 shows the ROC curves based on the three detection measures d', A', and d_a of the normal decision condition for known targets (curves with higher HR for a given FAR)

Table 2 Mean (*SD*) of the normal and liberal decision condition and the effect size (standardized difference) of the decision condition for hit rate (HR), false alarm rate (FAR), and detection measures d', A', d_a, and A_g

Decision condition	HR	FAR	d'	d_a	A'	A_g
	Known targets					
Normal decision	.79 (.10)	.09 (.08)	2.25 (0.61)	2.03 (0.57)	.916 (.044)	.894 (.072)
Liberal decision	.90 (.10)	.25 (.13)	2.01 (0.58)	2.08 (0.61)	.899 (.049)	.906 (.073)
Effect size			-0.40	-0.08	-0.42	0.23
	Novel targets					
Normal decision	.58 (0.14)	.09 (.08)	1.63 (0.41)	1.28 (0.38)	.851 (.040)	.799 (.082)
Liberal decision	.71 (0.13)	.25 (.13)	1.27 (0.44)	1.19 (0.43)	.817 (.074)	.793 (.076)
Effect size			-0.70	-0.19	-0.50	-0.07

Table 3 Response times [ms] for correct responses

	Normal decision		Liberal decision	
	M (*SD*)	*Mdn*	*M* (*SD*)	*Mdn*
Target-present	6,000 (2,407)	4,295	8,018 (4,331)	6,291
Target-absent	6,813 (2,798)	5,873	11,162 (6,872)	9,464

Note. The reported means and standard deviations are based on individual mean response times, and the reported medians on individual median response times

and novel targets (curves with lower HR for a given FAR). Because this figure is based on pooled data, it should be interpreted with caution: The aggregation of individual ROC curves can distort their shape, and the figure is therefore not a one-to-one illustration of the tested hypotheses (Yonelinas & Parks, 2007; see the Appendix for a discussion of pooling).

Discussion

In Experiment 1, we instructed X-ray screeners for one condition to visually inspect X-ray images in the same manner used when they performed their job. For another condition, they were instructed to apply a more liberal decision criterion. Half of the target-present trials contained target items known from training, the other half contained novel target items. As can be seen in Fig. 3, the resulting four points defined by the pooled HR and FAR fit the ROC curve implied by d_a that was set to a slope of 0.6, as suggested by previous research (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). The permutation tests revealed that d' and A' values decreased when screeners were instructed to apply a more liberal decision, which casts doubt on the validity of these detection measures in the context of X-ray image inspection. By contrast, d_a with a slope of 0.6 and A_g did not change significantly between the two experimental conditions.

The fact that the instructed, more liberal criterion caused a decrease in d' and A' is in line with previous findings of changes in d' when target prevalence manipulations induced a shift in the criterion (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). The results of these studies also suggest that d' and A' can lead to wrong conclusions when used to decompose a unidirectional change of HR and FAR into sensitivity and criterion changes.

When trying to induce a criterion shift using experimental manipulation, there is a risk that the manipulation might also affect sensitivity (Yonelinas & Parks, 2007). In our experiment, the given instruction to decide more liberally slowed the response times. Similarly, studies that manipulated target prevalence also found slower responses in high target prevalence conditions (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). Our main findings should be robust regarding a potential change in sensitivity for two reasons: First, we found no difference in the share of images with target fixation between the two decision conditions. This supports the assumption that the observed change in HR and FAR was caused by a change in decision making and not a change in search errors (McCarley, 2009; Rich et al., 2008). Second, if the manipulation affected sensitivity, then one would expect higher sensitivity in the liberal decision condition in which response times were longer (following the line of argument in Wolfe et al., 2007). Such an accidental effect on sensitivity could therefore not explain the decrease we found in d' and A'.

Table 4 Mean (*SD*) share of images per subject with a recorded fixation within the area of interest

	Share AOI fixations	
Image type	Normal decision	Liberal decision
Known target	.713 (.237)	.740 (.258)
Novel target	.742 (.165)	.730 (.180)

Experiment 2

In Experiment 1, we calculated d', A', and d_a, for which we set the slope to 0.6 based on previous findings (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). d_a was found to be a more valid detection measure than d' and A'. However, estimations of the slope parameter with the data from Experiment 1 resulted in large confidence intervals. Further, ten of the participants were excluded because they failed the manipulation check, which might have biased the sample. Experiment 2 was therefore intended to provide a more precise estimation of the slope parameter and to further investigate the validity of detection measures using another methodological approach: multiple ROC points were obtained by analyzing confidence ratings. In comparison to Experiment 1, the criterion was not manipulated directly, and the test therefore included more trials per participant and condition.

Methods

Participants

A total of 124 professional, certified cabin baggage screeners (68 female) from an international airport participated in

Table 5 Effect size (standardized difference) [and 95% confidence intervals] of target novelty (known vs. novel targets)

	d' / d_a		A'		A_g	
Normal decision	1.60	[1.21, 2.10]	1.72	[1.34, 2.15]	1.24	[0.84, 1.64]
Liberal decision	1.98	[1.20, 3.02]	1.73	[1.11, 2.48]	2.20	[1.35, 3.04]

Experiment 2. The participants were between 22 and 64 years old (M = 44.3, SD = 11.2; one participant did not report his/her age) and they had up to 29 years of work experience (M = 7.1, SD = 5.6; seven participants did not report their work experience). The research 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 each participant.

Stimuli and materials

The test consisted of 128 X-ray images of real passenger bags. Half of these images contained a prohibited item. The merging of the prohibited items into the bag images was performed in the same manner as in Experiment 1 using a validated algorithm (Mendes et al., 2011). Four categories of prohibited items were used: 16 images contained a gun, 16 images a knife, 16 images an IED, and 16 explosive material. Each prohibited item appeared twice, once in an easy view and once rotated. None of the prohibited items were part of the training system used at the particular airport. The 128 images were equally divided into two blocks with each block containing the same number of targets per category and view. Images were presented in a random order within the block. The order of the two blocks was counterbalanced across the participants.

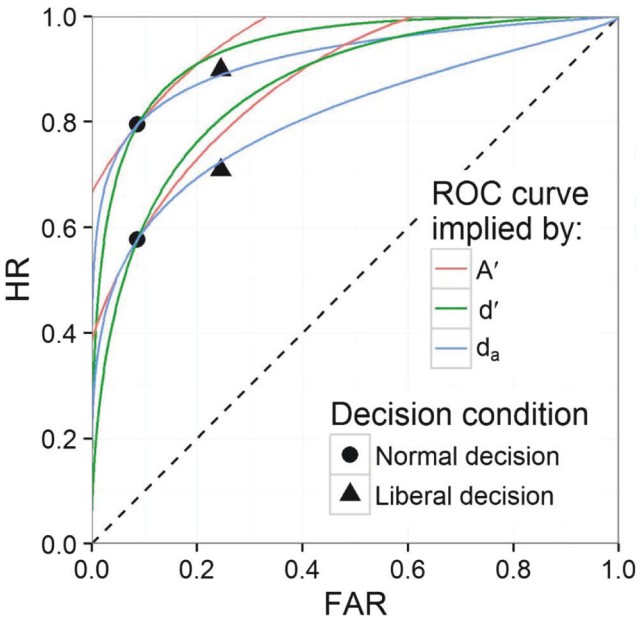

Fig. 3 Receiver operating characteristic (ROC) curves implied by d', A', and d_a estimated by the pooled hit rate (HR) and false alarm rate (FAR) of the normal decision condition for known prohibited items (higher HR) and novel prohibited items (lower HR)

Procedure

The participants were tested in groups of maximally six screeners at a time. The screeners had to inspect the X-ray images for prohibited items. If they detected a prohibited item, they had to mark its location in the image (this was conducted for another study). They had to press a key to decide whether the bag was harmless or not, and they then had to assign a confidence rating on a 5-point scale ranging from 1 (*very unconfident*) to 5 (*very confident*). To become familiar with the test, the instruction was followed by eight practice trials, on which the screeners received feedback on the correctness of the responses. During the test itself they did not receive feedback. Participants were allowed to take a short break after the first half of the test that lasted for 1 min in average. Participants took about 20 min to complete the test.

Data analysis

For each participant, the HR and FAR were calculated for the different levels of confidence rating according to Macmillan and Creelman (2005, pp. 51–54), resulting in nine ROC points per participant.

To estimate individual slope parameters based on the confidence ratings, we used the maximum likelihood estimation algorithm LABROC4 developed by Metz, Herman, and Shen (1998). Because the slope parameter is the ratio of two differences in two variables, it is inappropriate to directly calculate its mean (because steep slopes result in large numbers, a horizontal zROC, for example, has a slope of zero and a vertical zROC has a slope of infinity and the mean of the two slopes would only consider the vertical slope). We therefore arctan-transformed the slope parameters into angles of incline before averaging, and then transformed them back for interpretability.

Results

One participant provided the maximum confidence level for all trials and was therefore excluded. A second participant had to be excluded because all derived ROC points for FAR were either zero or one, not allowing for a maximum likelihood estimation of the slope parameter. The remaining 122 participants achieved a mean HR of .70 (*SD* = .07) with a mean FAR of .07 (*SD* = .05). The response time (time from the onset of the image display until the submission of the decision by the participant) is summarized in Table 6 for correct responses by image type (target-present trials vs. target-absent trials).

Figure 4 shows individual *z*ROC points and the averaged *z*ROC curves based on confidence ratings (for a discussion of pooling ROC curves see the Appendix). The averaged *z*ROC curves seem to better fit the *z*ROC curve predicted by d_a based on a slope of 0.6 than those predicted by d' or A' (one exception is the mean of the leftmost *z*ROC point, which, however, is distorted downwards as a result of the necessary exclusion of ROC points with a false alarm of zero that are not defined in *z*ROC space).

Arctan-transformed individual slope parameters (i.e., angles of incline) estimated using the LABROC3 algorithm (Metz et al., 1998) are illustrated in Fig. 5. When transformed back, they show a mean of 0.54 (95% BCa-CI [0.50, 0.60]) and median of 0.50 (95% BCa-CI [0.46, 0.55]).

Discussion

In Experiment 2, the participants completed an X-ray baggage inspection task providing confidence ratings for each image. The pooled *z*ROC points and the estimated *z*ROC slopes of around 0.5–0.6 confirm the findings of Experiment 1 that d' and A' overestimate HR, or underestimate FAR when the criterion is shifted and becomes more liberal. The pooled *z*ROC curves were approximately linear, which supports the validity of d_a for the X-ray baggage inspection task in line with the results of Wolfe and Van Wert (2010). The results show a mean slope of 0.54, close to other studies that reported *z*ROC slopes of around 0.6 (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007) and another study that reported a slope of 0.56 (Wolfe & Van Wert, 2010).

Table 6 Response times [ms] for correct responses

	M (*SD*)	*Mdn*
Target-present	4,781 (1,087)	3,816
Target-absent	5,079 (1,959)	4,008

Note. The reported group means and standard deviations are based on individual mean response times, and the reported medians on individual median response times

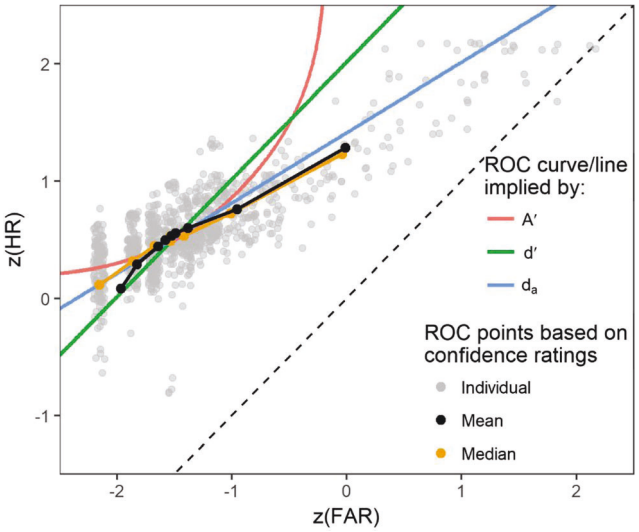

Fig. 4 Individual (grey; jittered) and pooled (black) empirical *z*ROC curves, the lines corresponding to the mean A', d', and d_a with a slope of 0.6, and the chance line (dashed)

Despite the similar *z*ROC slopes found in these studies, one should be cautious to always adopt d_a with a slope of 0.5–0.6 for any X-ray baggage inspection or other visual search task. A non-unit slope *z*ROC implies that there is a point at which the ROC curve falls below the chance line, where the FAR exceeds the HR (Macmillan & Creelman, 2005, p. 68). When sensitivity is sufficiently high, this becomes negligible because it only concerns values very close to the limits of the ROC space. However, for low sensitivity (e.g., for difficult items or inexperienced X-ray screeners), a *z*ROC with a slope of 0.5–0.6 implies below-chance performance for a possibly relevant range of the decision criterion (see Fig. 1e). It would therefore be reasonable to assume that the *z*ROC slope converges to a unit slope with decreasing sensitivity. Such a convergence has been found repeatedly in research on recognition memory (Brown & Heathcote, 2003; Glanzer, Kim, Hilford,

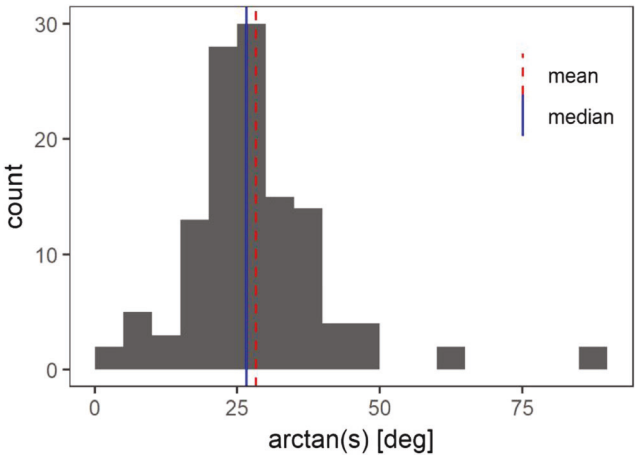

Fig. 5 Distribution, mean (red dashed line), and median (solid blue line) of arctan-transformed individual slope parameters

& Adams, 1999; Hirshman & Hostetter, 2000; Ratcliff, McKoon, & Tindall, 1994).

In addition to the level of sensitivity, other factors might influence the slope parameter. There is some empirical evidence that the *z*ROC slope might vary between different implementations of the X-ray baggage inspection tasks or depending on the participants: Alongside our findings and other studies reporting *z*ROC slopes around 0.5–0.6 (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010), one study found a lower d' for lower target prevalence (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013), which indicates a *z*ROC slope larger than one. There are also a few studies that show an effect of target prevalence on HR and FAR without a significant effect on d' (Godwin, Menneer, Cave, Helman, et al., 2010b; Ishibashi et al., 2012) or A' (Godwin, Menneer, Cave, Thaibsyah, & Donnelly, 2015). They therefore do not contradict a unit-slope *z*ROC. To summarize, whereas it is reasonable to infer that a *z*ROC slope is around 0.5–0.6 for many visual inspection, visual search, and decision tasks with X-ray images, this might not be always true. In the following section we discuss how this issue can be addressed in future studies.

General discussion

To investigate the validity of two detection measures commonly used in visual search and decision tasks such as airport security and medical screening, we conducted two studies with different methodological approaches. Experiment 1 manipulated the criterion by direct instruction, whereas Experiment 2 used confidence ratings to generate multiple ROC points. For both studies, d' and A' were found to be invalid detection measures for the investigated X-ray baggage inspection tasks. More specifically, d' and A' would have wrongly indicated lower sensitivity for a more liberal decision criterion.

Studies investigating the effect of target prevalence on X-ray baggage inspection tasks also found d' to indicate lower sensitivity for more liberal decision criteria where equal or lower sensitivity would be expected (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). Our studies extend this research by showing that this phenomenon is not specific to the effect of target prevalence but also holds for other means of manipulating the criterion, and therefore seems to be a property of the ROC curve of the X-ray baggage inspection task in general.

Despite A' not making any assumptions about the underlying decision processes, A' implies a very specific and symmetric ROC curve (Macmillan & Creelman, 2005). It should therefore not be expected to have an advantage over d', which the results of our studies confirmed. The general discussion and our recommendations will therefore focus on d' and d_a.

When lifting the assumption of equal variance, the Gaussian SDT model is extended by an additional parameter: the ratio s between the standard deviation of the signal-plus-noise (target-present) and noise (target-absent) distribution. The Gaussian SDT model assumes an ROC curve that becomes a straight line when *z*-transformed with parameter s as its slope. For detection measure d_a, which corresponds to this model, to be valid for X-ray baggage inspection tasks, *z*ROC curves should be approximately linear. In line with a study from Wolfe and Van Wert (2010), the results of Experiment 2 show approximately linear pooled *z*ROC curves. In our experiments, the slope parameter was around 0.5–0.6, which corresponds well with the findings in other experiments that investigated the X-ray baggage inspection task (Godwin, Menneer, Cave, & Donnelly, 2010a; Wolfe et al., 2007; Wolfe & Van Wert, 2010). However, the slope parameter might depend on the level of sensitivity and might vary between different implementations of the X-ray baggage inspection tasks or depending on the participants.

To better understand what factors influence the slope parameter, a better understanding of the inspection process would be useful and should be the focus of future studies. From the perspective of Gaussian SDT, a *z*ROC slope smaller than one implies that the signal-plus-noise distribution has a higher standard deviation than the noise distribution. A possible explanation for this is that prohibited items can vary strongly in how well they can be recognized – for example, depending on item category (Halbherr et al., 2013; Koller et al., 2009) and the exemplar within categories (Bolfing, Halbherr, & Schwaninger, 2008; Schwaninger et al., 2007). The SDT framework might have to be extended to provide a better model of the visual inspection process. For instance, Wolfe and Van Wert (2010) described the task as successive decisions for single items within the X-ray image. This model assumes that the observer makes a decision according to SDT for one item after the other until the observer either decides that an item is prohibited or a quitting threshold is reached. Conceptually, this is similar to the two-component model of visual inspection by Spitz and Drury (1978), which has been applied to the visual inspection of X-ray images and consists of visual search and decision processes (Koller et al., 2009; Wales et al., 2009). For modeling recognition memory, SDT has been extended in various forms by assuming that recognition can be based on either recollection or familiarity (Yonelinas & Parks, 2007). Similarly, different types of recognition might apply in X-ray baggage inspection – some items might be recognized with certainty, whereas for other items, a decision has to be made under high uncertainty.

Our studies and the reviewed literature focus on the task of inspecting X-ray images of passengers' cabin baggage. Our findings do not necessarily directly translate to related domains, such as the inspection of medical X-ray images or other visual search tasks with artificial stimuli; however, such

related domains should also not expect d' and A' to be valid without further consideration. Future research should specifically investigate to what extent the findings we report also apply in related domains.

We hope that future research will provide more insights into the image inspection process; however, we suggest a critical yet pragmatic approach when investigating performance in image inspection tasks. As famously stated by Box (Box & Draper, 1987, p. 424), "all models are wrong, but some are useful." In X-ray image inspection, the main use of a detection measure is to identify whether a unidirectional difference in HR and FAR (i.e., when both HR and FAR are higher in one group or condition) is only a difference in the decision criterion or also a difference in detection performance in terms of sensitivity. That is, a comparison of detection measures should answer the question of who would have the higher HR and lower FAR if everyone used a similar decision criterion.[1] For one-point detection measures, the implied ROC curve therefore needs to be approximately correct. Our studies and the reviewed literature show that for X-ray baggage inspection, this is often not the case for d' and A'. Instead, d_a with a zROC slope of 0.5 to 0.6 often seems to provide the better measure. However, while it is not clear what factors determine the zROC slope, we recommend testing d_a with a slope of 0.5 in addition to d_a with a slope of 1 (i.e., d') as the upper and lower bound, respectively. Another approach is to gather confidence ratings and use A_g as a detection measure. Whereas d', A', and d_a imply a specific shape of ROC curve, A_g is conceptually valid for any form of ROC curve. However, it requires the collection of confidence ratings, and is based on the assumption that these confidence ratings allow a prediction of alternative criterion locations at an individual level. Moreover, some methodological problems can arise because A_g estimates the AUC by linearly interpolating empirical ROC points (Pollack & Hsieh, 1969). This approach increasingly underestimates the AUC with a decreasing number of ROC points (Macmillan & Creelman, 2005, p. 64). A_g might therefore require a relatively high number of trials to be a valid detection measure. In Experiment 1, A_g performed acceptably well – it was not significantly affected by the manipulation of the decision condition, and differentiated between known and novel targets with statistical power comparable to d_a. However, this is only limited support for the measure, as the results are restricted to a within-subject comparison of a small sample. Future research might clarify whether confidence ratings allow a reliable prediction of criterion shifts induced by changes in target prevalence or instruction.

In conclusion, X-ray image inspection research and related domains will have to be cautious when using one-point estimates of sensitivity such as d' and A'. We recommend always starting by performing an analysis and discussion of the directly accessible HR and FAR. Estimating the sensitivity and criterion is often only necessary if HR and FAR are affected unidirectionally. In that case, it should be considered that a zROC slope can be expected to lie somewhere between 0.5 and 1 for X-ray baggage inspection tasks. With d_a, effects on sensitivity can be estimated for these two slopes separately to test the two limits of the assumption of constant sensitivity (where the upper limit with a zROC slope of 1 corresponds to d'). Collecting confidence ratings allows to directly estimate the zROC slope for the investigated task, to calculate A_g, which provides an additional estimation of sensitivity, and help to further understand the shape of the ROC curve in X-ray image inspection.

Appendix

Pooling and ROC curves

When investigating receiver operating characteristic (ROC) curves based on the framework of signal detection theory (SDT), in almost all experiments of real interest, some type of averaging must be performed (Macmillan & Creelman, 2005, p. 331). For X-ray image inspection, combining different stimuli in an experiment seems reasonable because this is representative of this task in the real world. However, when responses from different subjects are averaged, the resulting ROC curve can deviate systematically from individual ROC curves, as we will illustrate in the following paragraphs.

Figure 6 assumes two subjects with an identical ROC curve in the shape assumed by Gaussian SDT. If these subjects differ in their decision criterion, their averaged ROC point (i.e., hit

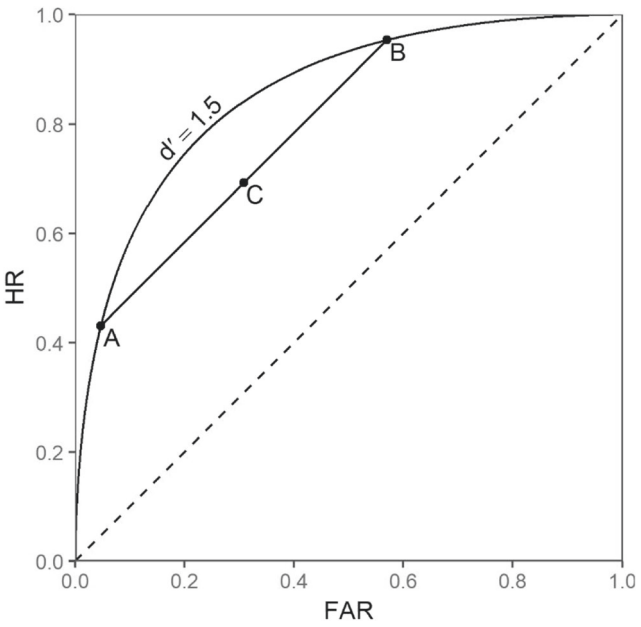

Fig. 6 When the two points A and B from the same receiver operating characteristic (ROC) curve are averaged, the resulting ROC point C is below the original ROC curve

[1] For different levels of sensitivity, it is conceptually not clear what constitutes an equal decision criterion (Macmillan & Creelman, 2005, pp. 36–44).

and false alarm rate) will lie in the middle of the line connecting their individual ROC points and therefore below their true ROC curve. How far away the averaged ROC point is from the true ROC curve depends on the difference between the decision criteria (i.e., the distance between the individual ROC points) and on the curvature of the ROC. When looking at pooled ROC points, it is therefore important to consider the between-subject variation in decision criteria. Plotting ROC curves based on confidence ratings now assumes that each level of the confidence rating could be a possible criterion and therefore each confidence level provides an ROC point (one of them is guaranteed to be at a HR and FAR of one, therefore k confidence levels result in k-1 meaningful ROC points). Figure 7 shows that for Experiment 2, the variation between the individual criteria is different between the confidence levels. Some of the ROC points based on pooled data should therefore be further away from the "true" ROC curve.

Figure 8 shows individual and pooled ROC points of Experiment 2 in comparison with the theoretical ROC curves based on the average d', d_a, and A'. As expected, particularly the two most liberal (i.e., rightmost) ROC points fall below the theoretical ROC curves.

To test whether the deviation from the theoretical ROC curves could be the mere result of pooling, we ran a simulation. The simulation assumed that the ROC curve based on d_a with a slope parameter of 0.6 holds true for each individual and, for simplification, that individuals deviate normally from the mean d_a of Experiment 2 (M = 1.37) with the standard deviation of Experiment 2 (SD = 0.26). Additionally, for the criterion c_a of each confidence level, it was assumed that subjects vary normally around the group's average, and again, these parameters were estimated using Experiment 2. According to these assumptions 10,000 observations were created for each confidence level and pooled. The result of this quite simple simulation is also depicted in Figure 8 and falls close to the pooled ROC points from the original data. This suggests that the pooled ROC points might simply deviate from the ROC curve based on d_a because of the variation in the criterion and sensitivity between subjects (however, this does not, of course, prove that the pooled ROC curve would look like the ROC curve based on d_a if all pooling artifacts were eliminated).

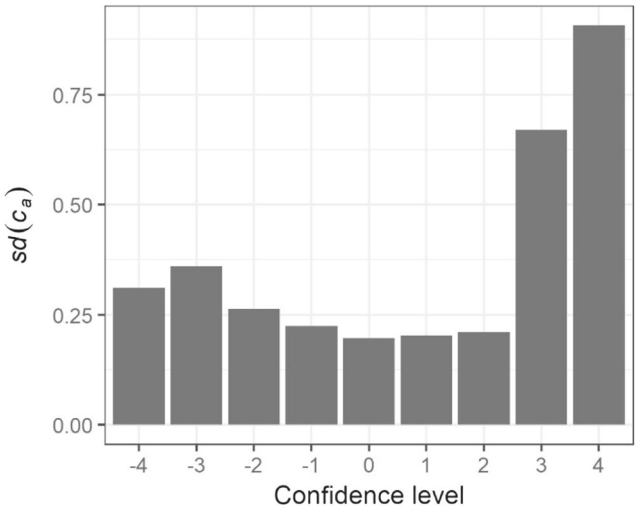

Fig. 7 Between-subject standard deviation of c_a (based on a slope of 0.6) for each confidence level

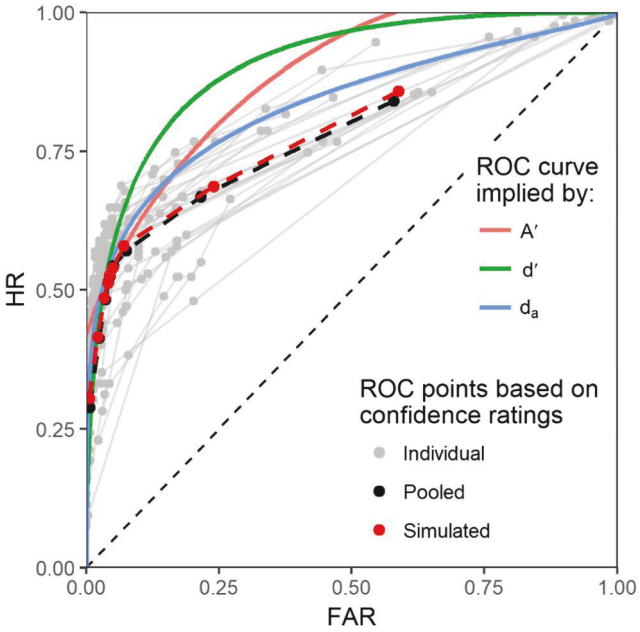

Fig. 8 Receiver operating characteristic (ROC) points based on individual (gray) and pooled confidence rating data of dataset 2 (black, dashed), created from a simulation (red, dashed), as assumed by the average d' (green), d_a (blue), and A' (red)

As illustrated, pooling ROC points can severely distort the shape of ROC curves. The illustrated problems of pooling should not occur if averaging is performed after z-transformation and the zROC curves are linear. However, z-transformation before pooling is often not fully possible because of FAR or HR values of zero or one on an individual level, for which the z-transformation (i.e., the inverse of the cumulative distribution function of the standard normal distribution) is undefined.

References

Appelbaum, L. G., Cain, M. S., Darling, E. F., & Mitroff, S. R. (2013). Action video game playing is associated with improved visual sensitivity, but not alterations in visual sensory memory. *Attention,*

Perception, & Psychophysics, *75*(6), 1161–1167. doi:https://doi.org/10.3758/s13414-013-0472-7

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. *Applied Cognitive Psychology*, *29*(1), 142–148. doi:https://doi.org/10.1002/acp.3083

Bolfing, A., Halbherr, T., & Schwaninger, A. (2008). How image based factors and human factors contribute to threat detection performance in X-Ray aviation security screening. In Holzinger A. (Ed.), HCI and usability for education and work. USAB 2008. Lecture Notes in Computer Science, vol 5298. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-89350-9_30

Box, G. E. P., & Draper, N. R. (1987). Empirical model building and response surfaces. New York, NY: John Wiley & Sons.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *35*(1), 11–21. doi:https://doi.org/10.3758/BF03195493

Brunstein, A., & Gonzalez, C. (2011). Preparing for novelty with diverse training. *Applied Cognitive Psychology*, *25*(5), 682–691. doi:https://doi.org/10.1002/acp.1739

Cain, M. Adamo, S. H., & Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Visual Cognition*, *21*(7), 899–921. doi:https://doi.org/10.1080/13506285.2013.843627

Chen, W., & Howe, P. D. L. (2016). Comparing breast screening protocols: Inserting catch trials does not improve sensitivity over double screening. *PLOS ONE*, *11*(10). doi:https://doi.org/10.1371/journal.pone.0163928

Commission Implementing Regulation (EU) (2015). Laying down detailed measures for the implementation of the common basic standards on aviation security 2015/1998 of 5 November 2015. Official Journal of the European Union.

Cooke, N. J., & Winner, J. L. (2007). Human factors of homeland security. *Reviews of Human Factors and Ergonomics*, *3*(1), 79–110. doi:https://doi.org/10.1518/155723408X299843

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), 1–36. doi:https://doi.org/10.1167/11.5.14

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. doi:https://doi.org/10.2307/2289144

Evans, K. K., Tambouret, R. H., Evered, A., Wilbur, D. C., & Wolfe, J. M. (2011). Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. *Archives of Pathology & Laboratory Medicine*, *135*(12), 1557–1560. doi:https://doi.org/10.5858/arpa.2010-0739-OA

Evered, A., Walker, D., Watt, A. A., & Perham, N. (2014). Untutored discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology*, *122*(3), 200–210. doi:https://doi.org/10.1002/cncy.21370

Gescheider, G. A. (1997). Psychophysics: The fundamentals. Mahwah, NJ: L. Erlbaum Associates.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500–513. doi:https://doi.org/10.1037/0278-7393.25.2.500

Godwin, H. J., Menneer, T., Cave, K. R., & Donnelly, N. (2010a). Dual-target search for high and low prevalence X-ray threat targets. *Visual Cognition*, *18*(10), 1439–1463. doi: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 Psychologica*, *134*(1), 79–84. doi:https://doi.org/10.1016/j.actpsy.2009.12.009

Godwin, H. J., Menneer, T., Cave, K. R., Thaibsyah, M., & Donnelly, N. (2015). The effects of increasing target prevalence on information processing during visual search. *Psychonomic Bulletin & Review*, *22*(2), 469–475. doi:https://doi.org/10.3758/s13423-014-0686-2

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. New York, NY: Wiley.

Halbherr, T., Schwaninger, A., Budgell, G. R., & Wales, A. W. J. (2013). Airport security screener competency: A cross-sectional and longitudinal analysis. *The International Journal of Aviation Psychology*, *23*(2), 113–129. doi:https://doi.org/10.1080/10508414.2011.582455

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. doi:https://doi.org/10.3758/BF03203619

Hirshman, E., & Hostetter, M. (2000). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. *Memory & Cognition*, *28*(2), 161–166. doi:https://doi.org/10.3758/BF03213795

Hofer, F., & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *Proceedings of the 38th IEEE International Carnahan Conference on Security Technology*, 303–308. doi:https://doi.org/10.1109/CCST.2004.1405409

Huang, L., & Pashler, H. (2005). Attention capacity and task difficulty in visual search. *Cognition*, *94*(3), B101–B111. doi:https://doi.org/10.1016/j.cognition.2004.06.006

Ishibashi, K., & Kita, S. (2014). Probability cueing influences miss rate and decision criterion in visual searches. *I-Perception*, *5*(3), 170–175. doi: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, & Psychophysics*, *74*(1), 115–123. doi: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. doi: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. *Journal of Transportation Security*, *1*(2), 81–106. doi:https://doi.org/10.1007/s12198-007-0006-4

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, *13*(3), 175–181. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/711391

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

Madhavan, P., Gonzalez, C., & Lacson, F. C. (2007). differential base rate training influences detection of novel targets in a complex visual inspection task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(4), 392–396. doi:https://doi.org/10.1177/154193120705100451

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. doi:https://doi.org/10.1080/00140130802376059

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? *Proceedings of the 45th IEEE International Carnahan Conference on Security Technology*. doi:https://doi.org/10.1109/CCST.2011.6095881

Mendes, M., Schwaninger, A., & Michel, S. (2013). Can laptops be left inside passenger bags if motion imaging is used in X-ray security screening? *Frontiers in Human Neuroscience*, *7*(October), 1–10. doi:https://doi.org/10.3389/fnhum.2013.00654

Menneer, T., Donnelly, N., Godwin, H. J., & Cave, K. R. (2010). High or low target prevalence increases the dual-target cost in visual search. *Journal of Experimental Psychology: Applied*, *16*(2), 133–144. doi:https://doi.org/10.1037/a0019569

Metz, C. E., Herman, B. A., & Shen, J. H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*(9), 1033–1053.

Miyazaki, Y. (2015). Influence of being videotaped on the prevalence effect during visual search. *Frontiers in Psychology*, *6*. doi:https://doi.org/10.3389/fpsyg.2015.00583

Nakashima, R., Watanabe, C., Maeda, E., Yoshikawa, T., Matsuda, I., Miki, S., & Yokosawa, K. (2015). The effect of expert knowledge on medical search: Medical experts have specialized abilities for detecting serious lesions. *Psychological Research*, *79*(5), 729–738. doi:https://doi.org/10.1007/s00426-014-0616-y

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *Radiographics : A Review Publication of the Radiological Society of North America, Inc*, *7*(6), 1241–1250. doi:https://doi.org/10.1148/radiographics.7.6.3423330

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.

Pepe, M., Longton, G., & Janes, H. (2009). Estimation and comparison of receiver operating characteristic curves. *The Stata Journal*, *9*(1), 1. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20161343

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'e. *Psychological Bulletin*, *71*(3), 161–173. doi:https://doi.org/10.1037/h0026862

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

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763–785. doi:https://doi.org/10.1037/0278-7393.20.4.763

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision*, *8*(15). doi:https://doi.org/10.1167/8.15.15

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*, 1–18. doi:https://doi.org/10.3389/fnhum.2015.00439

Russell, N. C. C., & Kunar, M. A. (2012). Colour and spatial cueing in low-prevalence visual search. *The Quarterly Journal of Experimental Psychology*, *65*(July), 1327–1344. doi:https://doi.org/10.1080/17470218.2012.656662

Schwaninger, A. (2004). Computer based training: A powerful tool to the enhancement of human factors. *Aviation Security International*, *2*, 31–36.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *Proceedings of the 38th IEEE International Carnahan Conference on Security Technology*, 29–35. doi:https://doi.org/10.1109/CCST.2004.1405402

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners: Visual abilities & visual knowledge measurement. *IEEE Aerospace and Systems Magazine*, *20*(6), 29–35.

Schwaninger, A., Hardmeier, D., Riegelnig, J., & Martin, M. (2010). Use it and still lose it? *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *23*(3), 169–175. doi:https://doi.org/10.1024/1662-9647/a000020

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

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

Sterchi, Y., Hättenschwiler, N., Michel, S., & Schwaninger, A. (2017). Relevance of Visual Inspection Strategy and Knowledge about Everyday Objects for X-Ray Baggage Screening. *Proceedings of the 51th IEEE International Carnahan Conference on Security Technology, 23-26*. doi: https://doi.org/10.1109/CCST.2017.8167812

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception, & Psychophysics*, *71*(3), 541–553. doi: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. doi:https://doi.org/10.3758/BRM.41.3.937

Wickens, T. D. (2001). Elementary signal detection theory. New York, NY: Oxford University Press.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), Integrated models of cognitive systems (pp. 99–119). New York, NY: Oxford University Press.

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). doi:https://doi.org/10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare targets are often missed in visual search. *Nature*, *435*, 439–440. doi: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. *Journal of Experimental Psychology: General*, *136*(4), 623–638. doi: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. doi:https://doi.org/10.1016/j.cub.2009.11.066

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. Retrieved from doi:https://doi.org/10.1037/0033-2909.133.5.800

Yu, R., & Wu, X. (2015). Working alone or in the presence of others: Exploring social facilitation in baggage X-ray security screening tasks. *Ergonomics*, *58*(6), 857–865. doi:https://doi.org/10.1080/00140139.2014.993429

Relevance of Visual Inspection Strategy and Knowledge about Everyday Objects for X-Ray Baggage Screening

Yanik Sterchi*, Nicole Hättenschwiler, Stefan Michel and Adrian Schwaninger
School of Applied Psychology
University of Applied Sciences and Arts Northwestern Switzerland (FHNW)
Olten, Switzerland

*Email: yanik.sterchi@fhnw.ch

***Abstract*—The screening of passenger bags at airports can be understood as a visual inspection task that consists of visual search and decision. Security officers (screeners) visually search for prohibited items in X-ray images and decide whether secondary search (e.g. using manual search or explosive trace detection) is needed. A screener's decision can be explained with signal detection theory and its measures (hit rate, false alarm rate, sensitivity and decision criterion). In this experiment tested whether a specifically instructed visual inspection strategy can influence the hit and false alarm rate. In addition, it was investigated whether knowledge about the visual appearance of harmless everyday objects in X-ray images is relevant for the detection of prohibited items. To this end, 31 screeners of an international airport conducted a simulated X-ray baggage screening task with two different instructions (normal vs. liberal decision) on how to conduct visual inspection: In the normal decision condition, screeners were instructed to visually inspect the X-ray images like they were used to from their job. In the liberal decision condition, screeners were instructed to visually analyze each object in the X-ray image and only decide that the bag was harmless if each object in the image could be recognized as harmless The screeners knew half of the prohibited items from computer-based training while the other half were novel prohibited items. In addition, knowledge about the visual appearance of everyday objects in X-ray images was measured. The results show that screeners were able to change their decision criterion depending on the instructed visual inspection strategy. Knowledge about harmless everyday objects was positively associated with detection performance and most notably correlated with the hit rate for novel threat items in the liberal decision condition. Implications for improving X-ray screening at airports using a risk-based and adaptive approach are discussed.**

Keywords—aviation security, detection performance, everyday object recognition, visual inspection, visual search, X-ray screening

I. INTRODUCTION

Secure air transportation is essential for economy and society. Over the past decades, airports and governments have invested heavily into further development of airport security checkpoints. At these checkpoints, airport security officers (screeners) visually inspect passenger baggage with X-ray screening technology to make sure that no prohibited items (IEDs: improvised explosive devices, knives, guns, and other prohibited items) can enter the security restricted area of an airport.

Initial and recurrent training to detect known and novel prohibited items in X-ray images is an essential factor for screener performance. Several studies have shown the importance of computer-based training to learn which items are prohibited and what they look like in X-ray images, e.g. [1]–[3]. In addition to these so-called *knowledge-based factors*, studies also show the relevance *image-based factors* (rotation of the prohibited item, superposition by other items, complexity of the bag) for X-ray image inspection, e.g. [4], [5].

Screening of passenger bags can be understood as a visual inspection task that consists of visual search and decision [2], [6] inspired by the work of Spitz and Drury [7]. The decision whether an X-ray image of a passenger bag contains a prohibited item or not can be described with signal detection theory (SDT) [8], [9]. Important measures in this context are the hit rate (share of passenger bags with prohibited items correctly classified as containing prohibited items), and the false alarm rate (share of harmless bags falsely classified as containing prohibited items). SDT assumes that the hit and false alarm rate of a person result from his or her *sensitivity* and *criterion*. *Sensitivity* is the ability to differentiate between *noise* (in our case the harmless bag containing everyday objects) and *signal plus noise* (bag containing a prohibited item and everyday objects). The *criterion* is the response tendency that is assumed to be independent from sensitivity. A more *conservative* criterion is a tendency towards deciding in favor of noise, resulting in fewer false alarms but also fewer hits. A more *liberal* criterion is a tendency towards deciding in favor of *signal plus noise*, resulting in more hits but also more false alarms (see Fig. 1A). So the assumption is that for a given sensitivity, the criterion can be changed, leading to a change in both the hit and false alarm rate in the same direction. The thereby possible pairs of hit and false alarm rate are described by the so-called *receiver operating characteristic curve* (ROC curve; Fig. 1B).

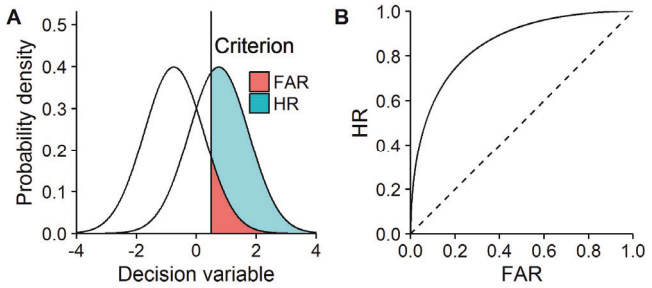

Fig. 1. Illustration of SDT. A: Noise and signal plus noise distribution, decision criterion and resulting hit rate (HR) and false alarm rate (FAR). B: Receiver operating characteristic curve resulting from shifting the criterion in Fig. 1A.

Measures used to estimate sensitivity and criterion are often derived from one hit and false alarm rate value (e.g. d' or A'). These measures assume a specific shape of ROC curve (for more information on detection measures and implied ROC curves see [8], [9]). However, some studies in the last ten years indicate that these assumptions might not apply to visual inspection of X-ray images [10]–[13]. As an alternative to the one-point measures, confidence ratings allow estimation of empirical ROC curves and use of the area under the curve (AUC) as sensitivity measure [8].

Sensitivity is high if the person who visually inspects X-ray images knows which items are prohibited and what they look like in X-ray images [1]–[3]. Knowing what everyday objects look like in X-ray images could further facilitate the differentiation between harmless and prohibited items as recently found by Hättenschwiler et al. [14]. The authors revealed a negative correlation between everyday object knowledge measured in an X-ray object categorization and naming test and false alarm rate in a simulated X-ray baggage screening task. An intuitive explanation of this result could be that once an item is identified as harmless, it can no longer be mistaken for a threat item and thereby not result in a false alarm. This assumption implies that screeners search an X-ray image and decide for one object after another whether it is harmless or not, in accordance with the model proposed by Wolfe and Van Wert [15]. This model is similar to the two-component model by [7], in which search continues until an inspector either finds what she or he is looking for (e.g. a prohibited item) or determines that enough time has been spent searching.

From an efficiency perspective, a low false alarm rate is desirable, as each false alarm requires resources for its resolution (e.g. using explosive trace detection and manual search of the bag). From a security effectiveness perspective, it would be interesting to investigate whether knowledge about everyday objects can also be used to increase the hit rate. According to SDT, this should be possible, if screeners can apply a more *liberal* criterion, i.e. increase their tendency to classify a bag as needing alarm resolution. This should increase both hit and false alarm rates. Assuming that the overall decision for a bag is based on decisions on the level of single objects within the bag, a more liberal criterion on bag level results from a more liberal criterion on the level of single items in the bag [15].

Based on the assumptions above, knowledge about everyday objects could be especially relevant for the detection of prohibited items that the screeners have never seen before. Since they lack the knowledge about their appearance (knowledge based factors), such novel prohibited items are harder to detect, when they less resemble known prohibited items. It is possible that screeners with good knowledge about everyday objects can detect novel prohibited items by an exclusion principle: They could only declare a bag as harmless if all contained objects are identified as harmless everyday objects, which in terms of SDT means the application of a very liberal decision strategy. If screeners can successfully be instructed to apply such a liberal decision criterion, this could allow for interesting practical applications, e.g. for increased effectiveness when screening bags of high-risk passengers.

To our knowledge, there is no study yet that investigated the effects of instructing such an inspection strategy on detection performance. We therefore pursue this question in this exploratory study.

II. METHOD AND PROCEDURE

A. Participants

A total of 31 screeners from one international airport completed this experiment (one participant dropped out after the first test due to illness). They were all certified screeners, meaning they were qualified, trained and certified according to the standards set by the appropriate national authority (civil aviation administration) consistent with European Regulation [16]. The participants were between 26 and 61 years old (M = 45.4, SD = 8.9) and had between 2 and 26 years of work experience (M = 8.4, SD = 5.5). 64.5% were female.

B. Experimental Design

The experiment used a mixed factorial design with two differently instructed inspection strategies (normal decision vs. liberal decision) as within-subjects factor and training of a new inspection instruction (short e-learning module vs. instruction only) as between-subjects factor. Since we were not sure whether the participants could apply the liberal decision strategy after only receiving a short instruction, the screeners were allocated into two groups. In addition to the instruction, one group received a short e-learning module explaining the liberal decision strategy in more detail to assist with switching from the normal decision strategy to the liberal decision strategy.

Performance measures and eye tracking data were calculated as dependent variables. Performance was assessed in terms of effectiveness (percentage detection of prohibited items, hit rate) and efficiency (false alarm rate, response times and scan paths).

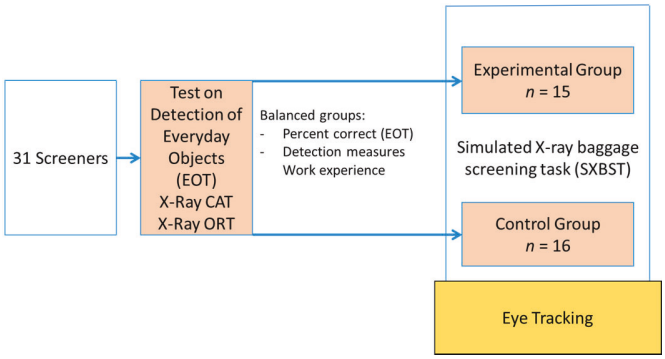

Fig. 2. Illustration of the study design.

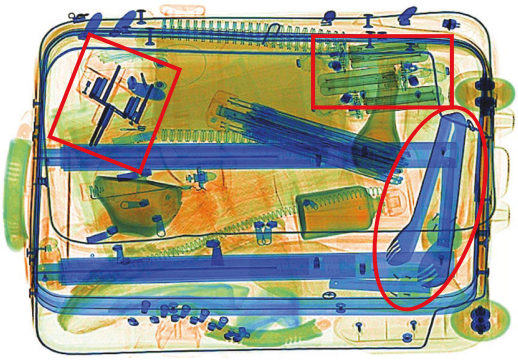

Fig. 3. Screenshot showing an X-ray image of a passenger bag from the everyday objects test with framed harmless everyday objects.

C. Procedure

All participants came to the test facilities twice. At the first test date, all screeners completed the same pre-tests (test on detection of everyday objects, X-Ray CAT [17] and X-Ray ORT [18]) to get an indication of their visual search performance. They were then divided into two groups counterbalanced regarding their detection performance scores and work experience. In the second test session, screeners were assigned to a simulated X-ray baggage screening task (SXBST) using eye tracking. One group completed an e-learning module right before starting the SXBST while the other group directly started with the SXBST.

In the normal decision condition, screeners were instructed to visually inspect the X-ray images like they were used to from their job. In the liberal decision condition, screeners were instructed to visually analyze each object in the X-ray image and decide that the bag is NOT OK if at least one object could not be recognized as harmless.

D. Materials

a) Everyday objects test (EOT): The EOT contains 32 X-Ray images of cabin baggage. In each image, three objects per bag were marked with a red frame (Fig. 3). Out of these objects, 17 were prohibited items out of the categories IEDs or other prohibited items and 79 were everyday objects. This resulted in 19 X-ray images containing only harmless everyday objects, nine X-ray images containing two harmless objects and one prohibited item, and four X-ray images containing one harmless object and two prohibited items. To solve the test, three items per X-ray image had to be categorized and named. For each item, participants had to click on one of three option buttons describing the categories: harmless everyday object, IED, and other prohibited item (e.g. gun, knife, electric shock device, etc.). After categorizing an object, participants had to enter the name of the object into a textbox and rate how confident they were in their decision. In case an object could not be named, participants left the corresponding textbox empty. There was no time limit and completing the test took about 45-60 minutes.

b) E-learning Module: The e-learning module consisted of a short definition of the new inspection strategy liberal decision followed by some examples with feedback. Screeners needed approx. 10 minutes to complete the module.

c) Simulated X-ray Baggage Screening Task (SXBST): 128 color X-ray images of passenger bags were selected by X-ray screening experts. In half of the X-ray images, one prohibited item was added using a validated X-ray image blending software [19]. Four categories of prohibited items were used (guns, knives, IEDs and other prohibited items). For each category, eight exemplars were used. Each exemplar was displayed once in canonical view (as defined by the two X-ray screening experts and the authors) and once rotated (around the horizontal or vertical axis by 85 deg.). For each category, half of the prohibited items were part of the training system used at this airport (*known items*). The other half were newly recorded (*novel items*) and visual comparison was used to make sure that they were different from the prohibited items contained in the training system. SXBST trials were structured as follows: After a fixation cross had to be fixated for 1.5 seconds, the X-ray images were displayed on the screen without time limit and screeners had to decide whether it was harmless or not by pressing a key, followed by confidence ratings on a scale from 0 to 10. The test was divided into four blocks. For two blocks screeners were instructed to visually inspect the X-ray images like they were used to from their job (*normal decision*). For the other two blocks, screeners were instructed to visually analyze each object in the X-ray image and only decide that the bag was harmless if each object in the image could be recognized as harmless (*liberal decision*). The order of the blocks was counterbalanced. There was no feedback on the correctness of responses and the participants took about 30 minutes to complete the test.

E. Eye Tracking Apparatus

Eye tracking was conducted using the SMI RED-m eye tracker with a gaze sample rate of 120 Hz, a gaze position accuracy of 0.5° and a spatial resolution of 0.1°. This non-invasive, video-based eye tracker was attached to a 22-inch screen that was placed 50 to 75 cm from the participant. The RED-m tracks both eyes (binocular) and works with two

infrared light sources, the reflection of which from the retina is recorded by a camera. Consequently, the participants could move freely in the limited area that the tracking system can record accurately. Two screen monitors were attached to a laptop: one showing the X-ray images to the participant, the other one showing the eye movements simultaneously to the facilitator.

F. *Analyses*

Confidence ratings were used to calculate AUC with the R-package pROC [20], [21]. All dependent variables were aggregated on individual level before statistical analysis. Since the dependent variables were substantially dispersed and not normally distributed, within-subject comparisons were tested for significance with the Wilcoxon signed-rank test and between-subject comparisons with the Mann-Whitney test.

III. Results

Fig. 4 displays the hit rate for novel and known prohibited items and the false alarm rate. As expected, the hit rate was higher for known than for novel prohibited items, $W = 1934$, $p < .001$. In comparison to *normal decision*, *liberal decision* resulted in a higher hit rate for known prohibited items, $W = 85$, $p = .02$, and for novel prohibited items, $W = 95.5$, $p = .02$. In addition, the false alarm rate was significantly higher, $W = 60.5$, $p < .001$. In contradiction to our expectation, these effects were not significantly larger for the group who received the e-learning, neither for the hit rate of known items, $U = 145$, $p = .16$, novel items, $U = 141.5$, $p = .20$, nor the false alarm rate, $U = 141$, $p = .21$.[1]

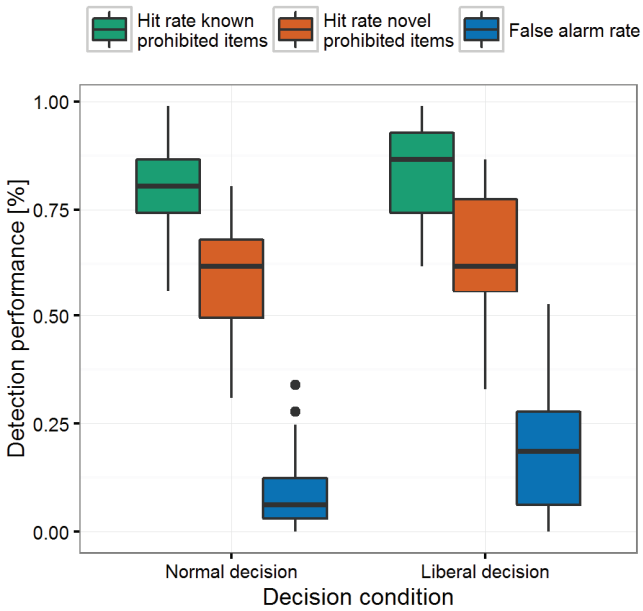

Fig. 4. Box plots of hit and false alarm rates depending on decision condition and prohibited item class (known vs. novel). (*Note: Performance values are multiplied by an arbitrary constant for security purposes.*)

Sensitivity (AUC values) did not differ between the two inspection strategies, neither for known, $W = 240$, $p = .88$ (normal decision: $M = .889$, $SD = .066$; liberal decision: M = .890, SD = .068) nor for novel items, $W = 262.5$, $p = .78$ (normal decision: $M = .794$, $SD = .079$; liberal decision: M = .789, SD = .067).

For the analysis of the eye tracking data, five participants had to be excluded due to technical difficulties that led to 20-100% of their trials without any recorded saccades or fixations. From the remaining participants, 38 of 3316 trials had to be excluded (again due to the lack of any saccades or fixations being recorded in these trials). Not surprisingly, the overall increased response times in the liberal decision condition were associated with on average (mean) 22% longer scan paths (measured in pixels) for target present trials, $W = 75$, $p = .009$, and 28% longer scan paths for target absent trials, $W = 49$, $p < .001$. However, this increase was disproportionate for target absent trials, leading to a shorter average scan path per response time, $W = 292$, $p = .002$, but not significantly so for target present trials, $W = 228$, $p = .23$. We further analyzed whether this slower scanning might be due to more frequent fixations or longer fixations, revealing that the number of fixations per response time actually decreased for target absent trials, $W = 271$, $p = .014$, but the average duration of these fixations increased, $W = 38$, $p < .01$.

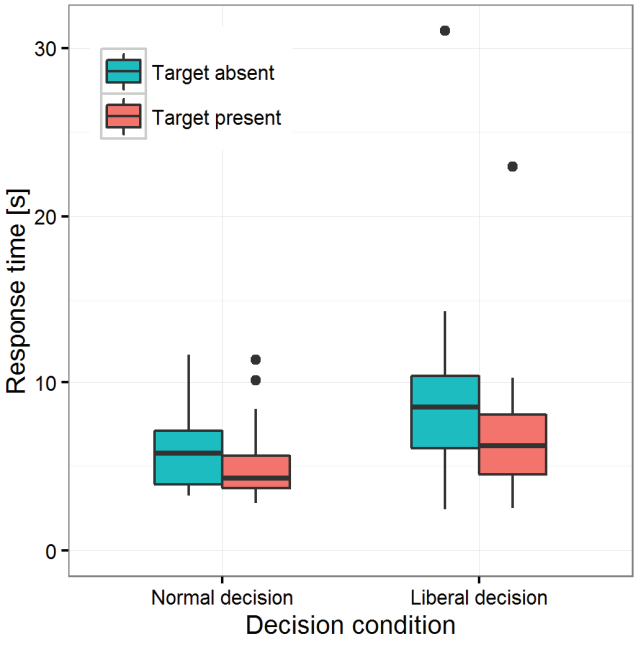

Fig. 5. Boxplot of individual median response times [s] depending on decision condition and separate for target absent and target present trials.

[1] We also analyzed whether e-learning affected sensitivity. There was no significant difference in AUC between the two groups neither for known, $U = 138.5$, $p = .48$ (e-learning: $M = .896$, $SD = .059$; control group: $M = .883$, $SD = .060$), nor novel items, $U = 121$, $p = .98$ (e-learning: $M = .787$, $SD = .059$; control group: $M = .789$, $SD = .051$)

In a next step, we analyzed whether everyday object knowledge was associated with a higher hit rate and a lower false alarm rate. Table I. shows based on rank correlations[2] that when instructed for *normal decision,* screeners with a high performance in the EOT also detected more known prohibited items, had a marginally significant lower false alarm rate, but did not detect more novel items. Looking at the condition *liberal decision*, the pattern changes: EOT performance was not associated with lower false alarm rates but with higher hit rates for novel prohibited items.

TABLE I. CORRELATIONS

Correlations EOT[a] and SXBST[b]	**SXBST[b] variable**		
	HR[c] known prohibited items	***HR[c] novel prohibited items***	***False alarm rate***
Normal decision	r_s = .430 p = .008	r_s = -.117 p = .735	r_s = -.298 p = .052
Liberal decision	r_s = .391 p = .015	r_s = .322 p = .038	r_s = -.018 p = .462

[a.] Everyday object test score
[b.] Simulated X-ray baggage screening task
[c.] Hit rate

IV. DISCUSSION

In our experiment, we investigated whether screeners can be instructed to apply a more liberal decision criterion when visually inspecting X-ray images of passenger bags resulting in higher hit rates at the cost of increased false alarm rates. Further, we explored whether knowledge about the appearance of everyday objects in X-ray images was associated with detection performance. The results show that an instruction to decide more liberal led to increased hit and false alarm rates. Sensitivity – estimated with the AUC based on confidence ratings – remained constant for the two inspection strategies. This implies that the observed change in hit and false alarm rates was due to a change in the decision criterion. We therefore conclude that screeners are generally capable to shift their criterion based on an instruction. However, this criterion shift also led to longer response times, especially for target absent trials, which are most relevant in practice, where the majority of images do not contain any prohibited items. It is not surprising that participants need more time when instructed to decide carefully for each object whether it is harmless or not. In this regard, it should also be noted that SDT does not explain how response times are linked to sensitivity or the criterion [22]. Reference [15] also found a criterion shift as the result of changes in target prevalence (share of target present trials) to influence response times without a change in sensitivity. In their proposed model, they explain the overall criterion as the result of the decision criterion on the level of single objects within the X-ray image (as already mentioned in the introduction) and in addition assume a quitting threshold. The assumption is that participants continue searching until they either come across an item that requires further inspection or until their quitting threshold is satisfied, which thereby governs the response time for target absent responses. As explained in the introduction, this is comparable to the model of [7]. Also [10], [12], [15] found response times to be longer when participants had a more liberal criterion due to higher target prevalence. Our results hence fall in line with criterion shifts induced by different levels of target prevalence.

The eye tracking data from our experiment shows that for images of harmless bags screeners have longer scan paths and more fixations. Nevertheless, at the same time scanning was slower and fixations longer. This suggests that applying the liberal decision not only extended the search duration but also affected underlying cognitive processes, e.g. [23].

In our experiment, we also investigated whether an e-learning module could assist with the application of the new inspection strategy. However, the liberal decision condition did not have a stronger effect for the e-learning group, neither for their hit rates, false alarm rates nor response times. This means that the e-learning module, as designed for this experiment, was not effective or necessary, since screeners without e-learning were also able to shift their criterion based on the instruction. Further, the e-learning module did also not interact with the effect of decision strategy on response times.

In the normal decision condition, screeners with more everyday object knowledge had lower false alarm rates (though only marginally significant), which is in line with the findings of [14]. In both decision conditions, screeners with more everyday object knowledge had higher hit rates for prohibited items known from computer-based training, possibly because these screeners had both more knowledge about everyday objects and about prohibited items included in training. Interestingly, when applying the liberal decision strategy, screeners with more everyday object knowledge no longer had lower false alarm rates but had higher hit rates for novel prohibited items. This is a first indication that good knowledge about the visual appearance of everyday objects might be useful for better detection of novel prohibited items.

V. SUMMARY, CONCLUSIONS AND LIMITATIONS

Our results show that the instruction of a more *liberal decision* for visual inspection of X-ray images led to an increased hit and false alarm rate without affecting sensitivity. This implies that the observed change in hit and false alarm rates was due to a change in the decision criterion alone. These findings are consistent with understanding visual inspection of X-ray images as a task consisting of visual search and decision, where the decision is made according to signal detection theory. Regarding practical implications, the instruction of visually inspecting X-ray images using a liberal decision on a *regular* basis is not advised because of increased false alarm rates and slower response times, which would reduce efficiency of X-ray screening at security checkpoints. However, visual inspection using a liberal decision strategy could be very valuable for increased effectiveness when screening bags of high-risk passengers and/or flights. This would be particularly useful to increase detection of novel threat items.

Since our findings regarding everyday object knowledge was merely correlative, future studies are needed to prove that everyday object knowledge can decrease false alarm rates and

[2] Due to the lack of linearity between the variable pairs, we refrained from Pearson correlations.

increase hit rates depending on decision strategy. In that case, a specific training of everyday objects would have a high potential to increase effectiveness and efficiency of X-ray screening.

REFERENCES

[1] S. M. Koller, D. Hardmeier, S. Michel, and A. Schwaninger, "Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-Ray image interpretation," *J. Transp. Secur.*, vol. 1, no. 2, pp. 81–106, 2008.

[2] S. M. Koller, C. G. Drury, and A. Schwaninger, "Change of search time and non-search time in X-ray baggage screening due to training.," *Ergonomics*, vol. 52, no. 6, pp. 644–56, 2009.

[3] T. Halbherr, A. Schwaninger, G. R. Budgell, and A. Wales, "Airport Security Screener Competency: A cross-sectional and longitudinal analysis," *Int. J. Aviat. Psychol.*, vol. 23, no. 2, pp. 113–129, 2013.

[4] A. Schwaninger, D. Hardmeier, and F. Hofer, "Measuring visual abilities and visual knowledge of aviation security screeners," *Proc. 38th Annu. Int. Carnahan Conf. Secur. Technol.*, pp. 29–35, 2004.

[5] A. Bolfing, T. Halbherr, and A. Schwaninger, "How image based factors and human factors contribute to threat detection performance in x-ray aviation security screening," *HCI and Usability for Educ. and Work, Lecture Notes in Comput. Sci.*, 5298, pp. 419-438, 2008.

[6] A. W. J. Wales, C. Anderson, K. L. Jones, A. Schwaninger, and J. A. Horne, "Evaluating the two-component inspection model in a simplified luggage search task.," *Behav. Res. Methods*, vol. 41, no. 3, pp. 937–943, 2009.

[7] G. Spitz and C. G. Drury, "Inspection of sheet materials - Test of model predictions," *Hum. Factors*, vol. 20, no. 5, pp. 521–528, 1978.

[8] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*, 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2005.

[9] D. M. Green and J. A. Swets, S*ignal detection theory and psychophysics*. New York: Wiley, 1966.

[10] H. J. Godwin, T. Menneer, K. R. Cave, and N. Donnelly, "Dual-target search for high and low prevalence X-ray threat targets," *Vis. cogn.*, vol. 18, no. 10, pp. 1439–1463, 2010.

[11] M. J. Van Wert, T. S. Horowitz, and J. M. Wolfe, "Even in correctable search, some types of rare targets are frequently missed," *Attention, Perception, Psychophys.*, vol. 71, no. 3, pp. 541–553, 2009.

[12] J. M. Wolfe, T. S. Horowitz, M. J. Van Wert, N. M. Kenner, S. S. Place, and N. Kibbi, "Low target prevalence is a stubborn source of errors in visual search tasks," *J. Exp. Psychol. Gen.*, vol. 136, no. 4, pp. 623–638, 2007.

[13] J. S. H. Lau and L. Huang, "The prevalence effect is determined by past experience, not future prospects," *Vision Res.*, vol. 50, no. 15, pp. 1469–1474, 2010.

[14] N. Hattenschwiler, S. Michel, M. Kuhn, S. Ritzmann, and A. Schwaninger, "A first exploratory study on the relevance of everyday object knowledge and training for increasing efficiency in airport security X-ray screening," *Proc. 38th Annu. Int. Carnahan Conf. Secur. Technol.*, pp. 25–30, 2015.

[15] J. M. Wolfe and M. J. Van Wert, "Varying target prevalence reveals two dissociable decision criteria in visual search," *Curr. Biol.*, vol. 20, no. 2, pp. 121–124, 2010.

[16] 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 J. of the European Union*.

[17] S. M. Koller and A. Schwaninger, "Assessing X-ray image interpretation competency of airport security screeners," *Proc. 2nd Int. Conf. Res. Air Transp.*, pp. 399–402, 2006.

[18] D. Hardmeier, F. Hofer, and A. Schwaninger, "Increased detection performance in airport security screening using the X-Ray ORT as pre-employment assessment tool," *Proc. 2nd Int. Conf. Res. Air Transp.*, pp. 393–397, 2006.

[19] M. Mendes, A. Schwaninger, and S. Michel, "Does the application of virtually merged images influence the effectiveness of computer-based training in x-ray screening?," *Proc. 45th Annu. Int. Carnahan Conf. Secur. Technol.*, pp. 1–8, 2011.

[20] X. Robin et al., "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[21] R Core Team, "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2016.

[22] T. J. Pleskac and J. R. Busemeyer, "Two-stage dynamic signal detection: A theory of choice, decision time, and confidence.," *Psychol. Rev.*, vol. 117, no. 3, pp. 864–901, Jul. 2010.

[23] R. N. Meghanathan, C. van Leeuwen, and A. R. Nikolaev, "Fixation duration surpasses pupil size as a measure of memory load in free viewing.," *Front. Hum. Neurosci.*, vol. 8, p. 1063, 2014.

Why stop after 20 minutes? Breaks and target prevalence in a one hour X-ray image inspection task

Daniela Buser*, Yanik Sterchi* and Adrian Schwaninger

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

Author Note

Correspondence concerning this article should be addressed to Daniela Buser, University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Riggenbachstrasse 16, CH-4600 Olten, Switzerland.

Email: daniela.buser@fhnw.ch, Phone: +41 62 957 27 92

*Joint first authorship between Daniela Buser and Yanik Sterchi

Abstract

Current EU regulations restrict the duration of X-ray inspection of passenger baggage at airport security checkpoints to 20 min as a precautionary measure to prevent performance decrements. However, this limitation to 20 min 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 baggage screening task. One group of screeners took 10-min breaks after 20 min of screening, whereas the other group worked without breaks. We varied target prevalence in order to determine a valid measure for detection performance in the X-ray inspection of passenger baggage. Results showed that d_a with a slope of 0.65 was a valid measure of performance. Moreover, there were no performance decrements over the course of 60 min. Breaks did not affect performance, but reduced the amount of subjective distress. However, there were high interindividual differences in the amount of distress reported by screeners working without breaks. These results provide a basis for designing field studies of prolonged screening durations, and open the discussion on whether to consider new break policies such as flexible work schedules.

Keywords: X-ray image inspection, visual search, time on task, breaks, detection measures, target prevalence effect.

1. Introduction

Throughout the world, passenger baggage is scanned at airports with X-ray machines, and security officers (screeners) inspect these X-ray images for prohibited items (guns, knives, bombs and other prohibited items). Current European regulations restrict X-ray image inspection of passenger baggage to a maximum of 20 min of continuous screening as a precautionary measure to prevent any decrease in detection performance (European Commission, 2015). Therefore, after 20 min of screening, screeners usually rotate to another

position at the airport security checkpoint where they carry out other tasks such as assisting passengers with loading trays or conducting body searches. A new technology called *remote cabin baggage screening* (RCBS), which is being employed increasingly by airports, creates operational challenges for the 20-min rule. With RCBS, security personnel visually inspect X-ray images in an office-like environment separate 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. Screening durations longer than 20 min would alleviate such concerns. 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. In another condition, screeners took 10-min breaks after each 20 min of screening. The following subchapters provide a theoretical overview on performance over time and the measurement of screener performance.

1.1. Performance decrements in X-ray screening

In X-ray image inspection screeners need to search for many different prohibited items among a great variety of harmless objects (Harris, 2002). Both prohibited and harmless objects look different in X-ray images from the way they look in reality, and usually many items superimpose each other in X-ray images (Schwaninger, 2003). Hence, it is not surprising that this task can be tiring, and this explains why European regulations limit X-ray image inspection to 20 min as a precautionary measure (European Commission, 2015).

However, the introduction of this limitation was not undisputed (Ref) and likely based on research into vigilance (personal communication with airport security expert), because research on this issue is scarce. Few studies exist investigating time on task in X-ray image inspection. . A study by Meuter and Lacherez examined the effect of time on task on screener performance in the field (Meuter & Lacherez, 2016). Said study analyzed 4 months of threat image projection (TIP) data from an Australian airport. TIP is a technology that projects prerecorded threat items onto real X-ray images of passenger baggage during baggage screening at airport security checkpoints (Cutler & Paddock, 2009; Hofer & Schwaninger, 2005). The TIP hit rate (or percent detected) refers to the proportion of projected fictional threat items that screeners have detected. 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 images screened per min during one session of continuous screening). 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 caveats when trying to use it to derive recommendations for time on task. First, the observed decrease in the hit rate was very small. Because the 20-min rule also applied to the screeners that participated in the study, it is unclear how performance would evolve for longer screening durations. It should also be noted that Meuter and Lacherez (2016) analyzed data from conventional 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. Another limitation of their study is that they were unable to analyze the false alarm rate, because the TIP system can tell only whether a rejected X-ray image contained a TIP, but not whether it contained an actual prohibited item. 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. A study conducted by Ghylin et al. (2007) provides more insight into this. In the study, airport security screeners completed a test with images of passenger carry-on bags over the course of four hours. Results were aggregated for each of the four hours. The study found significantly lower hit rates in the third and fourth hour compared to the first. For the false alarm rate, the difference between hour one and two, and hour two and four also attained significance. However, there was no significant change in the sensitivity measure *A'*. The mutual decline of hit and false alarm rate therefore suggests a shift in response tendency. Reaction times decreased from the first to the second and from the second to the third hour. The authors concluded that vigilance decrements occurred, which we will address in the next subchapter. Whereas the study of Ghylin et al. (2007) provides very interesting results, it only compares full hours and does not report hit rate, false alarm rate, or *A'* values. Conclusions about performance changes within the first hour of screening and an evaluation of the 20 min rule are therefore limited.

1.2. Performance decrements in vigilance tasks

The effect of time on task has been investigated quite extensively for vigilance tasks, which 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 & Parasuraman, 1982). In both tasks, the infrequent appearance of targets causes higher misses (Wolfe et al., 2007). In vigilance tasks, for very difficult tasks, performance decrements can already be observed as early as after 5 min (Nuechterlein, Parasuraman, & Jiang, 1983; Rose, Murphy, Byard, & Nikzad, 2002) Most studies have revealed decreases in vigilance within the first 15to30 min of the task (Mackworth, 1948; Teichner, 1974). Nonetheless, it is not clear whether the performance decrement within the first 15 to 30 min often found in

vigilance tasks can also be expected for X-ray image inspection, because the tasks differ in several respects. In vigilance tasks, a short distraction can lead to missing a target, whereas in an X-ray image inspection, one has to actively declare that no target is present in an image (Wolfe et al., 2007). Vigilance tasks present mostly one stimulus at a time, whereas detection tasks require the detection of a target among distractors (Wolfe et al., 2007). Also, in X-ray image inspection, only certain types of targets are very rare (e.g., bombs, guns), whereas other targets occur more frequently in carry-on baggage (e.g., bottles and laptops). In addition to the effect that time on task has on performance in vigilance tasks, research has also shown that people report more distress and less engagement after a vigilance task (Helton, 2004; Matthews et al., 2002).

1.3. The effects of breaks on performance

Further insight into the effect that time on task has on performance can be gained from inspecting research on the effect of breaks. Current research reveals mainly positive effects of breaks on the performance of a variety of different tasks (Arrabito, Ho, Aghaei, Burns, & Hou, 2015; Colquhoun, 1959; Kopardekar & Mital, 1994; Steinborn & Huestegge, 2016). Breaks have been found not only to have positive effects on performance but also to decrease subjectively perceived workload (Arrabito et al., 2015) as well as perceived fatigue and discomfort (Galinsky, Swanson, Sauter, Hurrell, & Schleifer, 2000). The frequency and length of breaks depends on the type, difficulty, and duration of the task (Tucker, 2003). Generally, short breaks can already help to restore attention and counter performance decline. Nonetheless, the current literature provides no clear indication on how frequent and long breaks should be for X-ray image inspection.

1.4. Measuring performance in X-ray image inspection

Some challenges emerge when investigating screener performance. Common measures for this task are the hit rate (percentage of detected prohibited items) and the false alarm rate (percentage of harmless baggage falsely sent to secondary search). Because the hit rate (HR) and false alarm rate (FAR) depend on the response tendency, it is recommended to use detection measures that are considered to be independent of response tendency (MacMillan & Creelman, 2013). Many of these detection measures are based on signal detection theory (SDT). This provides a general framework for defining detection performance, called *sensitivity*, independently from response tendency, called *criterion*.

Research in X-ray image inspection often uses d' as such a measure of sensitivity (Sterchi, Hättenschwiler, & Schwaninger, 2019)), which is calculated as

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

whereby z is the inverse of the cumulative distribution function of the standard normal distribution (Green & Swets, 1966)A commonly used measure for the criterion that is con

However, recent research has found d' to be invalid for X-ray image inspection. Several studies investigating the effect that the hit rate decreases when targets become 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 & 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, SDT assumes that target prevalence affects only the criterion and not sensitivity. Instead of assuming that sensitivity actually increases when target prevalence decreases,

Wolfe et al. (2007) have argued that X-ray image inspection does not fulfil the assumptions that underlie *d′*. SDT assumes a decision process in which target-present and target-absent trials each result in a distribution of evidence for the presence of a target, and *d′* assumes that these distributions have equal variance (MacMillan and Creelman, 2013). If the assumption of equal variance is not met, d_a offers an extension of *d′* with the slope *s* as an additional open parameter that is the ratio of the two standard deviations.

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

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 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 on X-ray image inspection (Godwin et al., 2010; Wolfe et al. 2007; Wolfe & Van Wert, 2010). Consistent with these findings, Sterchi et al. (2019) have reported slope parameters between 0.5 and 0.6 based on one experiment manipulating the criterion through instruction and another experiment using confidence ratings.

Our experiment investigates the target prevalence effect in order to determine which detection measure is valid when analyzing the effect that time on task has on detection performance. It is therefore also worth discussing how time on task and target prevalence interact. The target prevalence effect has been shown to depend on implicit learning: People rely more on the prevalence they experience rather than the prevalence they have been told to expect (Ishibashi, Kita, & Wolfe, 2012; Lau & Huang, 2010a)). Hence, the target prevalence

effect evolves over time, because people need to experience the prevailing prevalence (Ishibashi et al., 2012; Lau & Huang, 2010; Wolfe & Van Wert, 2010)).

In the current study, we investigated the effect of time on task on screener performance when X-ray images were analyzed for 60 min. One group screened for 60 min continuously, whereas the other group took 10-min breaks between 20-min screening blocks. Based on current evidence, we cannot formulate clear hypotheses on performance decrements depending on time on task. However, we assume, in line with previous research (Godwin et al., 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010), that *d'*, which assumes a slope parameter of 1, is an invalid measure of detection performance for this task and might therefore be affected by target prevalence. We expect the slope parameter to be around 0.6, and that d_a based on that slope will be more appropriate. We investigated this assumption by varying target prevalence. We also monitored the perceived stress of the task by asking screeners to complete the Short Stress State Questionnaire (SSSQ; Helton, 2004).

2. Methods

2.1. Participants

71 screeners working at a European airport completed the study. 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)[1], had 0.3 to 12 years of working experience (M = 2.08, SD = 2.23), and 46.38% of them were female. The study complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board of the University of Applied Sciences and Arts Northwestern Switzerland. Informed consent was obtained from all screeners prior to their participation.

[1] Two participants did not report their demographics.

2.2. Design

A 2 (break condition: with vs. without breaks) × 2 (prevalence condition: high vs. low prevalence) × 3 (time on task: 0–20 min, 20–40 min, 40–60 min) mixed factorial design was employed. The two break conditions *with breaks* and *without breaks* served as between-subject variable. All screeners completed the test twice, once in the low prevalence condition and once in the high prevalence condition (within-subject). 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 (Godwin et al., 2010) . Furthermore, it matches the prevalence of the screeners' training. In the low prevalence condition, one out of eight bags (12.5%) contained a prohibited item. This is higher than in practice, but it was necessary to collect enough target present trials within the experiment to calculate a reliable hit rate. The order of the two prevalence conditions was counterbalanced between test sessions. To analyze the effect of time on task, the test was broken down into three 20-min blocks: *0–20 min*, *20–40 min*, and *40–60 min*.

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 break condition and prevalence condition on the three factors of the SSSQ (distress, worry, engagement; Helton, 2004).

2.3. Materials

The test consisted of 864 X-ray images of passenger cabin (carry-on) baggage. For a subset of the images, prohibited items were merged into the bags using a validated X-ray image merging algorithm (Mendes, Schwaninger, & Michel, 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 test, each image of a passenger bag and each prohibited item appeared twice in the test. 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. Figure 1 shows two bags as an example. 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 (Schwaninger, 2003), were held at a medium level and not varied systematically.

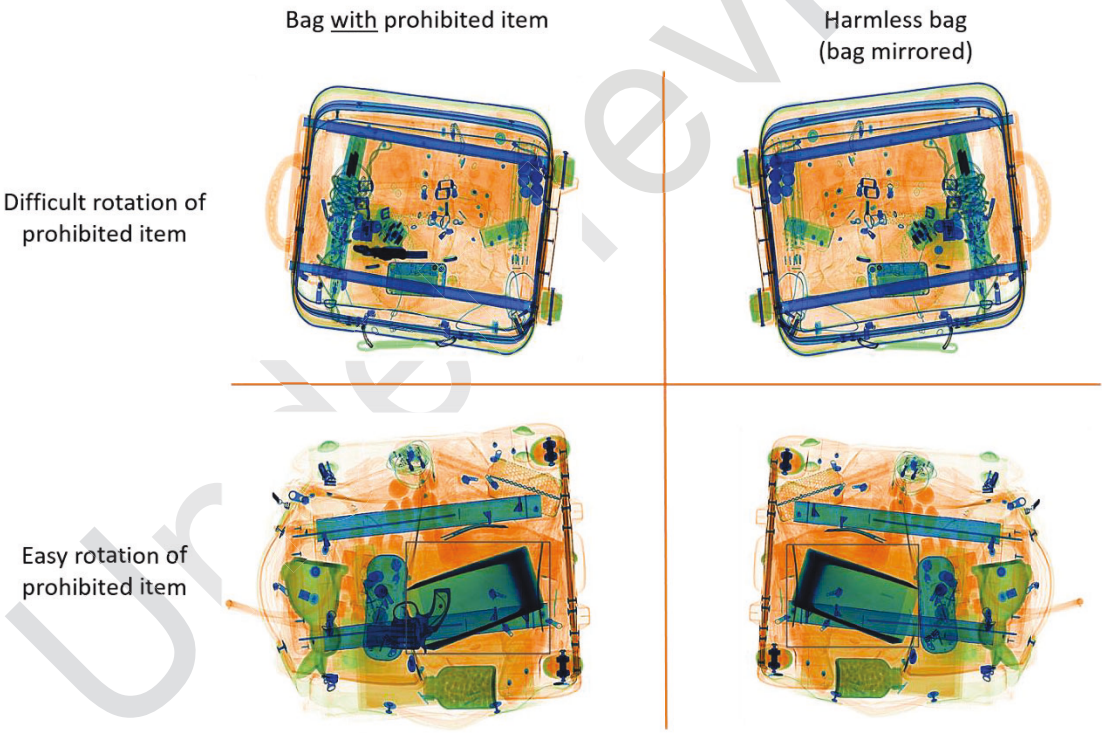

Fig. 1. Examples of images of passenger bags and prohibited items used in the test.

The test was so constructed as to allow all participants to be compared on the basis of the same images, regardless of their processing time and total amount of images analyzed during

the test. Therefore, the test was divided into 12 blocks, each containing 72 images in a fixed order. After analyzing images for 5 min, the system automatically switched to the next block. With a time restriction of 12 s per image, we expected each screener to analyze a minimum of 24 images within 5 min. Therefore, the first 24 images of each 5-min block were used to measure screening performance.

In the low-prevalence condition (12.5%), one gun, one knife, and one IED appeared among the first 24 images. In the high-prevalence condition (50%), four images per category were presented within the first 24 images of each block. The order of the 12 blocks was counterbalanced between participants with a Latin square design, ensuring that the images did not vary between the three 20-min blocks.

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

The airport's security provider scheduled groups of 6 to 12 screeners to participate in the experiment. These groups were then randomly assigned to either the group *with breaks* or the group *without breaks*.[2] Each participant completed the test twice, once with *low prevalence* and once with *high prevalence*. The two test times were separated by an interval of 3 to 5 weeks.

[2] A post-hoc comparison showed that the two groups were similar with regard to work experience, age, and gender.

Testing sessions took place in a training room at the airport. 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 of the respective test condition to avoid confusion.

Several groups belonging to the same break condition participated in one test session simultaneously. Due to organizational constraints, screeners from the group with breaks and screeners from the group without breaks had to complete the test simultaneously in the last two test sessions. Therefore, the participants in the two break conditions were separated spatially in the room. Each test session lasted about 1.5 hrs and was made up of three parts:

1. Practice trials. After receiving verbal and written instructions, screeners completed practice trials containing 16 images. They first completed these trials without time restriction, and then repeated the same training with the 12 s time restriction per image used in the actual test. The training was designed to familiarize participants with the interface and the procedure.

2. X-ray image inspection task. Screeners 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. For the X-ray image inspection, participants had to inspect the images as if they were working remotely. They were instructed 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*. Screeners had 12 s per image to decide whether the image was *OK* or *NOT OK*. Feedback was given in the same manner as provided by TIP systems: 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.

3. Questionnaire. After completing the X-ray image inspection 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 the first 24 images of each 5-min block and only images that appeared in both the high- and low-prevalence conditions were analyzed. For the descriptive statistics and analyses, dependent variables were first aggregated separately for each participant, and then separately for each 20-min block and prevalence condition.

The hit rate was calculated as the share of images correctly 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 & 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 of the images 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 & Geisser, 1959) was used where applicable and effect sizes are reported with η^2_p (partial eta squared). Cronbach alpha was calculated for the first test time point and each factor of the SSSQ seperately. In case of significant interactions between target prevalence and time on task, post hoc analyses were calculated comparing the first block (*0–20 min*) with the second block (*20–40 min*) and the second block with the third block (*40–60 min*) for both levels of target prevalence separately. Post hoc tests were Holm–Bonferroni corrected (Holm, 1979).

3. Results

Four participants were not able to take part in the second testing session and were therefore excluded from the analysis. 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 (PT); 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 PT as dependent variables.

3.1. Hit rate and false alarm rate

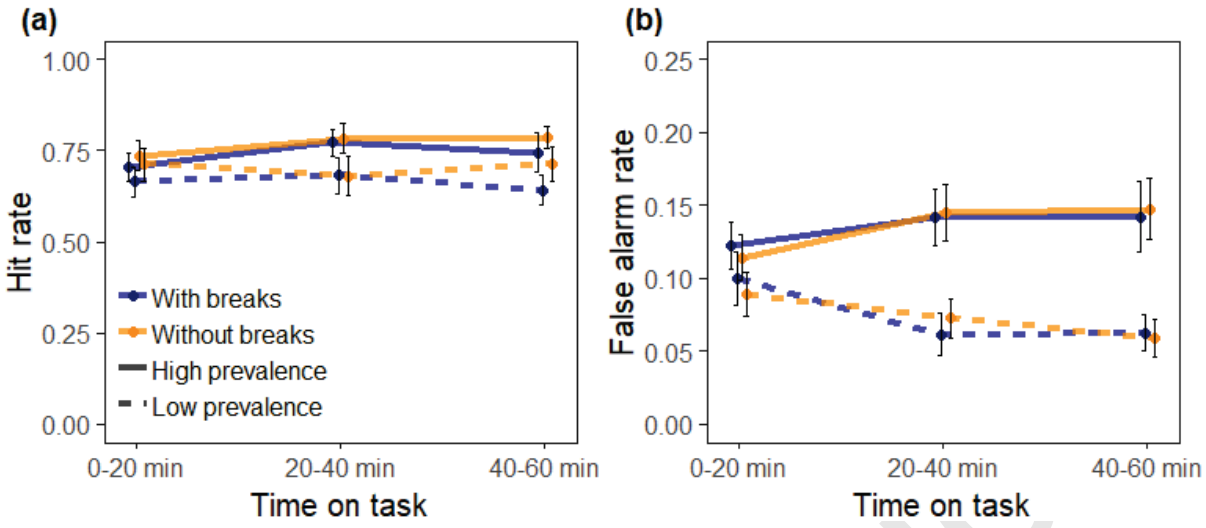

Fig. 2. Hit rate (a) and false alarm rate (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task. Standard errors are represented in the figure by the error bars.

Figure 2a shows the hit 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, $F(1, 69) = 37.99$, $p < .001$, $\eta_p^2 = .36$; but no effect of break, $F(1, 69) = 1.84$, $p = .180$, $\eta_p^2 = .03$, or time on task, $F(1.93, 133.17) = 1.78$, $p = .174$, $\eta_p^2 = .03$. None of the two-way interactions were significant: Break × Prevalence, $F(1, 69) = 0.25$, $p = .621$, $\eta_p^2 = .00$; Break × 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 did not attain significance either, $F(1.96, 134.94) = 0.31$, $p = .731$, $\eta_p^2 = .00$.

Figure 2b shows the false alarm rate for both break conditions and prevalence conditions as a function of time on task. 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 break, $F(1, 69) = 0.00$, $p = .957$, $\eta_p^2 = .00$; and no effect of time on task, $F(1.87, 129.37) = 0.24$, $p = .776$, $\eta_p^2 = .00$. The two-way interaction between Prevalence × Time on task attained significance, $F(1.97, 136.18) = 17.9$, $p < .001$, $\eta_p^2 = .21$. No other interactions

were significant: Break × Prevalence, $F(1, 69) = 0.01$, $p = .917$, $\eta_p^2 = .00$; Break × Time on task, $F(1.87, 129.37) = 1.23$, $p = .294$, $\eta_p^2 = .02$; Break × 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 x 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

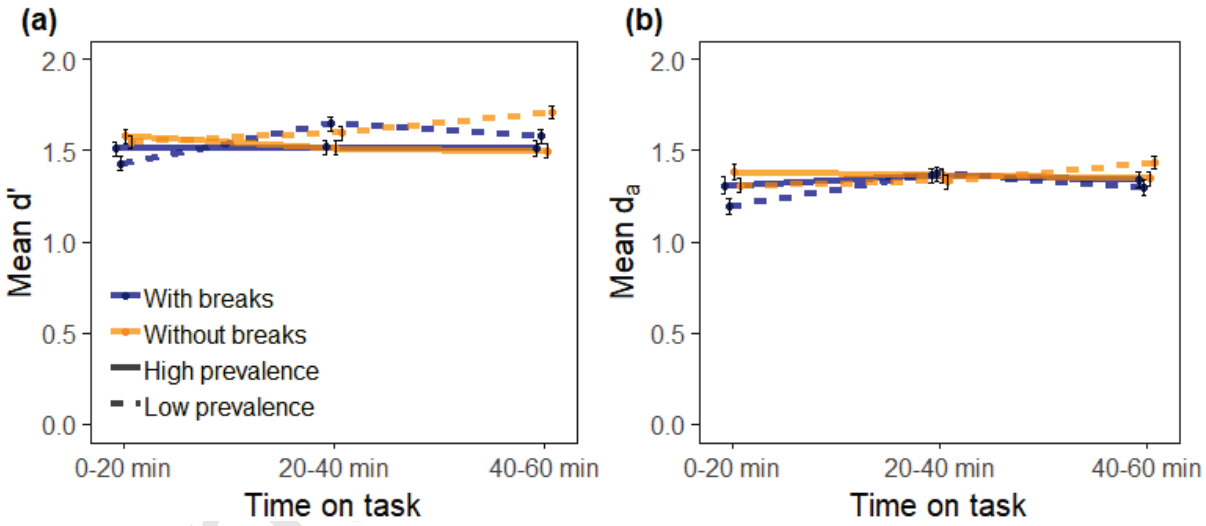

Fig.3. Sensitivity measure d' (a) and sensitivity measure d_a (b) for the group with breaks and the group without breaks for both prevalence conditions as a function of time on task.Standard errors are represented in the figure by the error bars.

Figure 3a shows the sensitivity measure d' 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, $F(1, 69) = 12.83$, $p = .001$, $\eta_p^2 = .16$; but no effect of break, $F(1, 69) = .80$, $p = .375$, $\eta_p^2 = .01$; or time on task, $F(1.99, 136.97) = 3.62$, $p = .030$, $\eta_p^2 = .05$. The

interaction Prevalence × Time on task attained significance, $F(1.93, 133.34) = 1.08$, $p = .340$, $\eta_p^2 = .02$. All other interactions were not significant: Break × Prevalence, $F(1, 69) = .15$, $p = .697$, $\eta_p^2 = .00$; Break × Time on task, $F(1.99, 136.97) = 2.77$, $p = .067$, $\eta_p^2 = .04$; and Break × Prevalence × Time on task, $F(1.93, 133.34) = 1.83$, $p = .166$, $\eta_p^2 = .03$.

Post hoc analyses revealed no significance in *d′* between *0–20 min* and *20–40 min* for the high prevalence condition ($p = .835$) nor for the low prevalence condition ($p = .060$). Also no significant difference was found between *20–40 min* and *40–60 min* in either the high-prevalence ($p = 1.000$) or low-prevalence ($p = 1.000$) condition.

The estimated slope parameter was 0.65 (95% BCa-CI [0.41, 0.89]) and thereby lower than the slope assumed by *d′*. Figure 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, $F(1, 69) = .65$, $p = .423$, $\eta_p^2 = .01$ (whereby the main effect of prevalence has no informative value, because this main effect was used to estimate the slope parameter) or break, $F(1, 69) = 1.03$, $p = .314$, $\eta_p^2 = .01$. No interactions attained significance: Break × Prevalence, $F(1, 69) = .22$, $p = .638$, $\eta_p^2 = .00$; Break × 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 Break × Prevalence × Time on task, $F(1.95, 134.72) = 1.53$, $p = .221$, $\eta_p^2 = .02$.

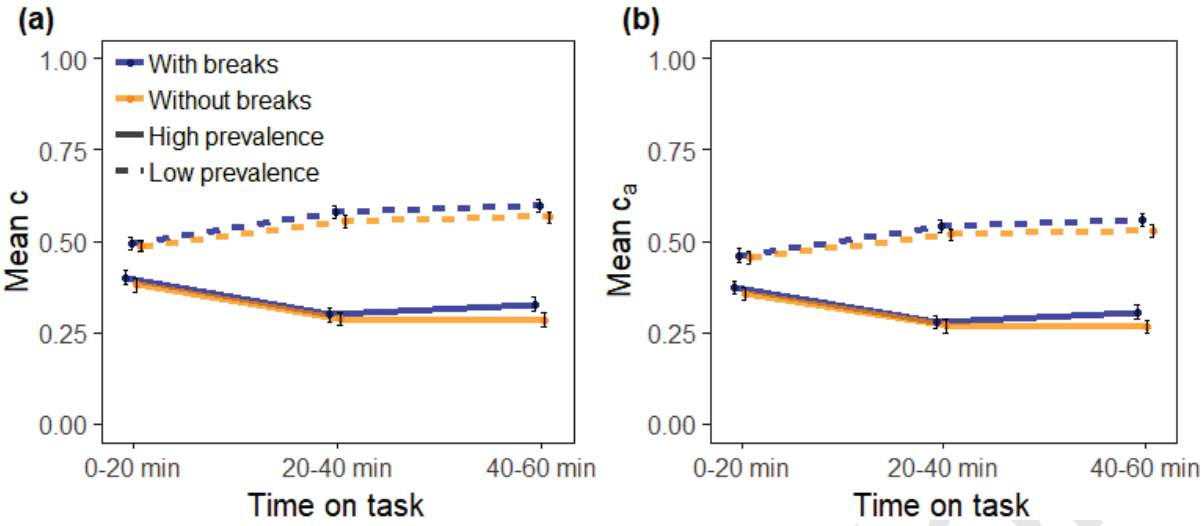

Fig. 4. Criterion c and c_a for groups with breaks and without breaks for both prevalence conditions as a function of time on task. Standard errors are represented in the figure by the error bars.

Figure 4 displays the criterion measures *c* and c_a. 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* and does not affect significance testing, ANOVA and post hoc results are identical for both *c* and c_a and are 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 break, $F(1, 69) = 0.96$, $p = .329$, $\eta_p^2 = .01$; or time on task, $F(1.93, 133.02) = 0.40$, $p = .665$, $\eta_p^2 = .01$. 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 Break × Prevalence, $F(1, 69) = .25$, $p = .619$, $\eta_p^2 = .00$; Break × Time on task, $F(1.93, 133.02) = .24$, $p = .782$, $\eta_p^2 = .00$; or Break × Prevalence × Time on task, $F(1.95, 134.28) = .02$, $p = .977$, $\eta_p^2 = .00$. 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).

3.3. Processing time

Figure 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.

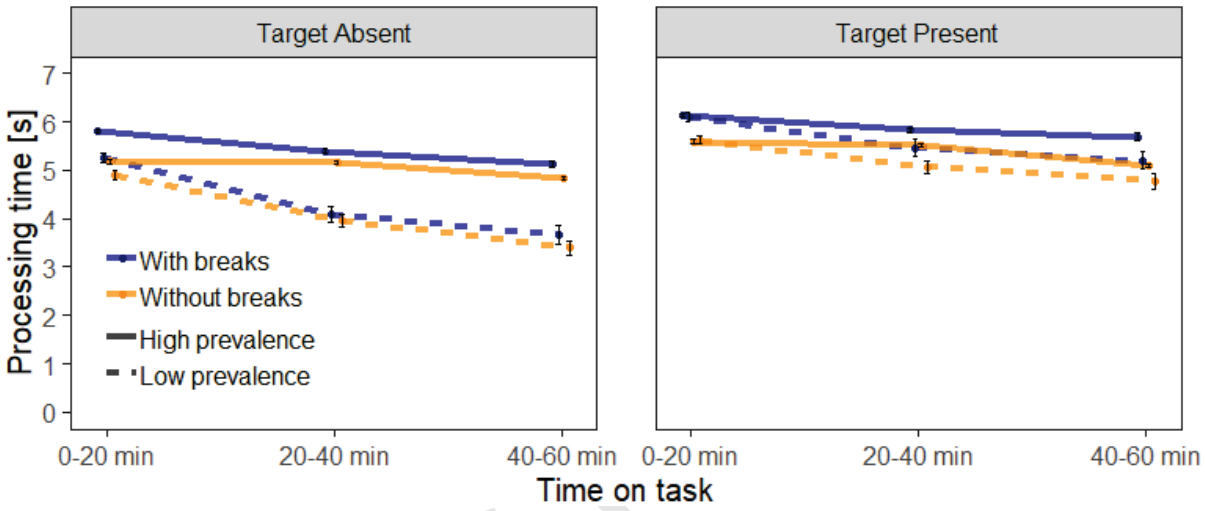

Fig. 5. Processing time for target-absent and target-present trials for both breaks and prevalence conditions as a function of time on task. Standard errors are represented in the figure by the error bars.

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 break was not significant, $F(1, 69) = 1.27$, $p = .264$, $\eta_p^2 = .02$. The interaction Prevalence × Time on task attained significance, $F(1.54, 105.92) = 37.07$, $p < .001$, $\eta_p^2 = .35$. The other interactions did not attain significance, Break × Prevalence, $F(1, 69) = .34$, $p = .560$, $\eta_p^2 = .00$; Break × Time on task, $F(1.69, 116.40) = 2.93$, $p = .066$, $\eta_p^2 = .04$; Break × 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).

For target-present trials, the ANOVA revealed significant main effects of prevalence, $F(1, 69) = 6.67$, p = .012, η_p^2 = .09; break, $F(1, 69) = 5.28$, p = .025, η_p^2 = .07; and time on task, $F(1.97, 136.22) = 26.98$, p < .001, η_p^2 = .28. The interaction Prevalence × Time on task also attained significance, $F(1.97, 135.76) = 4.61$, p = .012, η_p^2 =.06. The interactions Break × Prevalence, $F(1, 69) = .04$, p = .851, η_p^2 = .00; Break × Time on task, $F(1.97, 136.22)$ = .43, p = .649, η_p^2 = .01; and Break × Prevalence × Time on task, $F(1.97, 135.76) = .29$, p = .745, η_p^2 = .00, were not significant. Post hoc tests for the significant interaction of Prevalence × Time on task revealed no significant decrease in reaction time between *0-20 min* and *20-40 min* for high prevalence (p = .161) but a significant decrease for the low prevalence condition (p = .000). Also between *20-40 min* and *40-60 min* there was no significant decrease for the high prevalence condition (p = .071) but a significant decrease for the low prevalence condition (p = .016).

3.4. Subjective measures of distress, worry, and engagement

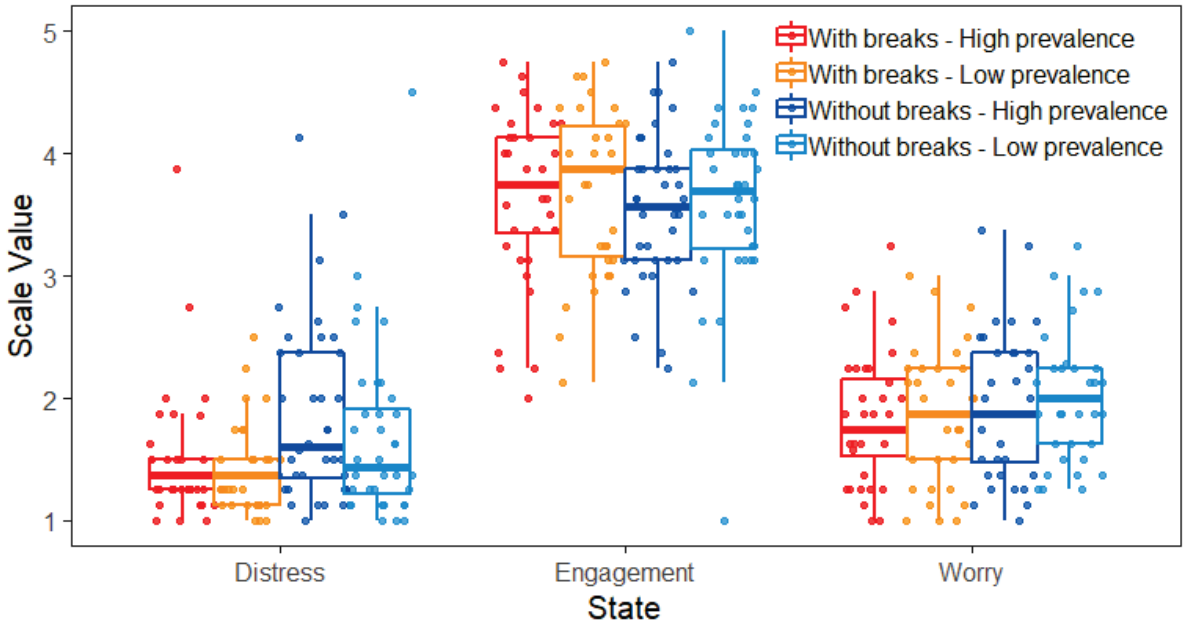

Fig. 6. Subjectively reported levels of *distress, worry*, and *engagement*. The lower and upper hinges correspond to the 25th and 75th percentiles. Each point represent a single measurement.

Figure 6 shows the reported levels of *distress, worry*, and *engagement* for both break and prevalence conditions. Two participants did not fill in the questionnaire and are therefore not in these results. Cronbach's alpha was .75 for *distress*, .81 for *engagement*, and .70 for *worry*. For the subjective stress levels, we calculated 2 (with vs. without breaks) × 2 (high vs. low prevalence) ANOVAs with the three levels of stress *distress, worry*, and *engagement* as dependent variables. For *distress*, the ANOVA revealed a significant main effect of break, $F(1, 66) = 9.17$, $p = .004$, $\eta_p^2 = .12$. Because the data do not meet the assumptions of normal distribution or homoscedasticity, a Wilcoxon rank sum test was carried out, which also revealed a significant difference between the two break conditions ($W = 1616$, $p = .003$). The main effect of prevalence, $F(1, 66) = 1.44$, $p = .234$, $\eta_p^2 = .02$, and the interaction Break × Prevalence, $F(1, 66) = 1.59$, $p = .212$, $\eta_p^2 = .02$, were not significant. For *worry*, the ANOVA revealed no significant effects: break, $F(1, 66) = 2.35$, $p = .13$, $\eta_p^2 = .03$; prevalence, $F(1, 66)$

= .58, p = .449 η_p^2 = .01; or Break × Prevalence, $F(1, 66)$ = .04, p = .847, η_p^2 = .00. For *engagement*, the ANOVA also revealed no significant effects for either break, $F(1, 66)$ = .70, p = .406, η_p^2 = .01, prevalence, $F(1, 66)$ = .56, p = .455, η_p^2 = .01, or for the interaction Break × Prevalence, $F(1, 66)$ = .04, p = .847, η_p^2 = .00.

4. Discussion

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

Because our findings on 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 false alarm rate in the low target prevalence condition compared to the high target prevalence condition. This is the typical effect of target prevalence on hit rate and false alarm rate: People adjust their criterion depending on the base rate with which targets occur. When comparing d' between the two target prevalence conditions over the full length of the test (i.e. the main effect of target prevalence), we found higher d' values for the low target prevalence condition in line with previous research (Godwin et al., 2010; Wolfe et al., 2007; Wolfe and Van Wert, 2010). At the same time, screeners needed less time to inspect an image in this condition.

In line with 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' 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. 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 (Godwin et al., 2010; Sterchi et al., 2019; Wolfe & Van Wert, 2010). Therefore, in line with these previous studies, d_a seems to be the appropriate detection measure here. A comparison of the two target prevalence conditions regarding the criterion again showed a clear prevalence effect. As mentioned, screeners needed less time to inspect an X-ray image in the low target prevalence condition. This was more strongly the case for target-absent trials in line with previous research (Godwin et al., 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010). In summary, we found that a lower target prevalence causes a shift in response tendency resulting in a lower hit rate and false alarm rate. Whereas d' would suggest that sensitivity also decreases, this is implausible with regard to shorter processing times.

Moreover, previous research suggests that d_a with a slope of around 0.6 is more appropriate. Consistent with this, our data suggest a *z*ROC slope of 0.65.

4.2. Interaction between target prevalence and time on task

Previous studies have found that the target prevalence effect depends on implicit learning 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 & Huang, 2010)). For the false alarm rate, we found a significant interaction between target prevalence and time on task. More specifically, within the first 20 min of the task the false alarm rate increased in the high target prevalence condition and decreased in the low target prevalence condition. This is consistent with the shift in criterion (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, which led to higher standard errors. Our analyses of the criterion, which takes the hit rate and the false alarm rate into account, clearly confirm that the effect of target prevalence increased within the first 20 min of the test. 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. 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 & Huang, 2010). In addition, consistent with the findings of Lau & 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 within the first 20 min of the test, which depended on the target prevalence condition. To discuss the effect of time on task on detection performance, it makes therefore sense to focus on the sensitivity measure d_a (with a slope of 0.65), which is not affected by this criterion shift. Whereas some of the previous research found performance decrements, we found a small increase in d_a over the first 20 min of the test and no change thereafter. The initial ramp-up could be due to accustomization to the task. It is possible that there is a warm-up phase in X-ray image inspection, during which the cognitive processes necessary for this task are fully activated, as can be observed in other recognition tasks (e.g. Monsell, 2003; Allport & Wylie, 1999). It is however 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 test. The change found by Meuter and Lacherez might also have been a shift in criterion. 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 where screening was performed in the lane, our study

investigated remote screening. It may be more difficult to maintain performance in an environment with more noise and distractors.

The pattern within the 60 min of screener performance from our study is similar to what Ghylin, Drury, Batta, and Lin (2007) found when testing how screener performance changed after the first hour up to hour four. They found the hit, false alarm rate, and reaction time to decrease while sensitivity (measured with *A'*) remained constant, concluding the presence of a criterion shift.

As we 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)) 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 image inspection as a typical vigilance task. One could argue that our study contrasts more strongly with vigilance tasks than the conventional X-ray image inspection task, because we used a higher target prevalence. However, whereas threats such as IEDs and guns are rare in practice, other prohibited items such as liquids and laptops left in baggage still provide quite common targets.

Another measure to consider regarding performance is processing time. Processing times decreased throughout the test. This decrease 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 test conditions and interface settings. We cannot be sure whether this effect would also occur in practice after screeners become familiar with the interface of the X-ray analysis software.

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). Whereas breaks have often had a positive effect on performance in previous studies (Arrabito et al. 2015; Balci and Aghazadeh, 2003; Colquhoun, 1959; Kopardekar and Mital, 1994; Steinborn and Huestegge, 2016), breaks are mainly thought to offer rest, recuperation, and prevention of fatigue (Tucker, 2003). Considering that participants who performed 60 min of continuous screening did not show a decrease in performance, there was no room for recuperation during breaks. 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 within the first 20 min and did not increase thereafter, indicating that it was not the result of the breaks themselves. Because the effect was not highly significant, it could also be coincidental. But if this effect of the break condition truly exists, it must be due to the instruction that there will or will not be breaks. Maybe knowing that there will not be breaks induced a bit of 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 stress. 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, increased distress could have an affect on performance. It has, however, to be noted that there was considerable variance between screeners in the condition without breaks. Whereas the longer screening without breaks caused distress in some participants, it did not in others.

4.5. Limitations and future research

Whereas our study has shown that screeners can maintain detection performance over 60 min without breaks, it is still too early to derive implications for practice.

Our results show that 60 min of continuous screening caused distress for some screeners. Considering that participants only did 60 min of screening twice with 3 to 5 weeks in between, it is unclear how prolonged screening would affect performance and distress if repeated multiple times a day and over months. One the one hand, distress levels could decrease with increasing practice, or could also induce more distress with time. This could in turn have a negative impact on well-being and on performance in the long term. Therefore, field studies are needed to determine how longer screening durations will affect performance and well-being of individuals in the long term. Such field studies would also tackle other limitations of our study. In our lab study, poor performance did not have any consequences whereas a miss can be disastrous in practice. This might make prolonged screening time more stressful in practice. Further, target prevalence is lower in practice and wherefore sustaining attention and performance could be more difficult.

Our results suggest that people react very differently to prolonged working sessions. Future studies should try to identify the reasons behind such interindividual differences and test whether flexible break schedules could provide a solution, although that might often be difficult in practice. Because interindividual differences seem likely, it is advisable to test different populations of security officers (at different airports and in different countries).

There is also a potential issue of researching different work settings with professional screeners in general. It would be reasonable for them to conclude that the results of such research is likely to affect their work in the long run and might let them act biased.

5. Conclusions

Our study shows that for X-ray image inspection of cabin baggage d_a with a slope of approx. 0.6 is a more valid measure of detection performance than d'. This performance measure is independent of the target prevalence effect that we have found to evolve at the beginning of the task. Further, it seems that the target prevalence effect is a shift in criterion rather than a loss in sensitivity. The examination of d' reveals that performance does not decrease in continuous X-ray inspection over the course of 60 min. Moreover, breaks do not influence performance. However, breaks do seem to have an effect on well-being, in the sense that screeners without breaks report more distress. It is also evident that people working without breaks evaluate the task quite differently with regard to the amount of distress caused by the task. These results provide the necessary evidence that longer screening durations are possible, and they allow implications for trials in the field. We conclude that trials of longer screening durations in the field 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 screening.. Further research will reveal clearer recommendations regarding optimal break schedules.

6. References

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

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. In *Proceedings - International Carnahan Conference on Security Technology*. https://doi.org/10.1109/CCST.2009.5335565

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

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*. 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. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, pp. 93–97). SAGE PublicationsSage CA: Los Angeles, CA. https://doi.org/10.1177/154193120705100209

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. 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*. https://doi.org/10.1007/BF02289823

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

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

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. *WIT Transactions on the Built Environment*. 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*. https://doi.org/10.3758/s13414-011-0225-4

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*. https://doi.org/10.1080/00140139408964946

Lau, J. S. H., & Huang, L. (2010a). 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

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

Mackworth, N. H. (1948). The breakdown of vigilance durning prolonged visual search.

Quarterly Journal of Experimental Psychology, *1*(1), 6–21. https://doi.org/10.1080/17470214808416738

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., … Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*. https://doi.org/10.1037/1528-3542.2.4.315

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? In *Proceedings - International Carnahan Conference on Security Technology*. 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. *Human Factors*, *58*(2), 218–228. https://doi.org/10.1177/0018720815616306

Nuechterlein, K. H., Parasuraman, R., & Jiang, Q. (1983). Visual sustained attention: Image degradation produces rapid sensitivity decrement over time. *Science*. https://doi.org/10.1126/science.6836276

Rose, C. L., Murphy, L. B., Byard, L., & Nikzad, K. (2002). The Role of the Big Five Personality Factors in Vigilance Performance and Workload. *European Journal of Personality*. https://doi.org/10.1002/per.451

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*. 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*. 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: The Journal of Human Factors and Ergonomics Society*. https://doi.org/10.1177/001872087401600402

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

Tucker, P. (2003b). 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

Wert, M. J. Van, Horowitz, T. S., & Wolfe, J. M. (2009). Frequently Missed, *71*(3), 541–553. https://doi.org/10.3758/APP.71.3.541.Even

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

Applied Ergonomics 72 (2018) 58–68

Contents lists available at ScienceDirect

Applied Ergonomics

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

Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection

Nicole Hättenschwiler*, Yanik Sterchi, Marcia Mendes, Adrian Schwaninger

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

ARTICLE INFO

Keywords:
Airport security X-ray screening
Explosives detection
Automation

ABSTRACT

Bomb attacks on civil aviation make detecting improvised explosive devices and explosive material in passenger baggage a major concern. In the last few years, explosive detection systems for cabin baggage screening (EDSCB) have become available. Although used by a number of airports, most countries have not yet implemented these systems on a wide scale. We investigated the benefits of EDSCB with two different levels of automation currently being discussed by regulators and airport operators: automation as a diagnostic aid with an on-screen alarm resolution by the airport security officer (screener) or EDSCB with an automated decision by the machine. The two experiments reported here tested and compared both scenarios and a condition without automation as baseline. Participants were screeners at two international airports who differed in both years of work experience and familiarity with automation aids. Results showed that experienced screeners were good at detecting improvised explosive devices even without EDSCB. EDSCB increased only their detection of bare explosives. In contrast, screeners with less experience (tenure < 1 year) benefitted substantially from EDSCB in detecting both improvised explosive devices and bare explosives. A comparison of all three conditions showed that automated decision provided better human–machine detection performance than on-screen alarm resolution and no automation. This came at the cost of slightly higher false alarm rates on the human–machine system level, which would still be acceptable from an operational point of view. Results indicate that a wide-scale implementation of EDSCB would increase the detection of explosives in passenger bags and automated decision instead of automation as diagnostic aid with on screen alarm resolution should be considered.

1. Introduction

Secure air transportation is vital for both the economy and society (Abadie and Gardeazabal, 2008). For several decades now, airplanes have been interesting targets for terrorists (Baum, 2016). Looking at the history of attacks against airplanes (both successful and near misses), one of the biggest concerns is bombs – that is, improvised explosive devices (IEDs; Novakoff, 1993; Singh and Singh, 2003; Baum, 2016). The Global Terrorism Database (2017) lists 893 attacks on airports or aircrafts with explosives, 247 of which occurred after 2001. Quite recently, on the 29th of July 2017, a terrorist plot was prevented at Sydney airport when an IED was found concealed inside a bag (Westbrook and Barrett, 2017). In response to heightened risk, especially since 9/11, airports and governments have increased their investments in aviation security (Gillen and Morrison, 2015). In the last few years, explosive detection systems for cabin baggage screening (EDSCB) have also become available (Sterchi and Schwaninger, 2015). Whereas a few countries such as the United States are using these systems (Neffenger, 2015), they have not been implemented widely in European countries and on other continents (Pochet, 2016). We investigated the benefits of EDSCB with two different levels of automation that are both being discussed currently by regulators and airport operators. We were able to recruit airport security officers (screeners) from two different European airports to work on two experiments using a simulated cabin baggage screening task. In this introduction, we first summarize previous research on visual inspection and conventional cabin baggage screening before going on to discuss automation and EDSCB.

1.1. Visual inspection and conventional cabin baggage screening

To prevent terrorist attacks and other acts of unlawful interference, passengers and their belongings have to be screened before they are allowed to enter the secure areas of airports and board airplanes (Thomas, 2009). Screeners visually inspect X-ray images of cabin baggage for prohibited items such as guns, knives, and improvised

* 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: nicole.haettenschwiler@fhnw.ch (N. Hättenschwiler).

https://doi.org/10.1016/j.apergo.2018.05.003
Received 15 December 2016; Received in revised form 4 May 2018; Accepted 5 May 2018

explosive devices (IEDs) as well as other items such as self-defence gas sprays or Tasers (Schwaninger, 2005). This inspection involves visual search and decision making (Koller et al., 2009; Wales et al., 2009; Wolfe and Van Wert, 2010). The challenges when performing visual search in X-ray baggage screening include a low target prevalence, the variation in target visibility, the search for an unknown target set, and the possible presence of multiple targets (for recent reviews, see Biggs and Mitroff, 2014; Mitroff et al., 2015). When deciding whether or not a bag contains a prohibited item, screeners need to know which items are prohibited and what they look like as X-ray images (Schwaninger, 2005, 2006). Whereas even novices can recognize certain object shapes such as guns and knives in X-ray images (Schwaninger et al., 2005), other prohibited items such as IEDs are difficult to recognize without training (Schwaninger and Hofer, 2004; Koller et al., 2008, 2009; Halbherr et al., 2013). An IED is composed of a triggering device, a power source, a detonator, and explosive that are usually all connected by wires (Turner, 1994; Wells and Bradley, 2012). Through computer-based training, screeners can learn to recognize these components, and they can achieve and maintain a high detection performance for IEDs (Schwaninger and Hofer, 2004; Koller et al., 2008, 2009; Halbherr et al., 2013; Schuster et al., 2013). In cabin baggage screening, bare explosives also pose a threat, because these could be combined with other IED components after passing an airport security checkpoint. Detecting bare explosives can be a challenge even for well-trained screeners, because they often look like a harmless organic mass (Jones, 2003). So far, no study has investigated how well screeners can detect bare explosives and whether automation and EDSCB can increase human–machine system performance in response to such threats. Before discussing automation and EDSCB as a specific application, it is worth considering important findings and concepts on automation and human–machine system performance in general.

1.2. Automation and human–machine system performance

Automation refers to functions performed by machines (usually computers) that assist or replace tasks performed by humans (for reviews, see Parasuraman and Wickens, 2008; Sheridan, 2011; Vagia et al., 2016). One form of automation assisting humans is the diagnostic aid (Wickens and Dixon, 2007). This provides support in the form of alerts or alarms and influences attention allocation (Cullen et al., 2013). Examples include collision warning systems for driving and air traffic control (Lehto et al., 2000; Abe and Richardson, 2006; Liu and Jhuang, 2012; Biondi et al., 2017) or aids assisting radiologists in making diagnostic decisions from mammograms (e.g. Vyborny et al., 2000; Fenton et al., 2007). Other examples are systems that indicate potentially threatening objects in X-ray images of passenger baggage. These systems have been investigated in laboratory studies with student participants (Wiegmann et al., 2006; Rice and McCarley, 2011). Common to this type of automation is that it categorizes events into target or non-target states (Wickens and Dixon, 2007). Signal detection theory (Green and Swets, 1966, 1972) provides a useful framework with which to describe the performance (reliability) of such diagnostic automation (Wickens and Dixon, 2007; Parasuraman and Wickens, 2008; Rice and McCarley, 2011). In signal detection theory, high performance (reliability) in terms of d' is achieved when targets are detected well (high hit rate) and the false alarm rate is low. The criterion (or response bias) is a threshold that can be changed while d' remains constant (Macmillan and Creelman, 2005). The criterion can be changed by adjusting thresholds for alerts, resulting in a trade-off between two types of automation errors: misses and false alarms (Parasuraman, 1987; Parasuraman and Riley, 1997; Wickens and Colcombe, 2007). Designers often set low thresholds, because the consequences of automation misses are considered to be more costly than false alarms (Parasuraman and Wickens, 2008). However, if the base rate of dangerous events to be detected is low, the result will be many false alarms and only few hits (Parasuraman and Riley, 1997). This can produce a 'cry wolf' effect with operators ignoring system warnings (Breznitz, 1983; Bliss, 2003). Such an effect can drastically reduce or even eliminate the benefits of automation when it is implemented as a diagnostic aid.

Alongside automation as a diagnostic aid, other levels of automation are possible. Sheridan and Verplank (1978) proposed a taxonomy with 10 levels of automation ranging from fully manual to fully computer automated. Parasuraman et al. (2000) proposed a taxonomy with four processing stages: 1) sensory processing, 2) perception/working memory, 3) decision making, and 4) response/action. Several other taxonomies for different levels of automation have been proposed (for a review, see Vagia et al., 2016). Kaber and Endsley (2003) have pointed out that specifying the 'best' level of automation is not as straightforward as one might think. Moreover, familiarity with automation can affect how people interact with it (Parasuraman and Manzey, 2010; Sauer et al., 2016; Strauch, 2016; Sauer and Chavaillaz, 2017). Indeed, deciding how best to organize human–machine function allocation and the level of automation remains a difficult task that can also depend on the specific application (Sheridan, 2011). Parasuraman et al. (2000) have suggested that appropriate criteria for selecting the level of automation for a particular application are human performance, automation reliability, and the cost associated with outcomes.

1.3. Automation and EDSCB

For X-ray screening of cabin baggage, regulators and airport operators are currently discussing two EDSCB implementation scenarios differing in their level of automation and human–machine function allocation: on-screen alarm resolution (OSAR) and automated decision (Sterchi and Schwaninger, 2015). In the OSAR scenario, automation is implemented as a diagnostic aid. Screeners visually inspect *every* piece of cabin baggage. During this inspection, EDSCB indicates potential explosive material by either marking an area on the X-ray image of a passenger bag with a coloured rectangle or highlighting it in a special colour (Nabiev and Palkina, 2017). Screeners then have to resolve this; that is, they have to visually inspect the X-ray image and decide whether the area indicated by the machine is harmless (EDSCB false alarm) or whether it actually could be explosive material, making it necessary to subject the baggage to a secondary inspection. This is also conducted at the airport security checkpoint and involves explosive trace detection, opening the bag, and manually searching it (Sterchi and Schwaninger, 2015). EDSCB systems with high hit rates (close to 90%) have false alarm rates in the range of 15–20% (personal communication with EDSCB experts, summer 2016). As mentioned above, system reliability can be described by d' from signal detection theory (Green and Swets, 1966, 1972). For example, an EDSCB with a hit rate of 88% and a false alarm rate of 17% would have a system reliability of $d' = 2.1$. In operation, most of the EDSCB alarms are cleared by screeners, leaving only a small percentage of bags on which EDSCB has raised an alarm that then requires a secondary inspection. Although OSAR is the scenario currently employed at airports that have already introduced EDSCB, its effectiveness can be questioned, because screeners might not be able to distinguish explosive material from benign material (as pointed out already by Jones, 2003). Moreover, EDSCB false alarm rates of 15–20% could result in a cry wolf effect leading screeners to potentially ignore system warnings (Breznitz, 1983; Bliss, 2003). Screeners might therefore be prone to mistakenly clearing bags that contain explosives. This would drastically reduce the effectiveness of EDSCB in the OSAR scenario. In other words, the probability of detecting explosives on the human–machine system level equals about 90% (EDSCB) *minus* the erroneously cleared alarms by screeners. This could result in a much lower detection rate.

The automated decision scenario uses a higher level of automation with different human–machine function allocation. Bags on which the EDSCB raises an alarm are sent *automatically* to secondary inspection using manual search and/or explosive trace detection (Sterchi and

Schwaninger, 2015). Because secondary inspection is time-consuming, EDSCB false alarm rates of 15–20% are not acceptable in this scenario. To be operationally feasible, EDSCB thresholds can be adjusted, which corresponds to moving the criterion in signal detection theory (Green and Swets, 1966; Macmillan and Creelman, 2005). For example, given a system reliability of $d' = 2.1$, like that in the OSAR scenario explained above, adjusting EDSCB thresholds to achieve a false alarm rate of 4% would result in an EDSCB hit rate of 63%. It is important to remember that in the automated decision scenario, screeners visually inspect all X-ray images on which the EDSCB does not raise an alarm. In the current example, this equals 96% of all bags (assuming a false alarm rate of the EDSCB of 4%). The probability of detecting explosives on the human–machine system level therefore equals 63% (EDSCB hit rate) *plus* detections by screeners on the 96% of bags on which the EDSCB has not raised an alarm. Therefore, in this example, the probability of detecting explosives on the human–machine system level equals 63% (EDSCB) *plus* the detections by screeners.

In summary, for a given EDSCB, the effectiveness of OSAR and the automated decision scenario depends finally on the screeners' ability to clear alarms by the EDSCB (in the OSAR scenario) and to detect explosives missed by the EDSCB (in both scenarios). Which scenario results in better human–machine system performance is difficult to predict and well worth investigating.

1.4. Present study

The present study examined the benefits of automated explosive detection systems for cabin baggage screening (EDSCB) in two realistic implementation scenarios differing in the level of automation and human–machine function allocation (EDSCB with OSAR vs automated decision). It addressed the following three research questions: 1) Does EDSCB lead to higher human–machine system performance for detecting IEDs and explosives? 2) Does this depend on the level of automation (OSAR vs automated decision)? 3) Is this dependent on screener work experience? To address these research questions, two experiments using a simulated baggage screening were conducted at different European airports with screeners differing in work experience.

Based on previous research, we derived three hypotheses: 1) EDSCB should improve human–machine system performance for detecting bare explosives because these often look like a harmless organic mass (Jones, 2003). 2) We expected better results for the automated decision scenario compared to OSAR, because clearing EDSCB alarms can be difficult (Jones, 2003) and false alarm rates of 15–20% in the OSAR scenario may result in a cry wolf effect with screeners ignoring system warnings (Breznitz, 1983; Bliss, 2003). 3) Effects should depend on screener work experience because previous research has shown that regular computer-based training, which is mandatory in Europe, results in large increases in IED detection during the first few years (Halbherr et al., 2013). Experiment 1 examined the first two hypotheses. The aims of Experiment 2 were to perform a replication, to address the limitations of Experiment 1, and to test all three hypotheses.

2. Experiment 1

2.1. Method

2.1.1. Participants

The current research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board of the University of Applied Sciences and Arts Northwestern Switzerland. Informed consent was obtained from all participants. The study was conducted with 61 screeners who had been qualified, trained, and certified according to the standards set by the appropriate national authority (civil aviation administration) in compliance with the relevant EU Regulation (Commission Implementing Regulation [EU], 2015/1998). Screeners had been employed for at least two years ($M = 7.68$, $SD = 4.85$) and were not familiar with automation aids for cabin baggage screening. They participated on a voluntary basis, were recruited by a security service provider at the airport, and compensated by regular salary. Their average age[1] was 42.5 years ($SD = 10.52$, range 24–60 years), and 57.37% of them were female.

2.1.2. Design

The experiment used a between-subjects design with condition (no automation as baseline, OSAR, and automated decision) as independent variable and hit rate (percentage detection of prohibited items) and false alarm rate of the human–machine system as dependent variables. The three experimental groups were balanced with regard to their detection performance score in a pre-test (X-ray CAT), age, and work experience (baseline, $n = 20$; OSAR, $n = 20$; automated decision, $n = 21$).

2.1.3. Materials

Pre-test: The *X-Ray Competency Assessment Test (X-Ray CAT)* is a reliable, valid, and standardized computer-based test used to assess the X-ray image interpretation competency of screeners (Koller and Schwaninger, 2006). It has been applied in several previous studies and is used for the mandatory X-ray screener certification at a number of European airports (e.g. Koller et al., 2008; Michel et al., 2007; Koller et al., 2009; Steiner-Koller et al., 2009; Halbherr et al., 2013). To solve the X-Ray CAT, screeners have to visually scan X-ray images for prohibited items and decide whether a bag can be considered either to be harmless (OK) or to contain a prohibited object (NOT OK). For a more detailed description of the X-Ray CAT, see Koller and Schwaninger (2006).

Main test: We measured human–machine system performance in the three automation conditions (baseline, OSAR, and automated decision) with 640 unique X-ray images of real passenger bags. These were selected by two experts (former screeners) from a pool of about 2000 X-ray images recorded during regular airport security screening operations. This selection procedure included making sure that no prohibited items were contained in the X-ray images. Target-present images were created by the screening experts using previously recorded prohibited items that were placed into 80 of the 640 X-ray images using a software- and image-merging algorithm that had been validated in previous studies (von Bastian et al., 2009; Mendes et al., 2011). This corresponds to a target prevalence of 12.5%. Five different threat categories were included in this test: IEDs, explosive materials, guns, gun parts, and knives (see Fig. 1 for examples).

The category gun parts was included to compare detection performance with explosives because the latter are parts of IEDs. Each category contained eight different prohibited items. As in the X-Ray CAT (Koller and Schwaninger, 2006), each item was depicted twice: once from an easier, canonical viewpoint, and once from a more difficult, rotated viewpoint.

Fig. 2 illustrates the three automation conditions. In the baseline condition (Fig. 2a), no automation is available and detecting prohibited items relies only on the screener. In the OSAR condition, automation is implemented as a diagnostic aid and red frames highlight areas in the X-ray image on which the EDSCB has raised an alarm (Fig. 2b). For the OSAR condition, an EDSCB was emulated by showing a red frame around 14 of the 16 IEDs and explosives and around 94 of the 560 images of harmless bags. The frames on the images were set manually by a screening expert and were based on available information and professional experience with existing EDSCB machines. The emulated EDSCB had a hit rate of 88% and an alarm rate of 17% (as mentioned in the introduction, EDSCB systems in service at airports using OSAR have hit rates close to 90% and false alarm rates of 15–20%).

For the automated decision condition, a set of images of 10 IEDs, 10

[1] One X-ray screener did not report her or his age.

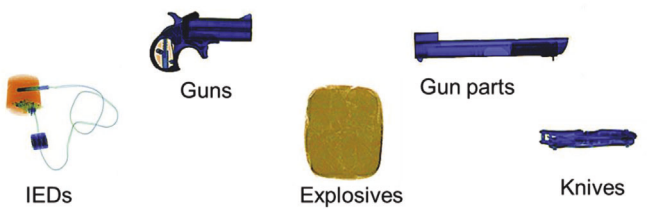

Fig. 1. Examples of the five threat categories.

explosives, and 20 harmless bags was randomly selected from the set of images with an alarm in the OSAR condition. These images were then removed (Fig. 2c) from the test (in order to emulate the implementation scenario in which bags that trigger an alarm by the EDS are sent directly to secondary inspection). The emulated EDS had a hit rate of 63% and a false alarm rate of 4%. This corresponds to the same system reliability of the EDS in terms of d' with a more conservative criterion (a requirement of the automated decision scenario, as explained in the introduction).

2.1.4. Procedure

All screeners came to the test facilities to conduct the pre-test (X-Ray CAT) and completed the main test on a second test date (mean interval between tests: 53 days, $SD = 11$). For the tests, eight laptops were set up in a normally lit room. Screeners sat approximately 60 cm away from the laptop screen. The X-ray images covered about two-thirds of the screen. Before starting the test, screeners were given general instructions on the number of images, the target prevalence, and the different prohibited item categories. They performed the test quietly, individually, and under supervision.

Screeners were instructed to inspect each image visually and report as quickly and accurately as possible whether a bag was harmless (OK) or not (NOT OK) by clicking on a button on the screen. In the OSAR condition, screeners were informed that they would be receiving support from an EDSCB that usually marks IEDs and explosives with a red frame. They were further instructed that red frames can also occur when the bag contains no IED or explosive (false alarm). In the automated decision condition, screeners were informed that this test condition would include support from an automated explosives detection system. They were told that if an IED or an explosive is detected by the EDSCB, the bag will be sent automatically to secondary inspection and will not be shown to the screener. They were further informed that in some cases, IEDs and explosives will not be detected by the EDSCB. After the instructions, all participants practiced on 20 sample images to familiarize themselves with the images and the task.

Following the European Commission (Commission Implementing Regulation [EU], 2015/1998) regulation, screeners have to take a break of at least 10 min after 20 min of continuous visual inspection of X-ray images. Therefore, the EDSCB test was divided into four equally long blocks, and screeners were asked to take a 10-min break after completing each block. Threat bags, threat categories, and harmless bags were distributed equally across the four blocks. The order of blocks was counter-balanced between conditions to minimize any training or order effects. Within a block, images appeared in random order. All participants completed the pre-test (X-Ray CAT) in less than 40 min and the main test in less than 2 h including breaks.

2.1.5. Analyses

All ANOVAs were conducted with SPSS version 22 and alpha was set at 0.05 unless otherwise stated. Post hoc comparisons were conducted with R version 3.22 (R Core Team, 2015) and Holm–Bonferroni corrections were applied (Holm, 1979). Effect sizes of ANOVAs are reported with η_p^2 (partial eta-squared); effect sizes of t tests, with Cohen's d.

ANOVAs were calculated using the hit and false alarm rate on the human–machine system level as dependent variables. Because hit and false alarm rates are bound between 0 and 1, normality and homogeneity of variances was generally not fully met. Traditionally, ANOVAs are assumed to be quite robust towards non-normality and homogeneity (e.g. Glass et al., 1972). However, because reviews question this robustness (Harwell et al., 1992; Erceg-Hurn and Mirosevich, 2008), all ANOVAs were also performed on scores that had been arcsine transformed for homogenization of variances and normalization (for more information on the application of arcsine transformations to proportion data, see McDonald, 2007). Results on transformed values are reported only when the transformation affected whether an effect attained significance.

2.2. Results

Fig. 3 shows the results of human–machine hit rate by prohibited item category and automation condition.

First, we conducted a univariate ANOVA with the hit rate of only the baseline condition. This revealed a significant effect of prohibited item category, $F(4, 76) = 83.03$, $p < .001$, $\eta_p^2 = 0.81$. Post hoc analyses revealed a significant effect between all category comparisons for prohibited items ($p < .017$) except for the comparison between knives and explosives ($p = .365$). Then, we conducted a 3 (prohibited item category: gun, gun parts, and knives) x 3 (condition: baseline, OSAR, and automated decision) ANOVA. We found no main effect of automation, $F(2, 58) = 1.05$, $p = .356$, $\eta_p^2 = 0.03$, and no interaction between prohibited item category and condition, $F(3.45, 100.05) = 0.63$, $p = .622$, $\eta_p^2 = 0.02$. To examine the benefits of EDSCB, we conducted a 2 (IED and explosives) x 3 (condition: baseline, OSAR, and automated decision) ANOVA. This revealed main effects for the prohibited item category, $F(1, 58) = 238.89$, $p < .001$, $\eta_p^2 = 0.80$, condition, $F(2, 58) = 34.74$, $p < .001$, $\eta_p^2 = 0.55$, and their interaction, $F(2, 58) = 37.06$, $p < .001$, $\eta_p^2 = 0.56$. For IEDs, there was a significant difference between OSAR and automated decision ($p = .041$) in favor of the automated decision condition. For explosives, direct post hoc comparisons showed a significant difference between the baseline condition and the automated decision condition ($p < .001$) as well as

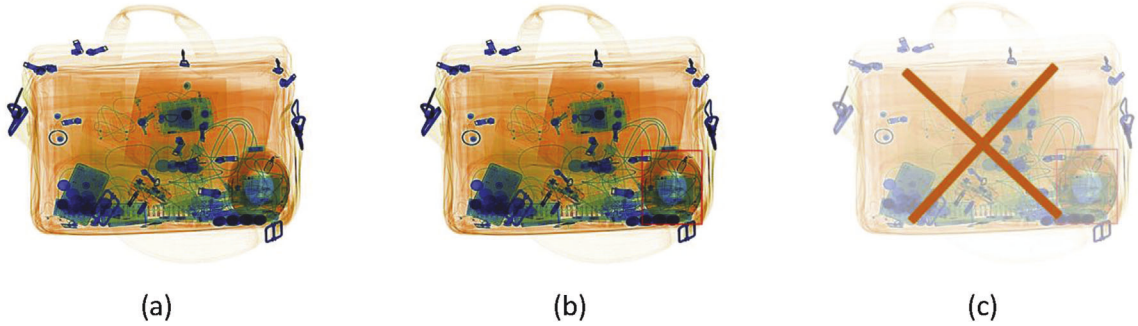

Fig. 2. Illustration of the three automation conditions: (a) baseline condition without automation, (b) OSAR, and (c) automated decision.

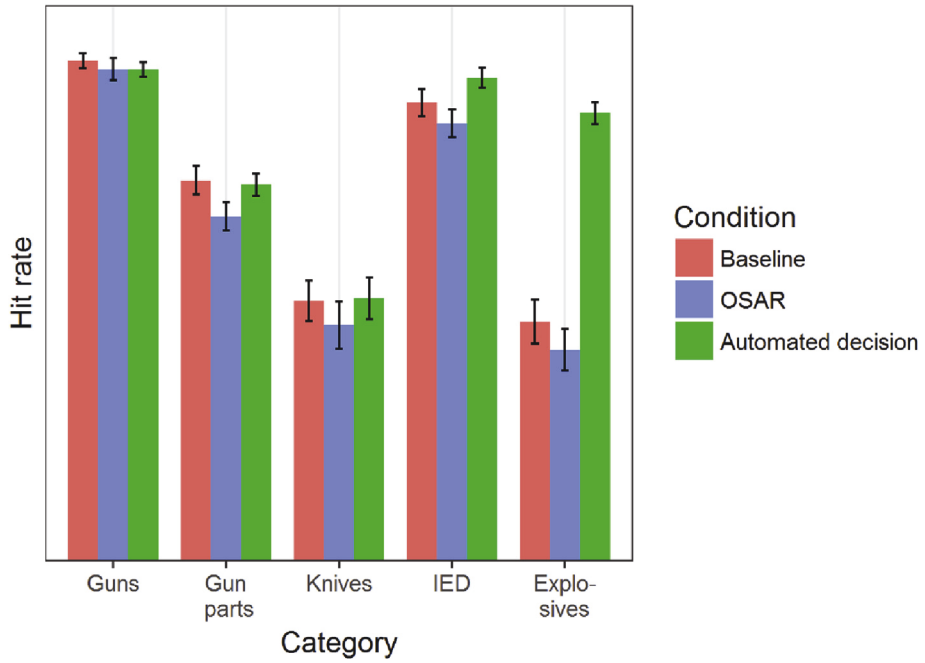

Fig. 3. Mean human–machine hit rates by condition (baseline, OSAR, automated decision) and prohibited item categories (guns, gun parts, knives, IEDs, and explosives). Absolute values of hit rate are not shown due to security restrictions in this project. Error bars are ± one standard error.

between OSAR and automated decision ($p < .001$).

Further analyses were conducted with the false alarm rate of the human–machine system as a dependent variable. A univariate ANOVA revealed a significant effect of condition, $F(2, 58) = 12.41$, $p < .001$, $\eta_p^2 = 0.30$. Post hoc pairwise comparisons using Holm–Bonferroni corrections revealed a significant difference between the baseline condition and automated decision ($p = .008$) as well as between OSAR and automated decision ($p < .001$). The false alarm rate in the automated decision condition was significantly higher than the false alarm rates in the two other conditions (see Fig. 4).

We further analysed whether automated decision affected human–machine system performance only through its direct contribution (i.e. producing hits and false alarms) or whether it also affected human performance. Therefore, the detection scores for images of IEDs and explosives shown to screeners in the automated decision condition (i.e. not sorted out by the automation aid) were compared with the detection scores for the same images from the baseline condition. Images that triggered the EDS alarm in the automated decision condition were excluded from this analysis for both conditions. Independent t tests were calculated for the hit rate for IEDs and for the hit rate for explosives. There were no significant effects for either IEDs, $t(39) = 0.40$, $p = .689$, or explosives, $t(39) = 0.34$, $p = .732$. Another t test was conducted with false alarm rate as the dependent variable. This revealed no difference between conditions, $t(39) = 0.64$, $p = .525$. In conclusion, it can be assumed that automated decision did not influence human performance.

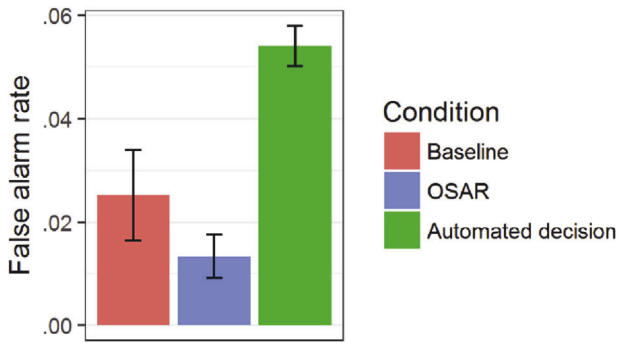

Fig. 4. Mean human–machine false alarm rates by condition (baseline, OSAR, and automated decision). Error bars are ± one standard error.

2.3. Discussion

The results for the baseline condition replicated those found in previous studies: guns were detected very well, IEDs only slightly less well, and knives came third (Koller et al., 2007, 2009; Halbherr et al., 2013). Gun parts were more difficult to detect than whole guns, presumably because configural representations of whole gun shapes cannot be accessed and only component representations of gun parts are available for recognition (Schwaninger, 2004). Explosives were difficult to detect, which could be due to the fact that they lack the diagnostic features of an IED and because explosive material often looks like a harmless organic mass (Jones, 2003). Automation had no impact on the detection of guns, gun parts, and knives. This is not surprising, because automation highlighted only potential explosives.

The screeners in Experiment 1 did not benefit from automation when OSAR was used with a realistically high false alarm rate of 17%. This is consistent with results found in earlier studies using different tasks indicating that automation with high false alarms can induce a cry wolf effect with operators ignoring system warnings (Bliss et al., 1995; Parasuraman et al., 2000). Results revealed a highly significant difference between the baseline condition and the automated decision condition – but only for explosives. Because the screeners' performance on detecting IEDs was already very high without the automated system (baseline), not much room was left for improvement. In Experiment 1, automated decision provided benefits only for the detection of explosives. This came at the cost of a higher false alarm rate, because all EDS alarms that are false alarms automatically add to the false alarms of screeners.

3. Experiment 2

The aims of Experiment 2 were to replicate Experiment 1 with screeners from a different airport, to address the limitations of

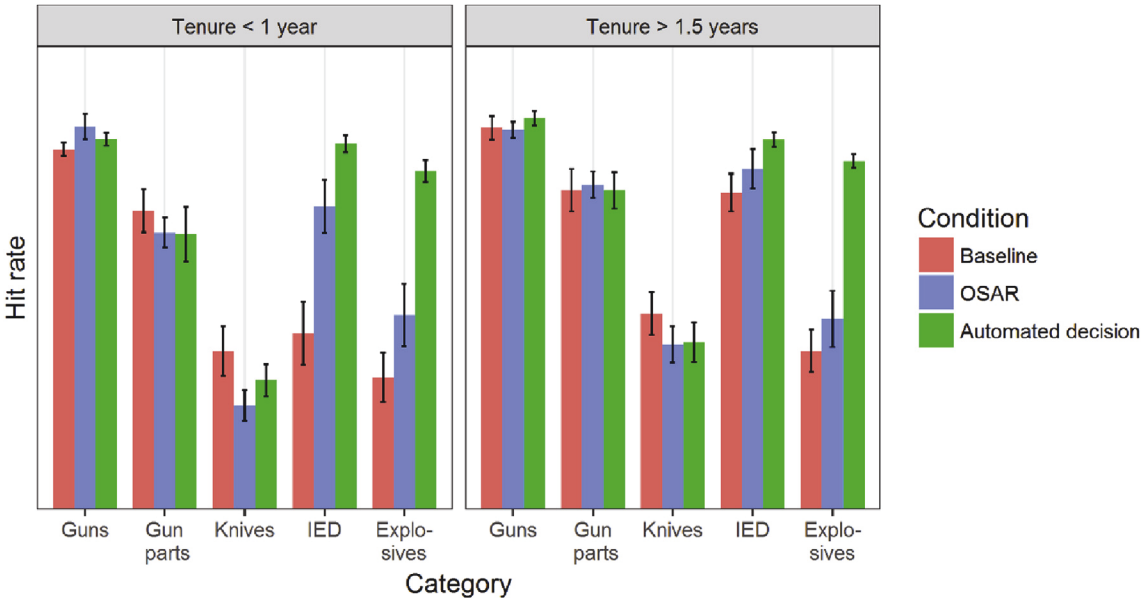

Fig. 5. Mean human–machine hit rates by condition (baseline, OSAR, automated decision), threat categories (gun, gun parts, and knives), and variation of work experience (tenure < 1 year and tenure > 1.5 years). Absolute hit rate values are not shown due to security restrictions in this project. Error bars are ± one standard error.

Experiment 1, and to test all three hypotheses.

The following limitations of Experiment 1 were addressed in Experiment 2: as described in the introduction, familiarity with automation can affect how people interact with it (Parasuraman and Manzey, 2010; Sauer et al., 2016; Strauch, 2016). Therefore, Experiment 2 was conducted at an international airport with screeners who were familiar with automation (this airport used EDSCB as diagnostic aid and screeners were familiar with OSAR). Moreover, screeners with less work experience and training might benefit when it comes to detecting IEDs and explosives in the OSAR condition due to their lower baseline performance (Halbherr et al., 2013). Therefore, Experiment 2 was conducted with two screener groups: experienced screeners (tenure > 1.5 years) and less experienced screeners (tenure < 1 year).

Experiment 2 addressed all three hypotheses: 1) As in Experiment 1, EDSCB should improve human–machine system performance for detecting bare explosives because these often look like a harmless organic mass (Jones, 2003). 2) We again expected better results for the automated decision scenario compared to OSAR, because clearing EDSCB alarms can be difficult (Jones, 2003) and because false alarm rates of 15–20% in the OSAR scenario may result in a cry wolf effect with screeners ignoring system warnings (Breznitz, 1983; Bliss, 2003). 3) Extending Experiment 1, we hypothesized for Experiment 2 that effects should depend on screener work experience because previous research has shown that regular computer-based training, which is mandatory in Europe, results in large increases of IED detection in the first few years (Halbherr et al., 2013).

3.1. Method

3.1.1. Participants

Experiment 2 was conducted with 77 screeners from another international European airport who were familiar with automation aids. As in Experiment 1, they had been qualified, trained, and certified according to the standards set by the appropriate national authority in compliance with the relevant EU Regulation (Commission Implementing Regulation [EU], 2015/1998). The screeners participated on a voluntary basis, were recruited by a security service provider at the airport, and compensated by regular salary. Informed consent was obtained from all participants. Group 1 (44 screeners, 14 females) was as well-trained and experienced as the screeners in Experiment 1 (years of work experience: M = 8.45 years, SD = 5.66). Their average age was 36.55 years (SD = 8.46, range 21–53 years). Group 2 (33 screeners, 19 females) had less work experience and training (less than one year). Their average age was 30.81 years (SD = 10.93, range 18–53[2] years).

3.1.2. Design

The experiment used a mixed design with condition (baseline, OSAR, automated decision) and years of work experience (tenure > 1.5 years or tenure < 1 year) as between-subjects independent variables and threat categories as within-subjects independent variables. The dependent variables were the hit rate (percentage detection of prohibited items) and false alarm rate of the human–machine system. As in Experiment 1, the three experimental groups were balanced according to their detection performance score in the pre-test (X-ray CAT) and the variables age and work experience within both tenure groups (> 1.5 years or < 1 year; baseline, tenure < 1: n = 10, tenure > 1.5: n = 14; OSAR, tenure < 1: n = 11, tenure > 1.5: n = 15; automated decision, tenure < 1: n = 12, tenure > 1.5: n = 15).

3.1.3. Materials, procedure, and statistics

The same tests and procedure were used as in Experiment 1. All participants completed the pre-test in less than 40 min and the main test in less than 2 h including breaks. The mean interval between the pre-test and the main test was 82.86 days (SD = 6.65). The same statistics were used as in Experiment 1.

3.2. Results

The same analyses were conducted as in Experiment 1 but with tenure as an additional between-subject factor. Fig. 5 shows human–machine system hit rates for both tenure groups by category and automation condition.

A two-way ANOVA on hit rates for the baseline condition with prohibited item category (guns, gun parts, knives, IEDs, and explosives) as within-subjects factor and work experience (tenure > 1.5 years vs. tenure < 1 year) as between-subjects factor revealed significant main

[2] Two screeners did not report their age.

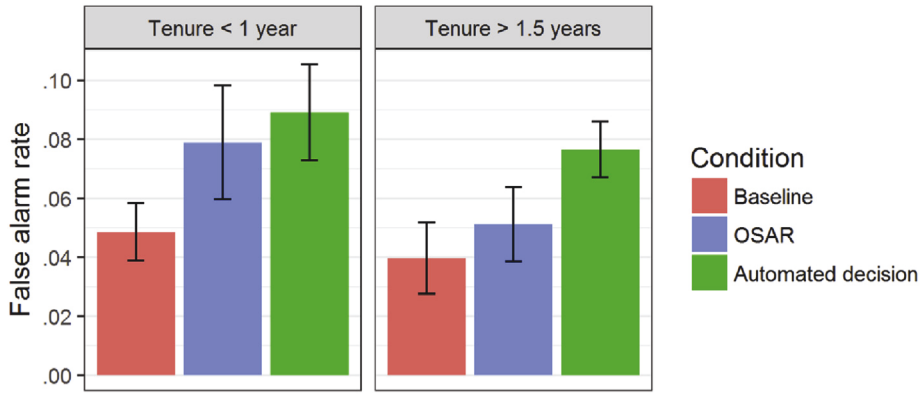

Fig. 6. Mean human–machine false alarm rates by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 1.5 years). Error bars are ± one standard error.

effects of the prohibited items category, $F(4, 88) = 63.38$, $p < .001$, $\eta_p^2 = 0.74$, and work experience, $F(1, 22) = 5.233$, $p = .032$, $\eta_p^2 = 0.19$, as well as their interaction, $F(4, 88) = 4.927$, $p < .001$, $\eta_p^2 = 0.12$. Post hoc pairwise comparisons were calculated separately for each screener group. For tenure > 1.5 years, there were significant comparisons between all threat categories ($p < .014$) except for those between gun parts and IEDs ($p = .943$) and between knives and explosives ($p = .277$). For < 1 year, all comparisons were significant ($p < .038$) except for those between knives and IEDs ($p = .699$), knives and explosives (p = .699), and IEDs and explosives ($p = .229$).

To rule out an effect of condition on the categories gun, gun parts, and knives, we calculated a 3 (prohibited item category: gun, gun parts, and knives) x 3 (condition: baseline, OSAR, and automated decision) x 2 (work experience: tenure > 1.5 years vs tenure < 1 year) ANOVA. This revealed no significant effect for condition, $F(2, 71) = 0.50$, $p = .610$, but a significant effect for work experience, $F(1, 71) = 8.07$, $p = .006$, $\eta_p^2 = 0.10$. This indicated that experienced screeners had a better detection performance on these three categories. Surprisingly, the interaction between category and condition was also significant, $F(3.88, 137.66) = 2.51$, $p = .047$, $\eta_p^2 = 0.07$. However, when we used the arcsine transformed scores, this effect no longer attained significance, $F(3.79, 134.57) = 2.37$, $p = .059$.

Furthermore, a 2 (categories: IEDs and explosives) x 3 (condition: baseline, OSAR, and automated decision) x 2 (work experience: tenure > 1.5 years vs tenure < 1 year) ANOVA for the hit rate revealed a significant effect of category, $F(1, 71) = 109.50$, $p < .001$, $\eta_p^2 = 0.61$, condition, $F(2, 71) = 39.28$, $p < .001$, $\eta_p^2 = 0.53$, and work experience, $F(1, 71) = 5.81$, $p = .019$, $\eta_p^2 = 0.08$, together with significant interactions between category and condition, $F(2, 71) = 15.66$, $p < .001$, $\eta_p^2 = 0.31$, and between category and work experience, $F(1, 71) = 9.55$, $p = .003$, $\eta_p^2 = 0.12$, as well as a significant three-way interaction, $F(2, 71) = 4.58$, $p = .014$, $\eta_p^2 = 0.11$. This shows that the effect of automation did not just depend on prohibited item category, but that this dependency also related to work experience.

In our next step, we calculated post hoc pairwise comparisons between the conditions within each screener group for IEDs and explosives separately. For IEDs, the less experienced screeners revealed a significant difference between the baseline condition and OSAR ($p = .039$) as well as between the baseline and automated decision ($p < .001$). In contrast, no comparison on the detection of IEDs was significant for experienced screeners. For explosives, there was a significant effect for the less experienced screeners between the baseline condition and automated decision ($p < .001$) as well as between OSAR and automated decision ($p < .001$). The same effects were found to be significant ($p < .001$) for explosives in experienced screeners.

Further analyses were conducted with false alarm rate as the dependent variable (see Fig. 6). A 3 (condition: baseline, OSAR, and automated decision) x 2 (work experience: tenure > 1.5 years vs tenure < 1 year) ANOVA revealed a significant effect for condition, $F(2, 71) = 4.043$, $p = .022$, $\eta_p^2 = 0.10$, but not for either work experience, $F(1, 71) = 2.19$, $p = .143$, or the interaction between work experience and condition, $F(2, 71) = 0.268$, $p = .76$. Moreover, post hoc pairwise comparisons within each screener group showed no significant difference between any two automation conditions.

Effect of OSAR. As reported above, the appearance of frames increased the hit rate for IEDs in less experienced screeners. Although there was no statistically significant increase in the false alarm rate between the baseline and OSAR condition, this does not mean per se that OSAR does not affect the false alarm rate in less experienced

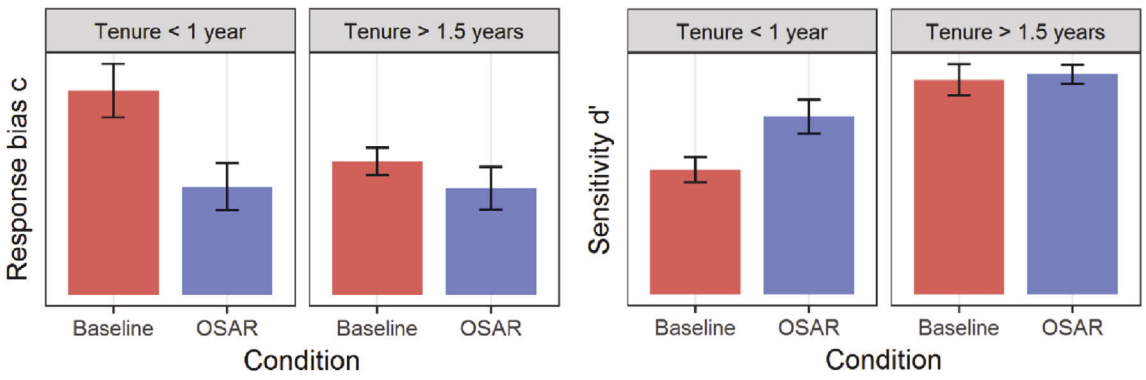

Fig. 7. (a) Mean response bias (*c*) of screeners by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 1.5 years). (b) Mean sensitivity measure *d'* of screeners by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 1.5 years). Error bars are ± one standard error.

screeners (see Fig. 6). Therefore, it is worth investigating whether there was a change in the response bias of the screeners (tendency to respond with NOT OK to images with frames) that can explain the increased hit rate for IEDs. In a next step, we compared the response bias *c* and the associated sensitivity measure *d'* (derived from signal detection theory using log-linear correction; Macmillan and Creelman, 2005) for IEDs between the baseline and OSAR condition in the less experienced screeners. This revealed an increase in both response bias, $t(18.33) = 2.68$, $p = .015$, $d = 1.18$, and sensitivity *d'*, $t(17.89) = -2.49$, $p = .023$, $d = 1.07$. Therefore, the results imply that OSAR leads to a higher sensitivity for detecting IEDs but is also responsible for a shift in response bias in less experienced screeners (see Fig. 7).

As in Experiment 1, we also tested whether human performance was affected by the implementation of automated decision by comparing only the images analysed by participants in both the baseline and automated decision condition. For the dependent variables hit rate for IEDs and hit rate for explosives, we calculated independent *t* tests separately for both experienced and less experienced screeners. Comparable to Experiment 1, automated explosives detection did not affect the detection of IEDs and explosives ($p > .182$). The same comparisons were made for false alarm rates, revealing no significant effects for either tenure group (tenure < 1 year: $t[20] = 0.27$, $p = .789$; tenure > 1.5 years: $t[27] = 0.15$, $p = .880$).

3.3. Discussion

In the baseline condition, the same results were found for well-trained and experienced screeners as in Experiment 1. Experiment 2 replicated the results from Experiment 1 while additionally revealing that screeners with less experience and training showed a lower detection of prohibited items than experienced screeners. This is consistent with previous research on the visual inspection of X-ray images without automation aids (Schwaninger and Hofer, 2004; Koller et al. 2008, 2009; Halbherr et al., 2013; Schuster et al., 2013). In the OSAR condition, results were as follows: as in Experiment 1, automation as a diagnostic aid (OSAR) did not increase detection performance for the experienced screeners in Experiment 2, despite their previous familiarity with such aids. The less experienced screeners detected more IEDs in the condition with OSAR, which was partly due to an increase in sensitivity and partly to a shift in response bias. The detection of explosives did not improve through OSAR. The use of automated decision resulted in the highest detection of explosives in both experienced and less experienced screeners. Less experienced screeners also detected the most IEDs in this condition, whereas it did not lead to any significant increase in experienced screeners. This was probably due to their already high level of performance as shown in the baseline condition. Regarding efficiency, results were consistent with Experiment 1; that is, automated decision resulted in a higher false alarm rate of the human–machine system, because screeners could not clear EDSCB alarms in this condition.

4. General discussion

This study examined the use of automation for the airport security screening of cabin baggage by testing two levels of automation that are currently being discussed by regulators and airport operators: on-screen alarm resolution (OSAR) and automated decision (Sterchi and Schwaninger, 2015). In the OSAR scenario, automated explosive detection systems for cabin baggage screening (EDSCB) assist airport security officers (screeners) by highlighting areas that could be explosive in X-ray images. This type of automation influences attention allocation and is comparable to diagnostic aiding used in other domains (Wickens and Dixon, 2007; Cullen et al., 2013). The automated decision scenario uses a higher level of automation and different human function allocation. Bags on which the EDSCB raises an alarm are sent automatically to secondary inspection, which involves manual search and/or explosive trace detection (Sterchi and Schwaninger, 2015). A simulated baggage screening task was used in two experiments with screeners working at two European airports who varied in their work experience. As expected, human–machine system performance varied between the two scenarios. In the following, we discuss both implementation scenarios in terms of their human–machine system performance.

4.1. Automation as diagnostic aid (OSAR)

Previous research has shown that fully functional improvised explosive devices (IEDs) can be detected very well by experienced and trained screeners even without automation (Schwaninger and Hofer, 2004; Koller et al., 2008, 2009; Halbherr et al., 2013). However, detecting bare explosives proves to be a challenge even for experienced screeners, because they often look like a harmless organic mass (Jones, 2003). Indeed, with automation as a diagnostic aid (OSAR), human–machine hit rates for bare explosives were similar to the baseline condition without automation. This is remarkable when it is considered that for OSAR, the EDSCB has a hit rate of 88% for explosives. In other words, using automation as a diagnostic aid, which means that screeners have to resolve EDSCB alarms, drastically reduces or even eliminates the benefits of EDSCB for detecting bare explosives.

However, the OSAR scenario is beneficial for the detection of IEDs but only for the less experienced screeners. We argue that the automation system with OSAR assists in the search component of X-ray image inspection by guiding attention (Cullen et al., 2013) to the relevant area – the first processing stage of sensory processing in the taxonomy proposed by Parasuraman et al. (2000). OSAR can further assist by providing relevant information and therefore support the decision component (i.e. an X-ray image that triggers an alarm is more likely to contain an IED or explosive). As explained, the main difference between IEDs and bare explosives is that screeners can learn to recognize IED components (triggering device, power source, detonator, and cables connecting these components to an explosive) in an X-ray image (Turner, 1994). In the presence of these components, less experienced screeners are able to profit from the attentional guidance provided by OSAR and increase their hit rate. Our further investigation of the increased hit rate for IEDs revealed an increase in sensitivity and simultaneously a decrease in response bias. This suggests that the automation system affects not only the visual search component but also the decision component in the less experienced screeners' inspection.

But, why did experienced screeners not profit from attentional guidance through OSAR? First, experienced screeners already achieved high hit rates for IEDs in the baseline condition without automation and this thereby does not leave much room for improvement through OSAR. In addition, experienced screeners may also have judged their own ability to detect prohibited items to be superior to the automation support – a reason for noncompliance also reported in other domains (e.g. Lee and Moray, 1992, 1994). However, as even experienced screeners could not profit from OSAR in regard to explosives, future research should explore whether specific training and familiarity with automation aids (Sauer et al., 2016) such as OSAR might provide screeners with a mental model of its capabilities. Such mental models could be crucial for an effective use of the automation aid (Strauch, 2016). Moreover, the low target prevalence in our study and, therefore, the low base rate led to many false alarms (Parasuraman and Riley, 1997). This probably led to a 'cry wolf' effect with experienced screeners, meaning that they might simply have ignored the system warnings (Breznitz, 1983; Bliss, 2003). This problem should be even more pronounced in practice where real IEDs and explosives almost never occur and almost all EDSCB alarms are false.

4.2. Automation as automated decision

We expected better results for the automated decision scenario

compared to OSAR, because clearing EDS alarms can be difficult (Jones, 2003) and the EDSCB false alarm rate of 17% in the OSAR scenario could result in a cry wolf effect with screeners ignoring system warnings (Breznitz, 1983; Bliss, 2003). Indeed, in both experiments, we found that screeners did not achieve high hit rates for bare explosives. However, EDSCB with automated decision was able to compensate for this, leading to better human–machine hit rates in both airports and both tenure groups. This came at the expense of higher false alarm rates (an increase by ca. 4 percentage points) – a rate that is still operationally feasible.

Because there was no direct interaction between the automation system and the screener, it is not surprising that the automated decision did not affect screener performance. Hence, the observed increase in detection performance was determined by the amount of explosives missed by screeners but detected by the EDSCB. This also explains why the detection of IEDs improved significantly only in less experienced screeners. As shown in the baseline condition, experienced screeners already detected IEDs well, and this left little room for improvement through EDSCB. As expected, automated decision showed a higher false alarm rate. Assuming that screener performance remains unaffected by the implementation of an automated decision when applying different hit and false alarm rates to the ones tested in this study, system hit and false alarm rates can be manipulated directly by the choice of the EDSCB machine and the machine settings (criterion of the machine) for a given screener performance. It is important to remember that with the EDSCB threshold settings used in our experiments, humans (screeners) still have an important role. They visually inspect all X-ray images on which the EDSCB does not raise an alarm. This would be 96% of all X-ray images in an operational environment (as the EDSCB alarms only on 4% of all bags).

4.3. Practical implications, limitations, and future research

Replication of psychological experiments is an important part of the scientific process – particularly in psychology (Rovenpor and Gonzales, 2015; Baker, 2016). This is why we regard the replication aspect of Experiment 2 as a specific strength. However, in addition to the replication, the effects in Experiment 2 also depend on screener work experience, as to be expected from previous research showing that regular computer-based training results in large increases of IED detection in the first few years (Halbherr et al., 2013).

Like most previous studies on visual inspection and automation, this study also uses laboratory experiments that simulate aspects of tasks that human operators conduct in the real world. Therefore, it is important to consider both the limitations of such simulations and their practical implications when discussing the similarities and differences between the baggage screening task used in this study and X-ray screening at airport security checkpoints. One difference is that airport security checkpoints are often noisy and stressful environments (Michel et al., 2014; Baeriswyl et al., 2016). Research in other domains (e.g. Sauer et al., 2013) has found that operators prefer higher levels of automation under noise than in quiet conditions. If this also proves to be the case for cabin baggage screening, it would generate further evidence in favor of automated decision instead of diagnostic automation (OSARP). Another difference is target prevalence; that is, the base rate of target-present events (Wolfe et al., 2007). In our study, one out of eight images contained a threat item and one out of 20 images either an IED or explosive. In practice, such threats are much less frequent. Assuming that airports conduct covert tests (Schwaninger, 2009) and use threat image projection, a technology that projects X-ray images containing threats during the routine X-ray screening operation (Hofer and Schwaninger, 2005), target prevalence would be about 2%. With regard to our findings, two expected effects of lower target prevalence need to be discussed. The first effect is that lower target prevalence probably leads to a shift in decision bias and therefore lower hit and false alarm rates in screeners (Wolfe et al., 2007, 2013). If detection of IEDs by screeners is lower in practice, this will leave more room for improvement through EDSCB. The second and much more important effect of lower target prevalence is a decrease in the positive predictive value (Meyer et al., 2014) of the EDSCB with OSAR. As a result, in practice, EDSCB alarms are very often false alarms. This accentuates the problem of the cry wolf effect and makes the successful implementation of OSAR more challenging. Another limitation of this study is the fact that it used single view imaging. This was because the participating screeners from the two European airports only had experience with single view X-ray machines. It would be interesting for a follow-up study to explore whether results would be different when using multi-view X-ray imaging.

Future research could also explore whether specific training and familiarity with the automation aid (Sauer et al., 2016) might provide screeners with a mental model of its capabilities. Such mental models could be important for an effective use of an automation aid (Strauch, 2016). These mental models could also be supported by artificially increasing the presence of IEDs and explosives in operation that interact with EDSCB in a realistic way by carrying out covert tests (Schwaninger, 2009) and using threat image projection (Hofer and Schwaninger, 2005) more frequently. Future studies should also use real EDSCB false alarms from an operational environment because screeners might learn to correctly resolve certain types of false alarms (e.g. those caused by certain types of harmless items).

Comparing automation as a diagnostic aid and a higher level of automation with automated decision could also be important in other areas such as diagnostic radiology in medicine. For example, automation as a diagnostic aid is also used for early detection of breast cancers from mammograms (e.g. Vyborny et al., 2000; Astley, 2004; Giger, 2004; Fenton et al., 2007). This task shares features with X-ray baggage screening that are relevant for selecting the appropriate level of automation such as imperfect automation performance, the prominence of false alarms due to a low target prevalence, and the potentially severe consequences associated with misses (Sampat et al., 2005; Nishikawa, 2007). Future research in different fields might provide a more detailed understanding of the optimal degree of automation depending on human and machine performance in different stages of information processing.

4.4. Conclusion

We investigated the benefits of automation for airport security screening of cabin baggage using two levels of automation that are currently being discussed by regulators and airport operators. Our three research questions can be answered as follows: We found that EDSCB improves human–machine system performance for detecting bare explosives. When comparing the two levels of automation, human–machine system performance using automated decision proved to be superior to automation as a diagnostic aid. EDSCB with automated decision has the potential to greatly increase the detection of explosives, but at the expense of some efficiency – depending on the criterion setting of the EDS algorithms. EDSCB as a diagnostic aid is false-alarm prone and results in a cry wolf effect with experienced screeners ignoring the system warnings; it is only beneficial for screeners with limited experience. Our results indicate that the wide-scale implementation of EDSCB can be recommended because it can greatly improve the detection of explosives in cabin baggage. The advantage of automated decision over automation as a diagnostic aid should be investigated further by also carrying out operational trials at airport security checkpoints.

Funding

This study was funded by the Ministry of Security and Justice, National Coordinator for Security and Counterterrorism, the Netherlands, and by the University of Applied Sciences and Arts

Northwestern Switzerland.

Acknowledgements

The authors particularly acknowledge the valuable contribution of Milena Kuhn throughout the entire course of the project. Further, we thank aviation security experts from the German Federal Police Technology Centre as well as the National Coordinator for Security and Counterterrorism (NCTV - Ministry of Justice and Security, The Netherlands) for their valuable expertise and support. We also thank NTCB Security Training Centre Netherlands for their help in recruiting screeners and data collection.

References

Abadie, A., Gardeazabal, J., 2008. Terrorism and the world economy. Eur. Econ. Rev. 52, 1–27.
Abe, G., Richardson, J., 2006. Alarm timing, trust and driver expectation for forward collision warning systems. Appl. Ergon. 37 (5), 577–586.
Astley, S.M., 2004. Computer-based detection and prompting of mammographic abnormalities. BJR (Br. J. Radiol.) 77. https://doi.org/10.1259/bjr/30116822.
Baeriswyl, S., Krause, A., Schwaninger, A., 2016. Emotional exhaustion and job satisfaction in airport security officers - work-family conflict mediator in the job demands-resources model. Front. Psychol. 7, 1–13. http://dx.doi.org/10.3389/fpsyg.2016.00663.
von Bastian, C.C., Schwaninger, A., Michel, S., 2009. The impact of color composition on X-ray image interpretation in aviation security screening. In: Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology, Zurich Switzerland, http://dx.doi.org/10.1109/CCST.2009.5335539. October 5–8, 2009.
Baum, P., 2016. Violence in the Skies: a History of Aircraft Hijacking and Bombing. Summersdale Publishers, Chichester, England.
Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533 (7604), 452–454. http://dx.doi.org/10.1038/533452a.
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 (1), 142–148. http://dx.doi.org/10.1002/acp.3083.
Biondi, F., Strayer, D.L., Rossi, R., Gastaldi, M., Mulatti, C., 2017. Advanced driver assistance systems: using multimodal redundant warnings to enhance road safety. Appl. Ergon. 58, 238–244.
Bliss, J., 2003. An investigation of alarm related accidents and incidents in aviation. Int. J. Aviat. Psychol. 13, 249–268.
Bliss, J., Dunn, M., Fuller, B.S., 1995. Reversal of the cry-wolf effect: an investigation of two methods to increase alarm response rates. Percept. Mot. Skills 80, 1231–1242.
Breznitz, S., 1983. Cry-wolf: the Psychology of False Alarms. Erlbaum, Hillsdale, NJ.
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.
Cullen, R.H., Rogers, W.A., Fisk, A.D., 2013. Human performance in a multiple-task environment: effects of automation reliability on visual attention allocation. Appl. Ergon. 44, 962–968.
Erceg-Hurn, D.M., Mirosevich, V.M., 2008. Modern robust statistical methods. Am. Psychol. 63 (7), 591–601.
Fenton, J.J., Taplin, S.H., Carney, P.A., Abraham, L., Sickles, E.A., D'orsi, C., Elmore, J.G., 2007. Influence of computer-aided detection on performance of screening mammography. N. Engl. J. Med. 356, 1399–1409. Retrieved from. http://www.nejm.org/doi/pdf/10.1056/NEJMoa066099.
Giger, M.L., 2004. Computerized analysis of images in the detection and diagnosis of breast cancer. Seminars Ultrasound, CT MRI 25 (5), 411–418.
Gillen, D., Morrison, W.G., 2015. Aviation security: costing, pricing, finance and performance. J. Air Transport. Manag. 48, 1–12. https://doi.org/10.1016/j.jairtraman.2014.12.005.
Glass, G.V., Peckham, P.D., Sanders, J.R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev. Educ. Res. 42, 237–288.
Global Terrorism Database, 2017. https://www.start.umd.edu/gtd (Accessed 15 November 2017).
Green, D.M., Swets, J.M., 1966. Signal Detection Theory and Psychophysics. Wiley and Sons, New York, NY.
Green, D.M., Swets, J.M., 1972. Signal Detection Theory and Psychophysics (Revised Edition). Krieger, New York, NY.
Halbherr, T., Schwaninger, A., Budgell, G.R., Wales, A., 2013. Airport security screener competency: a cross-sectional and longitudinal analysis. Int. J. Aviat. Psychol. 23 (2), 113–129.
Harwell, M.R., Rubinstein, E.N., Hayes, W.S., Olds, C.C., 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J. Educ. Behav. Stat. 17 (4), 315–333. http://dx.doi.org/10.3102/10769986017004315.
Hofer, F., Schwaninger, A., 2005. Using threat image projection data for assessing individual screener performance. WIT Trans. Built Environ. 82, 417–426. http://dx.doi.org/10.2495/SAFE050411.
Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.
Jones, T.L., 2003. Court Security: a Guide for Post 9-11 Environments. Charles C. Thomas, Springfield, IL.
Kaber, D.B., Endsley, M.R., 2003. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theor. Issues Ergon. Sci. 5 (2), 113–153. http://dx.doi.org/10.1080/1463922021000054335.
Koller, S., Drury, C., Schwaninger, A., 2009. Change of search time and non-search time in X-ray baggage screening due to training. Ergonomics 52 (6), 644–656.
Koller, S., Hardmeier, D., Michel, S., Schwaninger, A., 2007. Investigating training and transfer effects resulting from recurrent CBT of x-ray image interpretation. In: McNamara, D.S., Trafton, J.G. (Eds.), Proceedings of the 29th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 1181–1186.
Koller, S., Hardmeier, D., Michel, S., Schwaninger, A., 2008. Investigating training, transfer and viewpoint effects resulting from recurrent CBT of x-ray image interpretation. J. Transport. Secur. 1 (2), 81–106.
Koller, S., Schwaninger, A., 2006. Assessing X-ray image interpretation competency of airport security screeners. In: Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, pp. 399–402 June 24–28.
Lee, J.D., Moray, N., 1992. Trust, control strategies, and allocation of function in human machine systems. Ergonomics 22, 671–691.
Lee, J.D., Moray, N., 1994. Trust, self-confidence, and operators' adaptation to automation. Int. J. Hum. Comput. Stud. 40, 153–184.
Lehto, M.R., Papastavrou, J.D., Ranney, T.A., Simmons, L.A., 2000. An experimental comparison of conservative versus optimal collision avoidance warning system thresholds. Saf. Sci. 36 (3), 185–209.
Liu, Y.-C., Jhuang, J.-W., 2012. Effects of in-vehicle warning information displays with or without spatial compatibility on driving behaviors and response performance. Appl. Ergon. 43 (4), 679–686.
Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: a User's Guide, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.
McDonald, J.H., 2007. The Handbook of Biological Statistics. Sparky House Publishing, Baltimore, MD. http://udel.edu/~mcdonald/statpermissions.html, Accessed date: 2 March 2011.
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? In: Proceedings of the 45th IEEE International Carnahan Conference on Security Technology, Mataro Spain, October 18–21, 2011.
Meyer, J., Wiczorek, R., Günzler, T., 2014. Measures of reliance and compliance in aided visual scanning. Hum. Factors 56 (5), 840–849.
Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., Schwaninger, A., 2014. A multimethod approach towards identifying situational factors and their relevance for x-ray screening. In: Proceedings of the 48th IEEE International Carnahan Conference on Security Technology, Rome Italy, pp. 208–213. http://dx.doi.org/10.1109/CCST.2014.6987001. October 13-16.
Michel, S., Koller, S., de Ruiter, J., Moerland, R., Hogervorst, M., Schwaninger, A., 2007. Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners. In: Proceedings of the 41st IEEE International Carnahan Conference on Security Technology, Ottawa Canada, October 8–11.
Mitroff, S.R., Biggs, A.T., Cain, M.S., 2015. Multiple-target visual search errors: overview and implications for airport security. Pol. Insights Behav. Brain Sci. 2 (1), 121–128.
Nabiev, S.S., Palkina, L.A., 2017. Modern technologies for detection and identification of explosive agents and devices. Russ. J. Phys. Chem. B 11, 729–776. https://doi.org/10.1134/S1990793117050190.
Neffenger, P.V., 2015, October 22. Advanced Integral Passenger and Baggage Screening Technologies. Fiscal Year 2015 Report to Congress. US Department for Homeland Security, Transportation Security Administration Retrieved from. https://www.dhs.gov.
Nishikawa, R.M., 2007. Current status and future directions of computer-aided diagnosis in mammography. Comput. Med. Imag. Graph. 31 (4–5), 224–235. http://dx.doi.org/10.1016/j.compmedimag.2007.02.009.
Novakoff, A.K., 1993. FAA bulk technology overview for explosives detection. SPIE 1824, 2–12.
Parasuraman, R., 1987. Human-computer monitoring. Hum. Factors 29, 695–706.
Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: an attentional integration. Hum. Factors 52 (3), 381–410.
Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Hum. Factors: J. Hum. Factors Ergon. Soc. 39 (2), 230–253. http://dx.doi.org/10.1518/001872097778543886.
Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Syst. Hum. 30 (3), 286–297.
Parasuraman, R., Wickens, C.D., 2008. Humans: still vital after all these years of automation. Hum. Factors 50 (3), 511–520. http://dx.doi.org/10.1518/001872008X312198.
Pochet, G., 2016, August 22. Smart Security: Alternative Detection Methods and Unpredictability. ACI World Report - August 2016. Retrieved from. https://issuu.com/aciworld/docs/aci-world-report_august-2016/22.
R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rice, S., McCarley, J., 2011. Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. J. Exp. Psychol. Appl. 17 (4), 320–331.
Rovenpor, D.R., Gonzales, J.E., 2015, January. Replication in psychological science: challenges, opportunities, and how to participate in the replication process. Psychol.

Sci. Agenda 29 (1) Retrieved from. http://www.apa.org/science/about/psa/2015/01/replicability.aspx.
Sampat, M.P., Markey, M.K., Bovik, A.C., 2005. Computer-aided detection and diagnosis in mammography. In: Bovik, A.C. (Ed.), The Handbook of Image and Video Processing, second ed. Elsevier, New York, pp. 1195–1217.
Sauer, J., Chavaillaz, A., 2017. The use of adaptable automation: effects of extended skill lay-off and changes in system reliability. Appl. Ergon. 58, 471–481.
Sauer, J., Chavaillaz, A., Wastell, D., 2016. Experience of automation failures in training: effects on trust, automation bias, complacency, and performance. Ergonomics 59 (6), 767–780.
Sauer, J., Nickel, P., Wastell, D., 2013. Designing automation for complex work environments under different levels of stress. Appl. Ergon. 44 (1), 119–127.
Schuster, D., Rivera, J., Sellers, B.C., Fiore, S.M., Jentsch, F., 2013. Perceptual training for visual search. Ergonomics 56 (7), 1101–1115. http://dx.doi.org/10.1080/00140139.2013.790481.
Schwaninger, A., 2004. Computer based training: a powerful tool to the enhancement of human factors. Aviat Secur. Int. 2, 31–36.
Schwaninger, A., 2005. Increasing efficiency in airport security screening. WIT Trans. Built Environ. 82, 407–416.
Schwaninger, A., 2006. Airport security human factors: from the weakest to the strongest link in airport security screening. In: Proceedings of the 4th International Aviation Security Technology Symposium, Washington, DC, pp. 265–270.
Schwaninger, A., 2009. Why do airport security screeners sometimes fail in covert tests? In: Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology, Zurich Switzerland, http://dx.doi.org/10.1109/CCST.2009.5335568. October 5–8.
Schwaninger, A., Hardmeier, D., Hofer, F., 2005. Aviation security screeners visual abilities & visual knowledge measurement. IEEE Aero. Electron. Syst. 20 (6), 29–35.
Schwaninger, A., Hofer, F., 2004. Evaluation of CBT for increasing threat detection performance in X-ray screening. In: Morgan, K., Spector, M.J. (Eds.), The Internet Society 2004, Advances in Learning, Commerce and Security. WIT Press, Ashurst, England, pp. 147–156. http://dx.doi.org/10.13140/RG.2.1.4051.8649.
Sheridan, T.B., 2011. Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. IEEE Trans. Syst. Man Cybern. Syst. Hum. 41 (4), 662–667.
Sheridan, S., Verplank, W., 1978. Human and Computer Control of Undersea Teleoperators. Technical Report. MIT Man–Machine Systems Laboratory, Cambridge, MA.
Singh, S., Singh, M., 2003. Explosives detection systems (EDS) for aviation security. Signal Process. 83 (1), 31–55.
Steiner-Koller, S.M., Bolfing, A., Schwaninger, A., 2009. Assessment of x-ray image interpretation competency of aviation security screeners. In: Proceedings of the 43rd IEEE Carnahan Conference on Security Technology, Zurich Switzerland, http://dx.doi.org/10.1109/CCST.2009.5335569. October 5–8.
Sterchi, Y., Schwaninger, A., 2015. A first simulation on optimizing EDS for cabin baggage screening regarding throughput. In: Proceedings of the 49th IEEE International Carnahan Conference on Security Technology, Taipei Taiwan, http://dx.doi.org/10.1109/CCST.2015.7389657. September 21–24.
Strauch, B., 2016. The automation-by-expertise-by-training interaction. Why automation-related accidents continue to occur in sociotechnical systems. Hum. Factors 59 (2), 204–228. http://dx.doi.org/10.1177/0018720816665459.
Thomas, A., 2009. Aviation Security Management. Available at: http://works.bepress.com/andrew_thomas2/20.
Turner, S., 1994. Terrorist Explosive Sourcebook Countering Terrorist Use of Improvised Explosive Devices. Paladin Press, Boulder CO.
Vagia, M., Transeth, A.A., Fjerdingen, S.A., 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Appl. Ergon. 53, 190–202. https://doi.org/10.1016/j.apergo.2015.09.013.
Vyborny, C.J., Giger, M.L., Nishikawa, R.M., 2000. Computer-aided detection and diagnosis of breast cancer. Radiol. Clin. 38 (4), 725–740.
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. Behav. Res. Meth. 41 (3), 937–943. http://dx.doi.org/10.3758/BRM.41.3.937.
Wells, K., Bradley, D.A., 2012. A review of X-ray explosives detection techniques for checked baggage. Appl. Radiat. Isot. 70 (8), 1729–1746. http://dx.doi.org/10.1016/j.apradiso.2012.01.011.
Westbrook, T., Barrett, J., 2017, August 4. Islamic State behind Australians' Foiled Etihad Meat-mincer Bomb Plot: Police. Reuters Retrieved from. https://www.reuters.com/article/us-australia-security-raids/islamic-state-behind-australians-foiled-etihad-meat-mincer-bomb-plot-police-idUSKBN1AJ367.
Wickens, C.D., Colcombe, A., 2007. Performance consequences of imperfect alerting automation associated with a cockpit display of traffic information. Hum. Factors 49, 839–850.
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. http://dx.doi.org/10.1080/14639220500370105.
Wiegmann, D., McCarley, J.S., Kramer, A.F., Wickens, C.D., 2006. Age and automation interact to influence performance of a simulated luggage screening task. Aviat Space Environ. Med. 77 (8), 825–831.
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. J. Vis. 13 (3), 33. http://dx.doi.org/10.1167/13.3.33.
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.
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. http://dx.doi.org/10.1016/j.cub.2009.11.066.

A First Simulation on Optimizing EDS for Cabin Baggage Screening Regarding Throughput

Yanik Sterchi and Adrian Schwaninger
School of applied Psychology
University of Applied Sciences and Arts Northwestern Switzerland (FHNW)
Olten, Switzerland
and
Center for Adaptive Security Research and Applications (CASRA)
Zurich, Switzerland

***Abstract*—Airport security screening is vital for secure air transportation. Screening of cabin baggage heavily relies on human operators reviewing X-ray images. Explosive detection systems (EDS) developed for cabin baggage screening can be a very valuable addition security-wise. Depending on the EDS machine and settings, false alarm rates increase, which could reduce throughput. A discrete event simulation was used to investigate how different machine settings of EDS, different groups of X-ray screeners, and different durations of alarm resolution with explosives trace detection (ETD) influence throughput of a specific cabin baggage screening process. For the modelling of screening behavior in the context of EDS and for the estimation of model parameters, data was borrowed from a human-machine interaction experiment and a work analysis. In a second step, certain adaptations were tested for their potential to reduce the impact of EDS on throughput. The results imply that moderate increases in the false alarm rate by EDS can be buffered by employing more experienced and trained X-ray screeners. Larger increases of the false alarm rate require a fast alarm resolution and additional resources for the manual search task.**

Keywords—aviation security; explosive detection systems (EDS); human factors; discrete event simulation; throughput

I. Introduction

A secure air transportation system is essential for society and economy. Repeated terror attacks [1] have led to increased aviation security measures. All passengers and their belongings are screened at an airport security checkpoint (ASC) to ensure that they are not carrying any prohibited items (guns, knives, improvised explosives (IEDs), and other threat items). At a typical ASC passengers first have to divest themselves of their luggage and other items like pocket content, jackets, and headwear. These are then scanned by an X-ray machine and an airport security officer (ASO) searches the X-ray images for prohibited items. If a suspicious item is detected, the ASO reviewing the X-ray images (X-ray screener) declares the image as "not OK" (which is also referred to as an “alarm”) and the corresponding bag or item is redirected for further inspection (secondary search). This is referred to as "alarm resolution" and typically includes a manual search of the bag and/or explosives trace detection (ETD), which consists of taking a swab at several locations of the bag and then having a machine analyze that swab for traces of explosive residue [2], [3]. If the suspicious item turns out to actually be a prohibited item, the X-ray screener's alarm (i.e. declaring the item as "not OK") is called a "hit". If the suspicious item turns out to be harmless, the X-ray screener’s alarm is called a "false alarm". As described, both technology and humans are involved in the process. How well the whole system performs therefore depends on human factors, machine attributes, and the process defining their interaction [4]–[7].

In recent years, the detection of explosives has been increasingly in the focus of security in civil aviation [3]. Manufacturers of detection equipment have recognized the increasing threat by explosives and the difficulty to detect them without additional technological assistance [8]. Singh and Singh [9] or Wells and Bradley [3] provide a good overview of different technologies developed for the detection of explosives. As Sing and Singh [9] point out, X-ray technology is the most common method for luggage inspection at airports. In recent years, X-ray based EDS machines have been made available for cabin baggage screening. Such machines can use dual energy X-ray imaging to detect explosives via the estimation of the effective atomic number and material density [8], [9].

The introduction of EDS into cabin baggage screening is certainly an advantage security wise. But how does EDS affect throughput, i.e. the amount of items that can be screened within a certain time? Butler and Poole [2] argued that EDS can reduce throughput, but since then EDS machines have become faster and more reliable. It would therefore be interesting to examine effects of EDS on throughput taking into account up-to-date information on technology, humans and processes. In this study this was explored for one specific process using discrete event simulation. In addition, two measures to cope with potential negative effects on throughput were tested for their effectiveness.

II. Process Description and Assumptions

System performance of an ASC depends on technology, humans, and the process defining their interaction. The difference between a conventional X-ray machine and an X-ray machine with EDS is that the latter analyzes the X-ray image information for potential explosives before the X-ray image is displayed to the X-ray screener. The analysis by the EDS can

be quite fast and does not necessarily delay the reviewing of the X-ray images by the X-ray screener. Once the EDS generates an alarm, there are at least two different approaches to resolve these alarms. One is on-screen alarm resolution: When the ASO reviews the X-ray image of the bag that triggered the alarm by the EDS, a frame is displayed around the area of the X-ray image which might contain explosives. The ASO then decides whether the bag needs further alarm resolution (e.g. using ETD and/or manual search). Another approach is to increase the level of automation (for an overview of levels of automation see [10]) and *automatically* redirect items that caused an alarm for alarm resolution by ETD and/or manual search. In this study we restricted ourselves to this second approach, which will be referred to as "automatic decision scenario". Further, the alarm resolution process has to be specified. It seems to be more adequate to resolve alarms by the EDS using ETD instead of manual search, as ETD is specialized for detecting explosives. But ETD is not suited for detecting other prohibited items. If an X-ray image triggers an alarm by the EDS, the X-ray screener would still need to review this image for prohibited items other than IEDs and explosives and pass it on for manual search, if required. If for example there was a knife in a bag that set off the alarm by the EDS and this alarm was only resolved with ETD, the knife would pass undetected. For this study the process is specified as illustrated in Fig. 1.

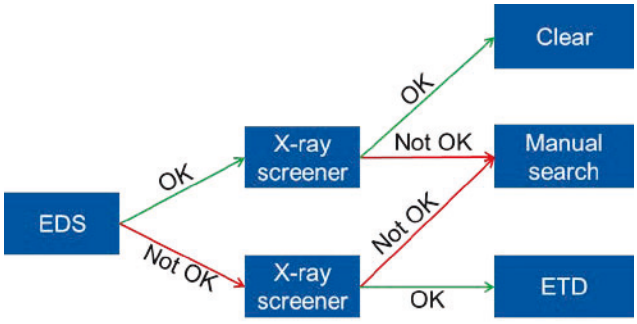

Fig. 1. Alarm resolution depending on alarm of EDS and X-ray screener. Note: Instead of only manual search, an X-ray screener could also decide for a manual search in combination with ETD, separation (i.e. unpacking the bag and X-ray screening certain items separately), or rescreening the bag with a different placement on the conveyor belt.

A. Relevant Variables of Machine and Human

With regard to ASC throughput the most important attributes of an EDS machine seem to be its ability to detect explosives, the amount of false alarms it generates, and how long it needs to process a bag. Comparable to an alarm by the X-ray screener, an alarm by the EDS can either be a hit (if an explosive is present) or a false alarm (if no explosive is present). If the EDS does not generate an alarm for a certain bag or item, this is called a "correct rejection" in the case that the decision was correct and no explosives were present and a "miss" in the other case. How likely these events are, depends of course on the machine, but also on the machine settings. If a more sensitive machine setting is chosen, the frequency of an alarm for bags containing explosives - the so called "hit rate" - rises. At the same time the frequency of an alarm for bags not containing any explosives - the so called "false alarm rate" - also increases. Choosing a more sensitive machine setting therefore leads to more alarms that need resolution. The hit rate of an EDS should not directly influence throughput, as real explosives are rarely encountered in operation. But the hit rate needs to meet regulatory requirements and the false alarm rate of an EDS machine can only be reduced to the point where its hit rate still meets the regulatory requirements. The false alarm rate of an EDS therefore depends on the intended hit rate and the machine itself, as different machines can have different false alarm rates for a predefined hit rate. This allows for a wide range of possible false alarm rates for EDSs. Authorities, who test and certify EDS machines, reported to us possible values of false alarm rates between 6% and 18.6%. But even lower false alarm rates are possible if only a share of the items is analyzed by the EDS. E.g. an EDS with a false alarm rate of 6% could randomly analyze half the items, which would then result in a false alarm rate of 3%. In this study, EDS false alarm rates ranging from 1% to 15% were considered.

To investigate the effect of EDS on ASC throughput other variables should also be taken into account, which are not directly related to the EDS machine's performance with regard to its false alarm rate and evaluation time. In our automatic decision scenario, false alarms by the EDS have to be resolved with an ETD conducted by the same ASO that also resolves alarms by the X-ray screener. The first and quite obviously relevant aspect of this process is the time needed for using an ETD to resolve the alarm by the EDS. It should be noted that modern ETD technology is fast, for example [11] report 5-8 s for their IONSCAN 500DT ETD machine. However, the overall time needed for alarm resolution using ETD depends strongly on where and how many trace samples are taken [2]. To take into account that different durations for alarm resolution with ETD are possible, three different scenarios were tested in the current study: One with a low average duration of 30 s, a second taking 60 s, and a third taking 120 s on average.

Beside the duration of the ETD, a second aspect is important with regard to resolving the alarms by the EDS: The number of the alarms by the X-ray screener. The more alarms (both false alarms and detected prohibited items, e.g. liquids or scissors) the X-ray screener generates, the more busy the ASO responsible for alarm resolution by secondary search will be and the less capacity he or she will have to resolve alarms by the EDS. There are several factors known to produce substantial variance in X-ray screeners' hit and false alarm rates. Especially important is initial and recurrent individually adaptive computer-based training, which has been shown to be an effective and efficient tool to learn which items are prohibited and what they look like in X-ray images [7], [12]–[14]. Such training has also been shown to reduce false alarm rates [6]. Several studies did not find an effect for job experience alone, if not accompanied by training [5], [15]. Age was found to have a negative effect [5], but a rather small one compared to the effect of training [16]. In order to take into account potential influences of differences in training, age, and job experience of X-ray screeners, the simulation will be based on data from three different X-ray screener populations, which vary with regard to training hours, job experience, age, and airport.

So far, attributes of technology and humans which seem the most important for the effect of EDS on throughput have been

discussed. A third aspect to consider is human-machine interaction. To test whether X-ray screeners reduce their false alarm rate when EDS is available, [17] conducted a laboratory experiment, where 150 certified X-ray screeners had to review X-ray images either with on-screen alarm resolution, automatic decision, or without any assistance by an EDS[1]. There were no significant differences in X-ray screeners' false alarm rates or evaluation times between the baseline condition without assistance and the condition with automatic decision. For the simulations of the present study we will therefore assume that the performance of X-ray screeners is not affected by EDS.

As explained above, the false alarm rate of the EDS machine, the false alarm rate of the X-ray screener, and also the duration for applying ETD are crucial for the effect of EDS on throughput in our automatic decision scenario, because these three factors affect the workload of the ASO responsible for resolving the alarms. In this study, two possible measures were examined on their effectiveness to reduce the workload of this ASO. The first and quite obvious measure is to assign a second ASO to the task of resolving alarms using manual search and/or ETD. This should double the rate at which alarms can be resolved (assuming there is sufficient room and equipment provided). Typically there is a limit to the number of items that can queue for alarm resolution. If this queue limit is reached, the X-ray screening process is interrupted until the responsible ASO has finished resolving one of the alarms and the queue is below its limit again. A second possible measure to expedite secondary search could be to instruct the X-ray screener to resolve one of the alarms as soon as the queue limit is reached and the screening process is interrupted. This allows the X-ray screener to use the time productively that he or she would otherwise spend waiting. A disadvantage of this measure is that the X-ray screener can still be busy resolving the alarm using manual search and/or ETD while the X-ray screening process is ready to be continued. Both these measures were tested in the simulation of the present study for their potential to reduce the impact of EDS on throughput.

III. Method and Procedure

The simulation was implemented in FlexSim, an off the shelf 3D modelling and discrete event simulation software. Fig. 2 shows a screenshot of the model ASC lane. The basic layout, processes, and parameters of the model were set in accordance with a specific ASC design of a European airport, hereafter referred to as "reference ASC". To gain insight into the processes and parameters of the reference ASC, data from a previous work analysis were used. Further information was provided by ASOs working at the reference ASC and by experts in the field of aviation security.

In this section the model assumptions are described in their order within the baggage screening process. TABLE I. gives an overview of the assumed model parameter values and distributions. At the beginning of the baggage screening process is the arrival of the passenger at the ASC. The model is set up to provide a constant flow of passengers to simulate capacity, i.e. the throughput that can be achieved if there constantly are passengers ready to be screened. In a second step, passengers place their baggage and other belongings on the conveyor belt with the help of an ASO. According to several interviewed ASOs working at the reference ASC, this should take only about 5 s per item, as most passengers prepare their belongings while waiting for their turn to place their items on the conveyor belt. In accordance with [18] divesting time was modeled to be gamma distributed. For the baggage screening process it is not directly relevant how many items each passenger carries; an average of 3 items per passenger with a minimum of 1 was assumed.

TABLE I. Model Parameters

Parameter	Distribution	Mean and Standard Deviation in Brackets
Placing item on conveyor	Gamma	5 s (5 s)
Items per passenger	Poisson (translated)	3 (√2)
Evaluation time X-ray screener	Empirical	CR [a]/Miss: 3.90 s (1.32 s)
		FA[a]: 5.13 s (2.67 s)
		Hit: 4.05 s (1.67 s)
Duration of alarm resolution with manual search	Lognormal	116 s (132 s)
Duration of alarm resolution with ETD	Gamma (translated, shape = 1)	Condition 1: 30 s (5 s)
		Condition 2: 60 s (10 s)
		Condition 3: 120 s (20 s)

[a.] CR: correct rejection; FA: false alarm

After the items have been placed on the conveyor belt, the items are screened by the X-ray machine. Then, the X-ray image is analyzed by the EDS and reviewed by the X-ray screener. For this component of the process the false alarm rate and evaluation time of both the X-ray screener and EDS machine have to be defined for the simulation. To model the alarm rate and evaluation time of the X-ray screener, empirical data from three groups of [17] was used.[2] TABLE II. shows how these reference groups differ with regard to their false alarm rate, training hours conducted with X-ray Tutor Version 3[3], work experience, age, and the airport they work at. Separate simulations were conducted for the false alarm rates of these three reference groups to explore how differences

[1] The study was conducted using an X-ray screening simulator software and the EDS detection performance and settings were provided by authorities who are responsible for testing and certifying X-ray screening equipment with EDS functionality.

[2] [17] tested four groups: from two different airports and with two different levels of work experience (less than one year or more than two years). Due to the low number of new ASOs at the first airport, the group of ASOs with less than one year work experience from the first airport was only tested for the control condition and is therefore not considered in this study.

[3] Information on the X-ray Tutor computer-based training software can be found at www.casra.ch and for an earlier version of the software in [14]. Reference group 3 received initial training using another computer based training; the number of training hours per ASO could not be determined.

between X-ray screener groups affect the relationship between EDS and throughput.

TABLE II. REFERENCE GROUPS

	Group Averages and Standard Deviations in Brackets				
	Airport	***False alarm rate baseline condition***	***Training hours***	***Tenure / work experience***	***Age***
Reference group 1	Airport 1	.025 (.039)	101.40 (31.47)	7.68 (4.85)	42.50 (10.52)
Reference group 2	Airport 2	.040 (.046)	28.56 (12.41)	8.24 (5.78)	36.55 (8.46)
Reference group 3	Airport 2	.049 (.031)	2.58[3] (2.00)	< 1 year (-)	30.81 (10.93)

In addition to false alarms there are some quite common prohibited items (e.g. liquids, gels, or scissors) that require alarm resolution. For the simulation, the empirical value from a work analysis at the reference ASC of 1.23% bags and trays containing detected prohibited items was taken.

At the reference airport, the X-ray screener has a minimum of 3.5 s to review an X-ray image before the next X-ray image appears on the screen. This minimum is defined by the belt speed, the required distance between the screened items, and the average size of these items. If an X-ray screener needs more than 3.5 s to evaluate an X-ray image, he or she can stop the belt temporarily. To model these durations, empirical evaluation time distributions of [17] were used and capped at the minimum of 3.5 s. These evaluation times did not differ much between the three reference groups, but depended on the decisions of X-ray screeners ("OK" or "not OK") and in case of "not OK" on whether there actually was a prohibited item present (hit) or not (miss). These differences were taken into account by using three different empirical distributions.

There are EDS machines available which have the same belt speed as the current X-ray machines at the reference ASC, meaning that an EDS would not necessarily affect the rate at which items are X-rayed. Hence the EDS was assumed not to require additional evaluation time in the simulation model. As explained in the previous section, EDS machines can greatly vary in their false alarm rates depending on technology, machine type, and targeted hit rate. Therefore, false alarm rates ranging from 1 to 15% were explored.

After the item leaves the X-ray machine, it can be picked up by the passenger if both the X-ray screener and the EDS have cleared the item (i.e. not generated an alarm). In case the X-ray screener, the EDS, or both produce an alarm, the item has to be redirected for alarm resolution. This can either be done manually or automatically. In the automatic decision scenario modeled in this study, the alarms were redirected automatically to not disrupt the X-ray screening process. If an item has been declared as "not OK" by the X-ray screener, it is assumed that the alarm resolution follows the same procedures as they are currently applied at the reference ASC. The time needed to resolve alarms of the X-ray screener was measured at the reference ASC and could be approximated well with a lognormal distribution with a mean of 116 s and a standard deviation of 132 s.

In case the EDS generates an alarm, this alarm is assumed to be resolved with ETD with an average duration of either 30 s, 60 s or 120 s, as explained in the previous section.[4] It is assumed that the X-ray screener also reviews X-ray images which triggered the alarm by the EDS in order to detect prohibited items other than explosives. If that occurs, the X-ray screener sends the screened item to secondary search including a manual search (see also Fig. 1).

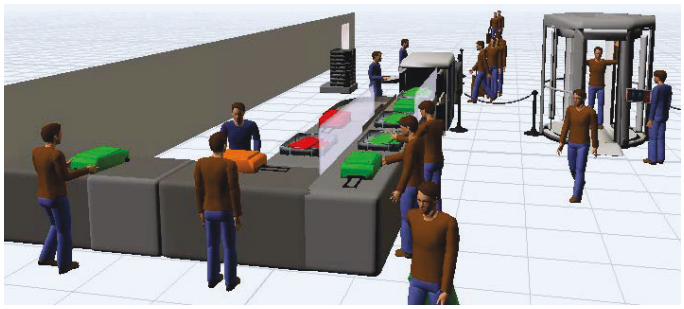

Fig. 2. Screenshot of the 3D model in FlexSim. Passengers have brown shirts, airport security officers have blue shirts. Passenger bags and other passenger belongings which have been judged as harmless are in green, bags and other passenger belongings that have been sent to secondary search are in red, and the bag currently being insepected using secondary search is in orange.

After the alarm has been resolved, the passenger can recollect his or her belongings.[5] How long the passenger needs for the recollection should not affect baggage throughput, as long as there is enough space available. Recollection was therefore modeled not to influence the baggage screening process in terms of throughput. At the reference ASC there is a limit of three items that can queue for alarm resolution and the X-ray screening process is interrupted if this limit is reached. This was modeled accordingly.

Separate simulations were run for each combination of the three reference groups, 15 false alarm rate levels of the EDS (1-15%), one level without EDS, and the three durations for alarm resolution using ETD (30 s, 60 s, 100 s). To test the effectiveness of the two measures described in the previous section, they were also run for each of the 15 false alarm rate levels of the EDS plus one level without EDS. To keep the results manageable, the measures were only tested in combination with the first reference group (which was

[4] It was assumed that alarm resolution by an ASO using ETD requires 25, 50 or 100 s respectively in most cases but that it can require longer in some cases. This was modeled with a gamma distribution with a shape factor of 1 and a mean of 5, 10 or 20 s respectively added to the mentioned minima.

[5] In rare occasions passenger screening can produce a disruption in baggage screening, e.g. if a security officer has to wait for the passenger before manual search of a bag can be conducted. As the focus of this study is on the baggage screening process, passenger screening will not be considered in the model, also because it is not directly affected by the introduction of an EDS. But the baggage screening process can be affected by the passenger screening process and can therefore not be expected to always achieve its full potential throughput in operation.

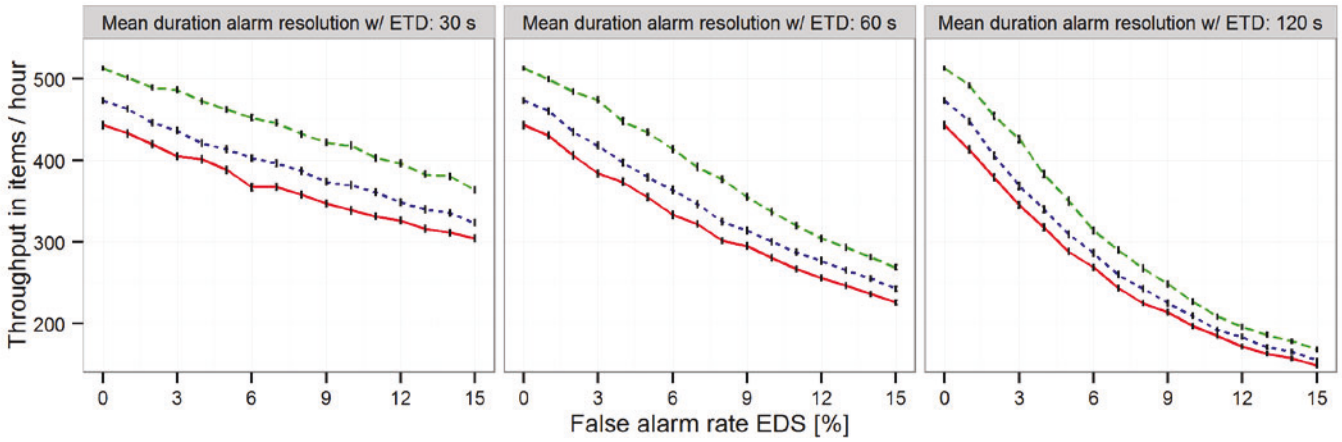

Fig. 3. Mean and standard error (over 200 simulation runs) of throughput in items per hour, depending on false alarm rate of EDS (zero representing the baseline without EDS) and on reference group, green dashed: reference group 1 (airport 1, tenure > 2 years), blue dotted: reference group 2 (airport 2, tenure > 2 years), red solid: reference group 3 (airport 2, tenure < 1 year), and mean duration of alarm resolution using ETD of 30, 60 and 120s.

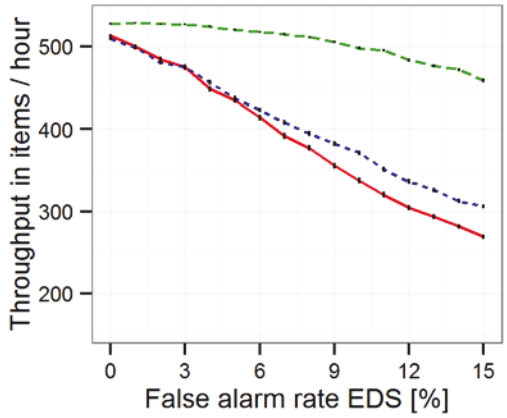

Fig. 4. Mean and standard error of throughput in items per hour, depending on false alarm rate of EDS, either with: red solid: single security officer resolving alarms, blue dotted: X-ray screener assisting with alarm resolution in case screening process is interrupted, green dashed: second security officer assigned to resolution of alarms.

recruited from ASOs working at the reference ASC and therefore seemed most adequate) and only for the medium duration of alarm resolution using ETD (60 s). For each of the resulting 192 conditions one hour of the baggage screening process was simulated 200 times (resulting in 38'400 simulated screening hours) and then aggregated.

IV. Results and Discussion

Fig. 3 shows the simulated throughput for a single ASC lane depending on reference group and false alarm rate of the EDS, whereby zero represents the absence of an EDS. As expected, capacity was negatively affected by the EDS's false alarm rate if no adaptions were made to cope with the increased workload due to additionally required alarm resolutions with ETD. This negative effect strongly depended on the average time required to resolve the alarms by the EDS using ETD. The results were also largely dependent on whether X-ray screeners were well trained and experienced.

As pointed out previously, a false alarm rate of 6% is feasible for certain currently available EDS equipment. For an average of 120 s for a secondary search with ETD, throughput was reduced by almost 40% compared to the baseline condition without EDS. If an alarm resolution with ETD only takes 60 s on average, then the reduction in throughput was much less, but also depended on the reference group of X-ray screeners: 19% for the first (i.e. the most experienced and best trained X-ray screeners), 23% for the second, and 25% for the third reference group (i.e. the least experienced and trained X-ray screeners). If the duration of alarm resolution with ETD was reduced to an average of 30 s, then throughput only decreased by 12% for the first, 15% for the second, and 17% for the third reference group. In this condition the well trained and experienced X-ray screeners of the first reference group still achieved a higher throughput than the third reference group with less than one year of job experience in the baseline condition without EDS (see Figure 2). Thus, the simulation results imply that a high throughput is still possible with EDS, if fast alarm resolution procedures using ETD can be implemented and if X-ray screeners are well trained.

Fig. 4 shows the relationship between capacity and the false alarm rate of the EDS for the standard security lane and the two measures that could be used to minimize negative effects on throughput as explained in the previous section. As could be expected, assigning a second ASO to the task of resolving alarms massively reduced the impact of the EDS's false alarm rate on throughput (assuming the tools and space for parallel alarm resolution are available). Within the simulation, instructing the X-ray screener to resolve one alarm while the screening process is interrupted only started having a positive effect on throughput at higher false alarm rate levels. In practice however, the X-ray screener might coordinate with the ASO responsible for alarm resolution and support him or her with tasks short enough not to delay the screening process too much. Therefore, having the X-ray screener assist with alarm resolution could be more useful in practice than it was found to be the case in the simulation.

V. Summary, Conclusions, and Limitations

The results of the discrete event simulation indicate that the baggage throughput of an ASC can strongly be affected by EDS. This effect is mainly due to the time needed for alarm resolution using ETD, which highlights the importance of fast ETD alarm resolution procedures (e.g. efficient trace sampling) and a short analysis time of the equipment. Not only the false alarm rate of the EDS machine and alarm resolution time of ETD but also the false alarm rates of the X-ray screeners were found to be very important. Training has been shown to reduce false alarm rates [6]. Potential decreases in baggage throughput due to the introduction of an EDS could therefore be at least partially compensated by having well trained X-ray screeners.

Having a second ASO to expedite alarm resolution could effectively reduce the negative impact of an EDS on throughput, while help by the X-ray screener with alarm resolution seems not to be a useful option based on the simulation results. A field study or a further work analysis combined with simulation could clarify if more coordinated assistance with alarm resolution by the X-ray screener (i.e. by only performing tasks that do not prolong the interruption of the X-ray screening process) has the potential to increase capacity.

Another limitation of the present model is that the false alarm probabilities of the items were assumed to be independent from each other. This does not necessarily need to be the case in practice. Certain passenger groups are likely to have increased or reduced alarm probabilities. Especially at small airports or decentralized ASCs, where passenger groups are less mixed, there might be periods requiring more alarm resolutions than other periods. This might mitigate average capacity due to the non-linear relationship between alarm rate and capacity.

Common cause failure of X-ray screeners and EDS machines could be investigated empirically and considered in future models. Also, more research is needed on the effect of EDS on X-ray screeners' performance under varying machine settings. The performance of the X-ray screeners working with EDS should also be examined over longer periods and in the field, as there is indication that the influence of automatic decision aids changes with experience [19] and training [20]. Another limitation concerns the 1.23% of detected prohibited items measured with a work analysis. The first reference group was a sample of the ASOs working at the reference ASC where the work analysis was conducted. It could however be expected that the other reference groups with less training and higher false alarm rates might also detect less prohibited items.

In sum, this first study on the effect of introducing EDS in cabin baggage screening provided important results, which could already have practical implications. However, more research including data collection from different ASCs as well as laboratory and field experiments are needed to validate and enhance these discrete event simulation results.

References

[1] T. Hunter, "Islamist fundamentalist and separatist attacks against civil aviation since 11th September 2001," in *Aviation security challenges and solutions*, F. Chau, Ed. Hong Kong: Avseco, 2011, pp. 35–54.

[2] V. Butler and R. W. Poole, "Rethinking Checked -Baggage Screening," *Reason Public Policy Inst., Policy Stud. No. 297*, Los Angeles, CA, 2002.

[3] K. Wells and D. a. Bradley, "A review of X-ray explosives detection techniques for checked baggage," *Appl. Radiat. Isot.*, vol. 70, no. 8, pp. 1729–1746, 2012.

[4] M. Mendes, A. Schwaninger, and S. Michel, "Can laptops be left inside passenger bags if motion imaging is used in X-ray security screening?," *Front. Hum. Neurosci.*, vol. 7, no. October, pp. 1–10, 2013.

[5] A. Schwaninger, D. Hardmeier, J. Riegelnig, and M. Martin, "Use It and Still Lose It?," *GeroPsych J. Gerontopsychology Geriatr. Psychiatry*, vol. 23, no. 3, pp. 169–175, 2010.

[6] S. M. Koller, C. G. Drury, and A. Schwaninger, "Change of search time and non-search time in X-ray baggage screening due to training.," *Ergonomics*, vol. 52, no. 6, pp. 644–56, Jun. 2009.

[7] A. Schwaninger, "Airport security human factors: From the weakest to the strongest link in airport security screening," in *Proc. 4th Int. Aviat. Secur. Technol. Symp.*, pp. 265–270, 2006.

[8] "Technical Data HI-SCAN 6040aTiX." [Online]. Available: http://www.siemens.ch/sbt/Sicherheit2011/Smiths_Detection_Bodyscanner/HI_SCAN_Brochure_Englisch.pdf. [Accessed: 25-Jun-2015].

[9] S. Singh and M. Singh, "Explosives detection systems (EDS) for aviation security," *Signal Processing*, vol. 83, no. 1, pp. 31–55, 2003.

[10] T. B. Sheridan and W. Verplank, "Human and Computer Control of Undersea Teleoperators," in *Technical Report, MIT Man-Machine Systems Laboratory*. Cambridge, MA, 1978.

[11] "IONSCAN 500DT." [Online]. Available: http://www.smithsdetection.com/index.php/products-solutions/explosives-narcotics-detection/61-explosives-narcotics-detection/ionscan-500dt.html#.VYvzAWP66sE. [Accessed: 25-Jun-2015].

[12] S. M. Koller, D. Hardmeier, S. Michel, and A. Schwaninger, "Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-Ray image interpretation," *J. Transp. Secur.*, vol. 1, no. 2, pp. 81–106, 2008.

[13] T. Halbherr, A. Schwaninger, G. R. Budgell, and A. Wales, "Airport Security Screener Competency: A Cross-Sectional and Longitudinal Analysis," *Int. J. Aviat. Psychol.*, vol. 23, no. 2, pp. 113–129, 2013.

[14] A. Schwaninger and F. Hofer, "Evaluation of CBT for increasing threat detection performance in X-ray screening," in *The Internet Society 2004, Advances in Learning, Commerce and Security*, K. Morgan and M. J. Spector, Eds. Wessex: WIT Press, 2004, pp. 147–156.

[15] A. Schwaninger, F. Hofer, and O. Wetter, "Adaptive computer-based training increases on the job performance of x-ray screeners," in *Proc. 41st Carnahan Conf. Secur. Technol.*, pp. 117–124, 2007.

[16] K. M. Ghylin, C. G. Drury, and A. Schwaninger, "Two-component model of security inspection: application and findings," in *16th World Congr. Ergon. IEA 2006*, 2006.

[17] M. Mendes, N. Hättenschwiler, Y. Sterchi, and A. Schwaninger, "Advanced cabin baggage (ACBS) study on human-machine performance and automation.," *Pap. Present. 61st Meet. ECAC Tech. Task Force, Paris, June 4-5*, 2015.

[18] A. Belyavin, "Simulating the impact of remote screening on search comb capacity," in *Proc. 47th IEEE Int. Carnahan Conf. Secur. Technol.*, pp. 1–6, 2014.

[19] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation bias: decision making and performance in high-tech cockpits.," *Int. J. Aviat. Psychol.*, vol. 8, no. 1, pp. 47–63, 1997.

[20] J. E. Bahner, A. D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Hum. Comput. Stud.*, vol. 66, no. 9, pp. 688–699, 2008.