UNIVERSITÄT BERN

Visual Search and Decision-making in Aviation Security X-Ray Screening

Commentary to a cumulatively published PhD thesis
submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy (Dr. Phil.) to the Department of Psychology,
University of Bern, Switzerland
by
Nicole Hättenschwiler
Bern, 2018

Advisors:
Prof. Dr. Fred W. Mast
Prof. Dr. Achim Elfering

Author:
Nicole Hättenschwiler
University of Applied Sciences and Arts, Northwestern Switzerland
Riggenbachstrasse 16
4600 Olten
nicole.haettenschwiler@fhnw.ch
0041 79 811 33 88

Acknowledgements

I am very thankful to the University of Bern and the School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland, for giving me the opportunity to pursue a PhD, and I am especially grateful to the many people I have met throughout my thesis.

Firstly, I would like to thank my supervisor, Prof. Dr. Adrian Schwaninger, for his guidance, advice and encouragement. I am deeply grateful to Prof. Dr. Fred Mast for agreeing to be my thesis advisor and opening the door to the University of Bern. I further like to thank Prof. Dr. Achim Elfering for agreeing to support me as my co-adviser.

Huge thanks go to my teammates Marcia Mendes and Yanik Sterchi who have been a part of my PhD-journey from day one and for their expertise, support and encouragement, as colleagues and friends. Further thanks goes to my team colleagues Sarah Merks and David Hügli who got my back whenever I needed them. I also like to thank the rest of my team colleagues who all had an important role during my PhD: Daniela Buser, Stefan Michel, Robin Riz à Porta, Thomas Wyssenbach, Melina Zeballos, Myrta Isenschmid, Kaspar Kaufmann, Marius Latscha and Vivienne Kunz. I also thank Milena Kuhn for her ideas and support when conducting the studies and all her inputs and corrections.

My biggest gratefulness and gratitude goes to my closest and dearest partners in crime, Kirsty Ng, Eliane Gilgen and Irina Baier who supported me with all of their strength. They are the kind of friends who help me becoming the best version of myself and I count myself extremely lucky to have them in my future. I also want to mention Halua Pinto, Barbara Kohler, Patrick Luder, Christina Wiedmer, Anja Wüthrich, Schirin Ibrahim and Joddy Appiah who were with me for an important part of this PhD-journey. There would be so many more to mention, who left their footsteps on my path.

Selina Ledergerber, my godchild, I know you will achieve so much in your life and I will guide, advise and support you with all what I've got to offer.

I dedicate this dissertation to my parents Regula and Pius, who believed in me and supported me with their love more than I could have ever asked for.

Abstract

In the present manuscript-based cumulative doctoral thesis, investigations were carried out on the topic of visual search and decision-making in aviation security X-ray screening. In particular, the question of how the overall system performance, including the interaction between human operators and technology, can be enhanced is being addressed. This thesis includes eight manuscripts which are divided into three different sections. The first section covers factors and challenges for visual search and decision-making in X-ray image inspection. Besides the discussion of known factors from literature, studies were conducted to evaluate the effect of knowledge about everyday objects and its relevance for visual inspection. Results of three experiments revealed interesting relationships between everyday object knowledge and performance in visual inspection. In the second section of the thesis, the impact of advanced technology for visual search and decision-making of improvised explosive devices (IEDs) in X-ray image inspection was investigated. Moreover, the application of newly developed technologies for cabin and hold baggage screening was evaluated in terms of their benefits on the overall human-machine system performance. Results showed promising benefits for detecting IEDs with the help of advanced technologies on the overall human-machine system performance in both, cabin and hold baggage screening. In the third section, studies were carried out to evaluate the validity of research about visual search and decision-making in X-ray image inspection. Results revealed insights on the generalizability and validity of different tasks, populations and detection measures in X-ray security screening. In order to obtain high external validity, all experiments conducted within the scope of this thesis used X-ray screeners working at European international airports as participants as well as realistic stimuli recorded and developed with aviation security experts. All results are discussed in the context of the corresponding literature and future research and implications are outlined. Conclusions are of high practical relevance and can thus deliver valuable contributions to the aviation security industry, including airports, appropriate authorities, regulators and manufacturers of X-ray technology.

Table of Contents

Introduction

Inspection research in the early days of the 20th century began with the assumption that the outcomes of said inspections were completely rigorous, elaborate and reliable. When studying the inspection process in depth in order to characterize and understand inspection errors, research in the 1950s disproved this assumption (Latorella & Drury, 1982, 1992; Wiener, 1986). In the 1970s, signal detection theory was applied to model inspection performance and various attempts were carried out to mathematically model the visual search phase of the inspection process (Drury & Fox, 1975; Harris & Chaney, 1969; Sheehan & Drury, 1971). In the 1980s, the focus was shifted to understanding individual differences in inspection performance and to derive techniques to select the best suited people for this job (Gallwey, 1982; Gallwey & Drury, 1986). Continued in the 80s, Drury and Sinclair (1983) were the first to carry out studies in the attempt to understand the use of automation for visual inspection. Following the Aloha Airlines incident in 1988, research regarding the reliability of aircraft maintenance and inspection drastically increased in the 1990s. The aforementioned scientific breakthroughs were accompanied by the introduction of computers to aid computer-based instruction and training for inspection tasks (Czaja & Drury, 1981; Drury & Gramopadhye, 1990; Drury & Kleiner, 1990; Shepherd & Parker, 1990; Thapa, Gramopadhye, Melloy, & Grimes, 1996). With 9/11, the application of findings from inspection research started to be applied in X-ray screening of passenger bags at airports (Drury, 2001). Inspection studies published after 9/11 investigated more thoroughly the use of automation using detection algorithms, new technologies including multi-view and 3D computer tomography (CT) imaging, the utility of virtual reality training techniques and many other factors (for reviews see Graves et al., 2011; McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Merry, 2015; Mounton & Breckon, 2015; Singh & Singh, 2003; Wells & Bradley, 2012). To summarize, research in this field is driven by the concerted effort to understand, minimize and cope with inspection errors.

X-Ray image inspection of passenger bags can be understood as a form of visual inspection (Drury, 1978; Koller, Drury, & Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, & Horne, 2009). Spitz and Drury (1978) assumed that the inspection task in general is composed of a search and a decision component, which are independent of each other. Thereby, visual search includes the visual scanning of an area and is terminated by either directing the attention to a suspicious part of an area (i.e. potential threat object) or by an upper bound on time spent on visual search. This search stage can be described as bottom-up, rapid, and global. Bottom-up processing relies on basic visual features that attract visual attention. Decision includes the fixation of the suspicious object, the matching of the visual stimulus with representations stored

in the visual memory, the decision itself (i.e. threat object vs. harmless) and the time to execute the response (OK vs. NOT OK). This decision stage of visual inspection involves both bottom-up (sensory) and top-down (cognitive) processes influenced by previous knowledge and expertise of the task. Thereby, signal detection theory (SDT) provides a framework to model this decision process. The basic idea of SDT is that when confronted with a binary decision task, cognitive information processing will ultimately result in some kind of one-dimensional subjective evidence variable for or against one of the alternatives (Wickens, 2002).

One major aspect in aviation security involves the mandatory process of baggage screening using X-ray machines. The implementation of X-ray machines to support baggage screening has been introduced over 40 years ago (Peil, 1972; Wetter, 2013). Instead of unpacking and hand searching every piece of baggage, X-ray images displaying the contents of a bag allow speeding up the inspection process. Based on the X-ray image, a screener who sits in front of a screen has to decide if a bag is OK and allowed to pass or whether it needs further inspection. As required by law (e.g. European Commission, 2015), screeners visually inspect every piece of *cabin* baggage at airport security checkpoints. Passengers' carry-on bags are stored in the cabin of airplanes during flight. Because this baggage can be accessed during a flight, guns, knives, improvised explosive devices (IED), and other items that could pose a threat (e.g., electric shock devices) are prohibited (Hancock & Hart, 2002; Harris, 2002; Schwaninger, 2005). Larger baggage is stored in the hold of an aircraft and processed differently (Shanks & Bradley, 2004). Passengers have to register such hold baggage at check-in stations before going through airport security checkpoints. At big airports, hold baggage is processed by a baggage handling system containing X-ray machines that have explosive detection systems (EDS) for detecting explosive material. Whereas in cabin baggage screening, there are multiple target types (guns, knives, IEDs, explosives, other threats), this is not the case in hold baggage screening. As passengers cannot access items stored in the hold of an aircraft, a gun or a knife does not pose a threat, and hold baggage screening targets only fully functioning IEDs (Bretz, 2002). Only X-ray images of hold baggage on which an EDS has raised an alarm are sent to remote screening locations for on-screen alarm resolution by screeners, thus visual inspection. If screeners decide that an X-ray image is suspicious, more time-consuming investigations follow, including rescreening with other X-ray technology, trace detection, explosive detection dogs and passenger investigation (Shanks & Bradley, 2004; Singh & Singh, 2003).

One conclusion has been drawn repeatedly and consistently since the initial investigations of visual search and decision-making: the final performance outcome relies on human inspectors who are not perfect. Or in other words and as Drury (1992) stated, inspection errors are a fact of life. They can be minimized with appropriate optimizations, but they cannot be

eliminated so far. Inspection errors can occur in the search and decision process of visual inspection. Wiener (1984) stated that inspection errors can either be omissions (missing a defect or a threat object) or commissive errors (so called false alarms—judging a harmless object as threat). In regard to airport security screening, missing a threat item can have severe consequences while producing false alarms has an influence on efficiency and costs (Dixon, Wickens, & McCarley, 2007). In other words, missing an item is mostly due to a search error that can occur due to factors such as target prevalence (frequency of occurrence), search strategy, visual-cognitive abilities or image-based factors (rotation/view of objects, superposition and bag complexity). While false-alarms are more often a result of errors in decision making that can be due to factors such as visual knowledge about objects (knowledge-based factors), task experience and training, environmental or situational factors and individual factors influencing the decision criterion.

Aim and Scope

A lot of research has already been conducted to investigate different factors and challenges influencing visual inspection performance, and therefore search and decision-making processes of X-ray screeners. Studying these factors will help to understand the occurrence of errors in search and decision-making of visual inspection and how to minimize and cope with them. Some factors already known from past research that have an influence on visual inspection are image- and knowledge-based factors (item view difficulty, superposition, complexity and visual knowledge: Schwaninger, 2003; 2005; 2006), individual factors such as visual-cognitive abilities, aptitudes, anxiety and motivation (e.g. Hardmeier, Hofer & Schwaninger, 2006; Mitroff, Biggs, & Cain, 2015; Schwaninger, 2006) as well as situational factors (Graves et al., 2011; Hättenschwiler, Michel, Kuhn, Ritzmann, & Schwaninger, 2015; Kraemer, Carayon, & Sanquist, 2009). Challenges further include target prevalence, variations in target visibility, the possible presence of multiple targets or the translation of results from lab-based research to the field and vice versa (Biggs & Mittroff, 2014; Clark, Cain, Adamo, & Mitroff, 2012; Godwin et al., 2010; Godwin, Menneer, Cave, Thaibsyah, & Donnelly, 2015; Mitroff, Biggs, & Cain, 2015). As new and unknown threats are emerging on a regular basis, it is important that screeners are well trained and constantly updated on what to actually search for in X-ray images. While advancements in technology can certainly enhance screening processes, recognizing objects in X-ray images remains a demanding and challenging task, which currently cannot be solved by machines solely.

This thesis aims to answer the question of what factors have an influence on search and decision-making (errors) in X-ray image inspection and therefore the overall system

performance, including both machine and human performance. This research question will be answered from three different angles involving (1) challenges and factors for visual search and decision-making, (2) the impact of advanced technology for visual search and decision-making and (3) the validity of results in visual search and decision-making in X-ray image inspection. Results from the conducted experiments shall give indications on how to cope with and minimize errors in search and decision-making.

(1) Challenges and Factors influencing Visual Search and Decision-making

Firstly, the detection of prohibited items, and therefore the search and decision process, can be affected by different image-based factors. Studies have identified three image-based factors which influence the probability of identifying a threat object in an X-ray image: viewpoint (rotation of threat items in a bag), superposition of a threat item by other items and bag complexity (Hardmeier et al., 2005; Hardmeier, Hofer & Schwaninger, 2005, 2006; Hofer, Hardmeier & Schwaninger, 2006; Schwaninger, 2003). Items which are presented from unusual or rotated viewpoints become more difficult for a person to identify (effect of viewpoint) (Palmer, Rosch & Chase, 1981). Similarly, the position of a prohibited item in a bag and its superimposition by other objects (effect of superposition), or the number and types of items in a bag which could attract attention (effect of bag complexity) also affect the difficulty of recognizing prohibited items. Bag complexity comprises the factors clutter (disarrangement, textural noise, chaos, etc.) and opacity (X-ray penetration of objects) (Schwaninger et al., 2008).

Studies have shown that coping with image-based factors is greatly dependent on a person's visual cognitive abilities and can thus only to a limited extent be improved through training (Hardmeier et al., 2005; Hardmeier et al., 2006b). Especially visual-cognitive abilities such as figure-ground segregation (related to superposition) or mental rotation (related to view difficulty) can play an important role for visual inspection performance. Such abilities are rather stable within a person (Hardmeier, Hofer & Schwaninger, 2006b) but vary substantially between people. Consequently, only people suited to conduct the screening task regarding the required visual-cognitive abilities should be recruited for this job (Hardmeier, Hofer & Schwaninger, 2005, 2006a; Harris, 2002). Therefore, investigating different cognitive abilities separately, such as tests on mental rotation, figure-ground segregation and visual search on complex backgrounds has been shown to be useful (see Bolfing & Schwaninger, 2009; Hardmeier & Schwaninger, 2008). In particular, logical thinking, spatial imaging and the ability to recognize a shape by ignoring irrelevant other features (figure-ground segregation) proved to be important predictors (Hardmeier & Schwaninger, 2008).

Secondly, previous studies in the field of X-ray security screening showed that the detection of forbidden objects in X-ray images of passenger bags depends on extensive visual knowledge, or so called knowledge-based factors (Schwaninger, Hardmeier & Hofer, 2005). Prohibited items are difficult to recognize without training because a) objects often look very different in X-ray images than in reality, b) certain prohibited items are not known from everyday experience (e.g. IEDs), c) some prohibited items look similar to harmless objects (e.g. a switchblade knife can resemble a pen), and d) when objects are depicted from unusual viewpoints, they become difficult to recognize (Schwaninger, 2003; 2005; 2006). During initial classroom, computer-based and on the job training, X-ray screeners learn how to interpret X-ray images in order to recognize everyday objects and prohibited items. Different studies could demonstrate, that the required knowledge-based factors for X-ray screening can be very well trained with appropriate computer-based training (e.g. Hardmeier et a., 2006b; Koller et al,. 2008). While these studies have provided converging on the importance to learn which items are prohibited and what they look like in X-ray images, the role of everyday object knowledge and the use of a specific search strategy to reduce search and decision errors for visual inspection has not yet been addressed thoroughly. These are interesting topics especially from an operational point of view but also regarding the emergency of new threats. Manuscripts ***1*** and ***2*** of this thesis therefore aim to answer whether the knowledge about everyday object and a specific search strategy have an influence on the visual inspection performance.

(2) The Impact of Advanced Technology on Visual Inspection Performance

The above mentioned factors focus on ways to mitigate search and decision errors and improve visual inspection performance of X-ray screeners. However, it is also vital to consider the technology applied and its interaction with the screener to continuously assess the effectiveness and efficiency of the combined system. It is for this reason that the technical and human elements of airport security systems must work together effectively, as weaknesses in either element can diminish the effectiveness of the overall system (Graves et al., 2011). To reduce challenges and errors of visual inspection of X-ray images, manufacturers of X-ray machines, regulatory bodies, airport research centers as well as scientists are investing into improvements of X-ray screening technology and therefore the overall human-system performance. From a technical perspective, these investments include enhancements of image quality, the provision of multiple or 3D rotatable views, and automated detection algorithms. Technological advancements bring along possibilities to increase effectiveness and efficiency of security X-ray screening by enhancing threat detection and supporting screeners in their decision. Nevertheless, it is crucial to thoroughly assess their effects on the overall system

performance and how well screeners can actually benefit from these new features. Despite the continuous advancements, the final *decision* on whether a bag is allowed to pass or not still relies on the human operators and depends on the image outcome provided by the machines (Graves et al., 2011; Koller et al., 2008).

Looking at the history of attacks against airplanes (both successful and near misses), one of the biggest concerns are bombs – that is, improvised explosive devices (IEDs; Baum, 2016; Novakoff, 1992; Singh and Singh, 2003). The Global Terrorism Database (2017) lists 893 attacks on airports or aircrafts with explosives, 247 of which occurred after 2001. The most recent attack involving an IED in hold baggage occurred on October 31, 2015 when Metrojet Flight 9268 was blown up during flight killing all 224 passengers (Baum, 2016). Just recently on July 29 2017, a terrorist plot was prevented at Sydney airport in which an IED was concealed inside a cabin bag (Westbrook and Barrett, 2017). Whereas even novices can recognize certain objects shapes such as guns and knives in X-ray images (Schwaninger et al., 2005), especially IEDs are difficult to recognize without respective training (Schwaninger and Hofer, 2004; Koller et al., 2008; Koller et al., 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 (Schwaninger and Hofer, 2004; Koller et al., 2008; Koller et al., 2009; Halbherr et al., 2013; Schuster et al., 2013). In cabin baggage screening however, bare explosives also pose a threat, because these could be combined with other IED components after passing an airport security checkpoint. Therefore, detecting bare explosives can be a challenge even for well-trained screeners, because they often look like a harmless organic mass (Jones, 2003).

In response to bomb threats, explosive detection systems (EDS) based on 2D imaging for hold baggage screening (HBS) were developed and introduced about 15 years ago (Caygill, Davis, & Higson, 2012; Harding, 2004; Singh & Singh, 2003). In the last few years, these explosive detection systems have become available for cabin baggage screening (EDSCB) as well (Sterchi and Schwaninger, 2015). These systems change the search and decision process of visual inspection. Manuscript ***3*** to ***6*** therefore aim to investigate the impact of these new technologies on the search and decision process and the overall system performance. Manuscript ***3*** focuses on the impact of automation using explosive detection systems for cabin baggage screening, while manuscripts ***4, 5*** and ***6*** investigate an explosive detection system using 3D computer tomography technology for hold baggage screening.

(3) Validity of Results in Visual Search and Decision-making in X-Ray Screening

Ideally, the knowledge gained from traditional research of visual search and decision-making could be directly applied to professionals and therefore provide high external validity. But several obstacles stand in the way of a straightforward translation; the tightly controlled visual searches performed in the traditional lab setting can be drastically different from applied X-ray image inspection (Clark, Cain, Adamo, & Mitroff, 2012). There are some major hurdles in directly translating results from traditional visual search to X-ray image inspection and vice versa: In comparison to traditional search tasks like feature or conjunction search (Eckstein, 2011), applied settings like X-ray image inspection differ in terms of the nature of the stimuli, the task itself, the environment in which the search is taking place, and the experience and characteristics of the searchers themselves (Clark et al., 2012). Research on traditional visual search certainly provides important theoretical insights for X-ray image inspection. Nevertheless, external validity should be questioned, as there are some limitations when interpreting traditional visual search results in terms of their implications for the X-ray inspection task at security checkpoints. Challenges of visual inspection in X-ray baggage screening compared to visual search in traditional search tasks further include low target prevalence, variation in target visibility, an unknown target set and the possible presence of multiple targets (for recent reviews see Biggs & Mitroff, 2014; Mitroff, Biggs, & Cain, 2015). To decide whether a bag contains a prohibited item or not, screeners need to know which items are prohibited and what they look like in X-ray images (Schwaninger, 2005, 2006). In the case of X-ray image inspection, a screener's task is to stay alert even though targets are hardly ever present. Further, screeners must search for different targets at a time from a diversity of categories. Traditional visual search tasks, on the other hand, normally include searching for only one target at a time, which often is known beforehand and occurs with high prevalence (e.g. 50%). Other important stimuli differences are target and distractor complexity. Targets in X-ray image inspection are not well-specified, not salient and not predictable by the context (Bravo & Farid, 2004), and distractors may additionally produce clutter and superposition (see image-based factors). Furthermore, it is often students performing traditional visual search tasks, lacking the experience of a professional even when completing test trials beforehand. Professionals have received directed visual search training and on-job experience with feedback. It might therefore be possible that professionals' experiences, above and beyond actual training, can significantly impact visual inspection behaviours and outcomes. Out of this, manuscript ***7*** of this thesis investigated underlying processes and the validity of visual search and inspection research in regard to whether results gained from a traditional visual search task can be directly translated to an X-ray visual inspection tasks and vice versa while testing students and professionals.

Next to the above-mentioned challenges, another important consideration is whether inspection measures used to map detection performance of standardized visual search tasks in the lab can be mapped to measure the task of visual inspection in X-ray screening. The outcome of the visual inspection task can be determined by screeners' decisions on whether an X-ray image of a passenger bag is harmless (target absent) or not because it might contain a prohibited item (target present). According to SDT (Green & Swets, 1966; Macmillan & Creelman, 2005) the proportion of bags that contain a prohibited item and are correctly classified as such is called hit rate (HR), while the percentage of harmless bags that are falsely classified to contain a prohibited item is the false alarm rate (FAR). There is a trade-off between the HR and FAR: if a screener heightens the tendency to respond with target present, both the HR and FAR will increase. It is therefore possible that one could always decide to respond with target present resulting in a HR and FAR of 100%. Individuals with the same ability to detect prohibited items can however show a different HR and FAR due to differences in their response tendency. Signal detection theory provides measures (such as d' and A') for detection performance in terms of sensitivity that are based on 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 visual inspection of passenger bags lead to an increased 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). D' and A' are also frequently used in related domains like the visual 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) or other 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, recent studies cast doubt on the validity of d' or A' for visual inspection tasks in X-ray screening due to differences in the underlying evidence distribution compared to simplified visual search tasks with artificial stimuli. Therefore, manuscript ***8*** of this thesis investigated the validity of commonly used visual search and inspection performance measures.

Outlook

Taken together, X-ray image inspection and the involved interaction of humans and technology is a research field of high practical relevance. The research studies conducted within the scope of this thesis thus deliver valuable contributions for the aviation security industry,

including airports, appropriate authorities, regulators and manufacturers of X-ray technology. After presenting the methods used to conduct the experiments, results of the conducted experiments will be discussed in relation to their practical relevance and implications.

Manuscripts in this Thesis

Factors and Challenges Influencing Visual Search and Decision-making in X-Ray Screening

1) Hättenschwiler, 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*, Taipei Taiwan, September 21-24, 2015, 25-30. doi: 10.1109/CCST.2015.7389652

2) 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*, Madrid Spain, October, 2017. doi: 10.1109/CCST.2017.8167812

Impact of Advanced Technology on Visual Search and Decision-making in X-Ray Screening

3) 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.

4) Hättenschwiler, N. Mendes, M., & Schwaninger, A. (2018). Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging. *Human Factors*, doi:10.1177/0018720818799215

5) Hättenschwiler, N., Merks, S., & Schwaninger, A. (2018). Airport security screening of hold baggage: 2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol. *Proceedings of the 52th IEEE International Carnahan Conference on Security Technology*, Montreal Canada, October, 2018.

6) Merks, S., Hättenschwiler, N., Zeballos, M. & Schwaninger, A. (2018). X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging?. Proceedings of the 52th IEEE International Carnahan Conference on Security Technology, Montreal Canada, October, 2018

Validity of Results in Visual Search and Decision-making in X-Ray Screening

7) Hättenschwiler, N., Merks, S., Sterchi, Y., & Schwaninger, A. (n.d.). Traditional visual search vs. X-ray image inspection in students and professionals: Are the same visual cognitive abilities needed? Under Review in *Frontiers in Psychology: Cognition*.

8) Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (n.d.). Detection Measures for Visual Inspection of X-ray Images of Passenger Baggage. Under Review in *Attention, Perception, & Psychophysics*.

Methods

In the following, a short overview on the methods and measures relevant for understanding the research conducted within this thesis is provided. In order to obtain high external validity, all experiments conducted within the scope of this thesis used X-ray screeners working at European international airports as participants as well as realistic stimuli recorded and developed with aviation security experts.

Participants

Participants for all studies conducted within this thesis were professional X-ray screeners from different European international airports 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). For each conducted experiment, the 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.

Experimental Tools and Procedure

For each experiment, X-ray screeners were tested using realistic stimuli that was recorded or created with the help of aviation security experts. Aligned to experimental research of visual inspection, images were split between target-present images (containing a target of interest) and target-absent images (no target item present). Target-present images always contained one threat object from the categories guns, knives, improvised explosive devices (IEDs) or other threat objects except for manuscript 3 to 6 where target-present images only contained single IEDs. Target-absent images did not contain any threat object (clear bag) and sometimes represented a false alarm (e.g., cheese, chocolate, or certain liquids) for studies with simulated automation aids. A target prevalence of 50% was used except for manuscript 3 where the target prevalence was set at 12.5%.

For the testing situations, four to six participants performed the tests in each session while working individually, quietly, and under supervision. Participants received instructions before the start of each test informing them about the imaging systems, the number of images, and the target items used. To prevent a criterion shift (change of response bias) during the experiment, screeners were informed beforehand about the target prevalence in the experiment (see also McCarley, 2009; Rich et al., 2008). All tests were conducted without giving performance feedback.

For every experiment, participants task was to visually inspect each X-ray image as if they were working at the airport and to decide as accurately and quickly as possible whether or not the image contained a target by clicking on a target-present or a target-absent button on the simulator interface (a yes–no task in signal detection theory; see MacMillan & Creelman, 2005). After receiving their instructions, all screeners started the experiment with practice trials (target-absent and target-present images in random order). As the European regulations mandates that screeners have to take a break of at least 10 min after 20 min of continuous visual inspection of X-ray images (European Commission, 2015), all tests were divided into blocks and screeners were asked to take breaks of 10 to 15 min after completing each block.

Measures

The outcome of the task of visually inspecting X-ray images of passenger bags is determined by X-ray screeners *decisions* on whether a bag is harmless (target absent) or might contain a prohibited item (target present). *Table 1* presents the four possible combinations and the associated terminology from visual search (e.g. Eckstein, 2011), signal detection theory (e.g. Green & Swets, 1966) and X-ray security screening (e.g. Drury & Fox, 1975).

Table 1

Outcome of Response/Decision Depending on Stimulus Using the Terminology of Visual Search, Signal Detection Theory, and X-ray Baggage Screening

	Response/decision	
	Target absent	Target present
	No signal	Signal
Stimulus	Bag is harmless	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). 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). The other terms have been used in X-ray security screening studies (Cooke & Winner, 2007; Hättenschwiler et al., 2015).

In detection theory (Macmillan & Creelman, 2005), the probability of bags containing 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). For all the analyses of the manuscript presented within this thesis, detection

performance measures were calculated using SDT formulas (Green & Swets, 1966; MacMillan & Creelman, 2005).

Factors and Challenges for Visual Search and Decision-making in X-Ray Screening

As mentioned before, several studies have provided converging evidence on the importance of image-based and knowledge-based factors and therefore the need to learn which items are prohibited and what they look like in X-ray images. However, the role of everyday object knowledge has not yet been addressed specifically. This is an interesting topic especially from an operational point of view. In particular, one could assume that the knowledge on what everyday objects look like in X-ray images could result in fewer cases where an everyday object is confused with a prohibited item (e.g. pen can resemble a switchblade knife). This would result in less decision errors and therefore fewer false alarms, i.e. wrongly deciding that a bag contains a prohibited item. False alarms have to be resolved by secondary search which typically involves manual search and/or alarm resolution using explosive trace detection technology (Koller & Schwaninger, 2006). Due to the additional time needed for secondary search, high false alarm rates can have a strong negative impact on throughput (Koller & Schwaninger, 2006) and could also result in lower passenger satisfaction (Hardmeier, Hofer & Schwaninger, 2006).

Two experiments within this thesis investigated the role of everyday object knowledge and training and its relevance for effective and efficient X-ray image inspection. For all experiments, screeners conducted a newly created test to measure how well novices and X-ray screeners can categorize and name everyday objects in X-ray images as well as a simulated X-ray baggage screening task. In the first experiment, we examined whether there is a statistically significant and meaningful relationship between everyday object knowledge and false alarm rate in a simulated X-ray baggage screening task (published in Hättenschwiler et al., 2015). In a second experiment, we tested whether e-learnings can be used to learn everyday object knowledge effectively and efficiently (published in Hättenschwiler et al., 2015).

We found a negative relationship between the percentage of correct answers for the naming of harmless everyday objects in an object categorization and naming task and the false alarm rate in a simulated X-ray baggage screening task. We could further show promising results on e-learning as an effective and efficient tool for building knowledge of harmless everyday objects in X-ray images. 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 (2010). This model is similar to the two-component model by Spitz & Drury (1978), 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.

Based on this results, 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. In a third experiment, we therefore investigated whether the effects of instructing a new inspection strategy using the knowledge about everyday objects can effect effectiveness of X-ray screening (published in Sterchi, Hättenschwiler, Michel, & Schwaninger, 2017). Results showed that an instruction to decide more liberal, with the help of knowledge about everyday objects, led to increased hit and false alarm rates while sensitivity remained constant what implies that the observed change in hit and false alarm rates was due to a change in the decision criterion. 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.

Implications

Results from my studies and earlier literature show that the task of visual inspection in X-ray screening cannot be compared to a simple visual search task. Not only threat objects but also distractors have to be recognized to reach a high inspection performance. To this date, the mostly used learning strategy for threat detection is simple computer-based training. There is still a lot of potential to improve and adapt learning environments, methods and strategies for X-ray image inspection. Innovations concerning training systems or training modules could therefore support screeners in their challenging task of visual inspection but also enhance their motivation. Training systems could for example also include interactive modules concerning topics such as everyday objects, known components of IEDs based on realistic scenarios, behaviour patterns of passengers, coping strategies for stress situations and so on. Additionally, multi-modal learning like computer simulation or virtual reality software could provide a wider spread interface by including the implementation of situational factors (surrounding distractors and time pressure) which can also affect performance (McCarley, 2004; Michel et al., 2014). A

future goal should be to enable operators to train with more realistic scenarios including mobile apps and e-learning, job aids and simulations. Trainees need multiple touches and ways to consume information as these transform training from an event into an extended learning experience.

X-ray screeners not only have very high job demands (high responsibility, high workload, time pressure) and therefore a lot of pressure as missing a threat can have severe consequences. However, they also have limited resources in regard to external motivation factors (conduct a task of high monotony, low autonomy and flexibility, shift work) which are important predictors of job satisfaction. Future studies on X-ray image inspection could also look at the wider spectrum of the whole screening process at security checkpoints, including aspects such as work environments, motivational factors, team compositions and organizational or cultural differences. Interventions could be set to reinforce motivational factors to enhance job satisfaction for example by reducing the workload of screeners, enlarging the job responsibilities or promoting a supportive working environment.

Impact of new Technology for Visual Search and Decision-making in X-Ray Screening

The implementation of new technology is another important factor to reduce the occurrence of errors in search and decision making of visual inspection. Innovations will be discussed separately for cabin baggage screening (CBS) and hold baggage screening (HBS).

Cabin baggage screening

Explosive detection systems for cabin baggage (EDSCB) use automation that influences both aspects of visual inspection in X-ray screening: search and decision. Automation refers to functions performed by machines (usually computers) that assist or replace tasks performed by humans (for recent reviews, see Parasuraman and Wickens, 2008; Sheridan, 2011; Vagia et al., 2016). One form of automation is assisting a human operator with a diagnostic aid (Wickens and Dixon, 2007) that provides support in the form of alerts or alarms and influences attention allocation (Cullen et al., 2013). In X-ray security screening, these systems indicate potentially threatening objects in X-ray images of passenger baggage and aim reduce search errors in visual inspection.

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) provides a useful framework 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 et al., 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 it will result in many false alarms and only few hits (Parasuraman et al., 1997). This can produce a 'cry wolf' effect what leads operators to ignore 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 (higher) levels of automation are possible (for a review, see Vagia et al. 2016).

To investigated the possible benefits of EDSCB, a study within this thesis (published in Hättenschwiler, Sterchi, Mendes, & Schwaninger, 2018) compared two different levels of automation currently being discussed by regulators and airport operators: automation as a diagnostic aid with on-screen alarm resolution (OSAR) by screeners or EDSCB with an automated decision by the machine. We conducted two experiments to test and compare 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 IEDs even without EDSCB, while EDSCB with OSAR induced a cry-wolf effect. EDSCB increased only their detection of bare explosives. In contrast, screeners with less experience (tenure < 1 year) benefitted substantially from EDSCB in detecting both IEDs and bare explosives. A comparison of all three conditions showed that automated decision provided better overall human–system 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. Based on the results it can be indicated that an implementation of EDSCB increases the detection of explosives in passenger bags and automated decision instead of automation as diagnostic aid with OSAR should be considered.

Hold baggage screening

Explosive detection systems (EDS) in hold baggage screening assist screeners to visually inspect X-ray images of passenger bags before they are loaded into the hold of an aircraft (Wells & Bradley, 2012). They indicate areas in X-ray images that might be explosive by colored

frames or a specific surface color (Wells & Bradley, 2012). X-ray images on which an EDS has raised an alarm are sent to remote screening locations for on-screen alarm resolution (OSAR) by screeners. Compared to cabin baggage screening without automation, the task of hold baggage screeners is therefore mainly a *decision* tasks and involves the visual inspection of the alarmed areas in X-ray images and decide whether such EDS alarms are harmless (false alarms of the EDS) or whether the hold baggage needs further inspection because it might contain an IED.

Several years ago, advanced CT technology, which has been implemented highly successfully in medical imaging (Barrat, 2000), became available for hold baggage screening (Mouton & Breckon, 2015; Wetter, 2013). Compared to the 2D imaging X-ray systems used in HBS, state-of-the art CT scanners feature better automated explosive detection, slicing, and 3D-rotatable images (Mouton & Breckon, 2015; Wells & Bradley, 2012). This might result in better detection performance of screeners as it might be easier to recognize the different components of an IED that, in certain 2D views, would be displayed from a difficult viewpoint and/or superimposed by other items in a complex bag (Bolfing et al. 2008; Schwaninger et al., 2005). Further, exposure to 3D objects results in richer visual object representations (Tarr & Vuong, 2002; Vuong & Tarr, 2004) that could improve screeners' detection performance not only in 3D but also in 2D images. On the other hand, CT systems have lower image resolution compared to EDS-HBS with 2D imaging (Flitton, Breckon, & Megherbi, 2010: Flitton et al., 2013; Mouton & Breckon, 2015), and this could impair detection performance with 3D imaging.

To accompany this technological change, a study conducted within the scope of this thesis (published in Hättenschwiler, Mendes & Schwaninger, 2018) compared the visual inspection performance of screeners when screening hold baggage with newer 3D versus older 2D multi-view imaging. Screeners from two European international airports, some used to working with 2D imaging (2D screeners) and some used to working with 3D imaging (3D screeners) conducted a simulated hold baggage screening task with both types of imaging. The results revealed that despite lower image quality (we assessed the image quality of the imaging systems with the standard procedure for 2D imaging), detection performance of both screener groups with 3D imaging was similar to that with 2D imaging. 3D screeners revealed higher detection performance with both types of imaging than 2D screeners which leads to the conclusion that features of 3D imaging systems (3D image rotation and slicing) seem to compensate for lower image quality. Further, visual inspection competency acquired with one type of imaging seems to transfer to visual inspection with the other type of imaging. These results imply that replacing older 2D with newer 3D imaging systems can be recommended. Also, 2D screeners do not need extensive and specific training to achieve comparable detection performance with 3D imaging.

The above described results could be replicated by a further study within the scope of this thesis (published in Hättenschwiler, Merks & Schwaninger, 2018). Moreover, an on-screen alarm resolution protocol (OSARP) that was created to facilitate the *decision* making process of screeners using 3D imaging was assessed in regard to its effectiveness. Our study showed that screener's visual inspection performance could be increased when they followed this specific OSARP. Due to the OSARP training, screeners also shifted their response bias to be more neutral (more biased toward judgement of target present). This implies that screeners were very compliant with the protocol resulting in more hits but also more false alarm decisions in the OSARP condition. The same study also investigated visual-cognitive abilities and correlations to visual inspection performance with 2D and 3D imaging (published in Merks, Hättenschwiler, Zeballos & Schwaninger, 2018). The results suggest that screening with 3D imaging systems might also require different visual-cognitive abilities. The possibility of rotating the X-ray image of a bag and its content around 360 degrees seems to facilitate the recognition of prohibited items when depicted from unusual viewpoints, when superimposed by other items and when placed in visually complex bags. This might explain why fewer visual cognitive abilities become relevant for 3D screening compared to 2D screening.

Implications

Prior to actually applying new technologies in the airport setting, it is advisable to properly evaluate how these actually benefit human performance and the overall system. It is not only important to carefully analyze possible advantages but also disadvantages when implementing new technologies and automation features. New devices will not only demand different skills from screeners to conduct their job successfully, they should also be assessed in terms of possible adjustments in object recognition, knowledge- and image-based factors and visual-cognitive abilities, as well as selection and training procedures for screeners. For example, a discrepancy between trust in an automation aid and system reliability can influence how automation is used. Automation aids with high false alarm rates may induce a cry wolf effect and therefore reduced compliance with alerts of the automation (Bliss, Gilson, & Deaton, 1995; Dixon, Wickens, & McCarley, 2007; Parasuraman et al., 2000; Wickens & Dixon, 2007). This low reliability of the automation system may result in disuse or misuse of the automation aid (Meyer, Wiczorek, & Günzler, 2014; Parasuraman, Molloy, & Singh, 1993).

EDS in CBS and HBS is a good example of the trend to use higher levels of automation in diverse work environments, also outside of airport security screening. This brings up several questions of how future developments might look like: Will the human operator still be involved in the workflow process? How will the function allocation be divided between human and

machine? Will work processes still be holistic or is there a trend towards more specialization? Will the scope of action be reduced for humans?

Even though technological advancement might increase effectiveness and efficiency, it should still be considered what impact these advancements have on the human operator. Future research should investigate good concepts that promote an ideal function allocation while also considering the well-being of the human operator to conduct a holistic work process. It should be investigated whether Taylorism might be a misconception of an increase in productivity as it might come at the cost of job satisfaction and all its consequences on the human operator. All this mentioned points demonstrates the need for a joint evaluation of the performance of the human operator and the effect of advanced technology to investigate whether the system performance as a whole can be actually improved and search and decision errors reduced.

Validity of Results in Visual Search and Decision-making in X-Ray Screening

Over the past decades, psychological research has made tremendous effort in understanding the processes responsible for performing traditional visual search tasks and the mechanisms that allow for the successful identification of target items (Clark et al. 2015). However, it is important to note that the traditional visual searches performed with student participants can lead to significantly different results when compared to X-ray image inspection by professionals. In X-ray image inspection, missing a threat can have severe consequences, while falsely rejecting too many bags will result in inefficiency and long waiting lines. Also, the large variety of potential threat items and distracting objects in passenger bags make X-ray image inspection a difficult task. Compared to traditional visual search, in this real-world scenario the accuracy of professional searchers can have life-or-death implications. Whereas traditional visual searches can usually be titrated down to the search for a specific target in a specific scenario, professional searches are often more noisy and ambiguous. This difference highlights a potential hurdle between traditional cognitive psychology research and how searches are actually conducted in real life. A fundamental research question derived from these challenges and hurdles is to know how traditional visual search might translate to X-ray image inspection and vice versa and if the same visual-cognitive abilities can predict performance of these tasks.

Comparing a traditional visual search task with an X-ray image inspection task is important to gain a better understanding regarding the external validity of the large number of studies that used traditional visual search paradigms with rather simple and artificial stimuli. Another study within the scope of this thesis (Hättenschwiler, Merks, Sterchi, & Schwaninger, submitted) therefore investigated this research questions by comparing two populations (students vs.

professionals) using the same experimental stimuli, including both a traditional visual search task and an X-ray image inspection task as well as an assessment of visual-cognitive abilities of these populations. Results showed that the same visual-cognitive abilities predict performance for both tested tasks and populations. A group difference was only found for the X-ray image inspection task where professionals performed faster and more accurately but not for the traditional visual search task. Based on the tasks and populations tested in this study, results can be seen as transferrable to a certain point. The understanding that a traditional visual search tasks without domain-specific knowledge can be generalize from students to professionals can be a huge advantage, allowing for certain applied studies to be run with relatively easier accessible populations. Further, comparing visual-cognitive abilities and their influence on performance showed that the same specific visual-cognitive abilities were able to predict performance and response times in both tasks. As the same cognitive processes are underlying the tested tasks, future searchers can be differentiated based on some very specific abilities. However results should be interpreted with caution if domain- and knowledge-specific tasks with unknown features of stimuli are performed, as students differ from professional populations due to differences in knowledge and experience.

To model performance for the traditional visual search task vs. X-ray image inspection task in the submitted manuscript (Hättenschwiler, Merks, Sterchi, & Schwaninger, n.d.), we used a different performance measure for each task implying different evidence distributions. An 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 fully defines the detection ability (sensitivity), while the decision criterion captures the response tendency. However, recent studies indicate that d' (and also A') is not always a valid measure for visual inspection tasks in X-ray screening. 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 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 signal-plus-noise (target-present) and noise (target-absent) distribution, the resulting zROC (z-transformed receiver operating characteristic) has slope s. This assumption makes sense as the prohibited items that have to be detected vary and therefore bring additional variation into the X-ray image. Therefore, Wolfe (2010) 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.

Two experiments within the scope of this thesis (Sterchi, Hättenschwiler & Schwaninger, submitted) investigated the validity of different measures of detection performance (d', A', da). In the first experiment, 31 screeners completed a simulated X-ray visual inspection task, while response tendency was manipulated directly by an instruction. Results showed that for two different levels of task difficulty, d' and A' decreased when the criterion became more liberal and were therefore not valid measures of detection performance for the investigated task. In contrast, another measure implying unequal variance distribution (da) with a slope parameter of 0.6 did not change significantly. In the second experiment, confidence ratings data from 122 screeners conducting a simulated X-ray baggage inspection task were used to estimate ROC curves and slope parameters. Repeatedly, results could confirm that *da* is a solid measure for the investigated X-ray baggage screening task. Therefore within both experiments, d' and A' were found to be less valid sensitivity measures for the investigated X-ray baggage inspection tasks. More specifically, d' and A' did wrongly indicated lower sensitivity for a more liberal decision criterion. While the two experiments and the reviewed literature focus on X-ray image inspection, related domains such as the inspection of medical X-ray images or other visual search tasks with artificial stimuli should also not expect d' and A' to be valid without further consideration. Future research should specifically investigate to what extent the reported findings apply to other related domains.

Implications

Results from my studies and earlier literature show that the task of visual inspection in X-ray screening cannot be directly compared to a simple visual search task. The reported research shows the need for further considerations to enhance validity, internal and external, in visual search and decision-making in general. Concerning research about X-ray image inspection, attempts could be made to promote generalizations to different populations, settings, times and measures. Very often, university students are used as the main participants in traditional visual search research. Whilst this provides an easy accessible sample, it will inevitably result in selection bias, reducing the ability to generalize to a specific population. Testing more than one population could provide more variance for the results. Studies of visual inspection in X-ray screening chould therefore test screeners at multiple airports and report results in multi-experiment papers. Future research could further investigate to what extent reported findings from traditional visual search and X-ray image inspection apply to other related domains like X-ray inspection in radiology. Can the same conclusions be drawn? Are results transferable from students to radiologists? Are professionals themselves comparable regarding their visual-

cognitive abilities in visual inspection of X-ray images in security vs medical settings? However, most of these visions are questions of resources and costs.

Another important approach to higher external validity in this research area and in general should be the aspiration to report multiple measures. This could prevent researchers from only reporting measures suiting their expected results. Many of these points could for example be made as requirements to publish in certain journals to higher quality of research.

Summary and Conclusion

Addressing the question of how performance of the overall human-machine system in X-ray screening can be further enhanced while reducing search and decision-making errors, the experiments reported within this thesis provide important insights on factors and challenges influencing *search* and *decision-making* in visual inspection performance. The summarized experiments demonstrate that security and efficiency can be improved if the human factor is considered, evaluated and supported appropriately. Further, the process of using higher levels of automation should be accompanied with research about different levels of system reliability and how these affect performance indicators. Also, when developing new technological features, it needs to be considered how these may affect the task of the human operators and how the mentioned factors and challenges might change. As further advancements in screening technology are to be expected in the future, the mentioned points demonstrates the need for a joint evaluation of the performance of the human operator and the effect of advanced technology. It needs to be investigated whether the system performance can be improved and training systems, selection of screeners and supervision procedures might need adaptions. Besides performance indicators of a security checkpoint such as efficiency and effectiveness, future studies should also take the employee (screeners) and passengers into account when evaluating the outcome of the whole human-system performance by including aspects such as the work environments, motivational factors, team compositions and organizational or cultural differences.

To get more insights for applied research in the field of aviation security and visual inspection of X-ray images, laboratory research might be a good starting point. However, differences and shortcomings must be accounted for and there are still several hurdles when directly translating research from the lab to the field (e.g. Clark et al., 2012). Studies with professionals using real stimuli in preferably realistic settings are therefore important and field studies assessing the effectiveness of measures and system performance are also required (e.g. Hofer & Wetter, 2012).

To this date, the human factor remains an indispensable element for visual inspection of X-ray images despite the constant improvements in X-ray screening technology. By addressing the factors and challenges of visual search and decision-making in X-ray image inspection, this thesis treats a topic of high practical relevance and delivers important insights for the aviation security industry. Based on the results and implications, not only future research can be initiated but also security regulations can be enhanced accordingly, complying with both technological standards and human capabilities.

“The important thing is to not stop questioning. Curiosity has its own reason for existing.”

Albert Einstein (1879 –1955)

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. https://doi.org/10.3758/s13414-013-0472-7

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

Barrat, H. H. (2000). *Handbook of medical imaging*. Bellingham, WA: SPIE Press.

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Erlbaum.

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

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

Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology, 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. *Perceptual and Motor Skills, 80,* 1231–1242.

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*, Zurich Switzerland, October 5-8, 2009.

Bravo, M. J., & Farid, H. (2004). Search for a category target in clutter. *Perception, 33*, 643-652.

Bretz, E. A., (2002). Slow takeoff. *IEEE Spectrum, 39(9*), 37–39.

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

Caygill, J. S., Davis, F., & Higson, S. P. (2012). Current trends in explosive detection techniques. *Talanta, 88,* 14-29.

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*). https://doi.org/10.1371/journal.pone.0163928

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 *The Influence of Attention, Learning, and Motivation on Visual Search*, (pp. 147-181). Springer New York.Cooke & Winner, 2007

Cullen, R. H., Rogers, W. A., and Fisk, A. D., (2013). Human performance in a multiple-task environment: Effects of automation reliability on visual attention allocation. *Applied Ergonomics, 44*, 962–968.

Czaja, S. J., & Drury, C. G. (1981). Training programs for inspection, *Human Factors,23(4)*, 473–484.

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),* 564–72. https://doi.org/10.1518/001872007X215656

Drury, C. G. (1978). Integrating human factors models into statistical quality control. *Human Factors, 20,* 561-572.

Drury, C. G. (1982). *Improving inspection performance*. In G. Salvendy (Ed.), Handbook of industrial engineering. New York: Wiley.

Drury, C. G. (1992). *Inspection performance*. In G. Salvendy (Ed.), Handbook of Industrial Engineering (2nd ed.) (pp. 2282-2314). New York: John Wiley & Sons.

Drury C. G. (2001). A unified model of security inspection. In: *Proceedings of the FAA's Third International Aviation Security Technology Symposium,* November 27-20, Atlantic City, NJ.

Drury, C. G., & Fox, J. G. (1975). *Human Reliability in Quality Control*. London, Taylor & Francis, Ltd.

Drury, C. G., & Gramopadhye, A.K. (1990). Training for visual inspection. *Paper presented at the Third FAA Conference on Human Factors in Aircraft Maintenance and Inspection: Training Issues*, Atlantic City, NJ.

Drury, C. G., and Kleiner, B. M. (1990). Training in industrial environments. In *Proceedings of the 23rd Annual Conference of the Human Factors Association of Canada*, pp. 99–108.

Drury, C.G., & Sinclair, M.A. (1983). Human and machine performance in an inspection task. *Human Factors, 25*, 391-399.

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

Eilbert, R. F. (2009). Chapter 6 – X-ray technologies. In M. M. Marshall & J.C. Oxley (Eds.), *Aspects of Explosives Detection* (pp. 89–130). Oxford: Elsevier. doi:10.1016/B978-0-12-374533-0.00006-4

European Commission (2015, November 5). 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. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri =CELEX:32015R1998&from=EN

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. 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. https://doi.org/10.1002/cncy.21370

Flitton, G., Breckon, T., & Megherbi, N. (2010). Object recognition using 3D SIFT in complex CT volumes. In *British Machine Vision Conference* (pp. 11.1–11.12). Aberystwyth, Wales: BMVA Press. doi:10.5244/C.24.11

Flitton, G., Breckon, T., & Megherbi, N. (2013). A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery. *Pattern Recognition, 46(9),* 2420–2436. doi:10.1016/j.patcog.2013.02.008

Franzel, T., Schmidt, U., & Roth, S. (2012). Object detection in multi-view X-ray images. In A. Pinz, Th. Pock, H. Bischof, F. Leberl (Eds.), *Pattern Recognition* (pp. 144-154). Lecture Notes in Computer Science, Vol. 7476. Springer Berlin Heidelberg

Gallwey, T. J., (1982). Selection test for visual inspection on a multiple fault type task. *Ergonomics, 25(11),* 1077–1092.

Gallwey T. J., and Drury, C. G. (1986). Task complexity in visual inspection, *Human Factors, 28(5)*, 595–560.

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

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.

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:10.3758/s13423-014-0686-2

Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parson, K., Kuester, N., McCormac, A., Jaques, P., & Johnson, R. (2011). The role of the human operator in image-based airport security technologies. In L.C. Jain, E.V. Aidman & C. Abeynayake (Eds.), *Innovations in Defence Support Systems 2: Socio-technical Systems* (pp.147-181).

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley Halbherr, Schwaninger, Budgell, & Wales, 2013

Hättenschwiler, 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*, Taipei Taiwan, September 21-24, 2015, 25-30. doi: 10.1109/CCST.2015.7389652

Hättenschwiler, N. Mendes, M., & Schwaninger, A. (2018). Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging. *Human Factors*. 2018 Sep 24:18720818799215. doi:10.1177/0018720818799215

Hättenschwiler, 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. In *Proceedings of the 52th IEEE International 52th Carnahan Conference on Security Technology*, Montreal, Canada, October, 2018.

Hättenschwiler, N., Merks, S., Sterchi, Y., & Schwaninger, A. (n.d.). Traditional visual search vs. X-ray image inspection in students and professionals: Are the same visual cognitive abilities needed? Under Review *Journal of Experimental Psychology: Applied*.

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. https://doi.org/10.1016/j.apergo.2018.05.003

Hancock, P. A., & Hart, S. G. (2002). Defeating terrorism: What can human factors/ergonomics offer? *Ergonomics in Design, 10(1),* 6–16.

Harding, G. (2004). X-ray scatter tomography for explosives detection. *Radiation Physics and Chemistry, 71*, 869–881. doi:10.1016/j.radphyschem.2004.04.111.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006a). 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*, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 393-397.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006b). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. *Proceedings of the 4th International Aviation Security Technology Symposium*, Washington, D.C., USA, November 27 – December 1, 2006, 338-342.

Hardmeier, D., & Schwaninger, A. (2008). Visual cognition abilities in x-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation*, ICRAT 2008, Fairfax, Virginia, USA, June 1-4, 2008, 311-316.

Harris, D. H. (2002). How to really improve airport security. *Ergonomics in Design, 10(1),* 17-22

Harris, D. H., & Chaney, F. B. (1969). *Human Factors in Quality Assurance*. New York, John Wiley and Sons.

Hofer, F., Hardmeier, D., & Schwaninger, A. (2006). Increasing airport security using the X-ray ORT as effective pre-employment assessment tool. *Proceedings of the 4th International Aviation Security Technology Symposium*, Washington, D.C., USA, November 27 – December 1, 2006 303-308.

Hofer, F., & Wetter, O. E. (2012). Operational and human factors issues of new airport security technology—two case

studies. *Journal of Transportation Security, 5(4),* 277-291.

Huang, L., & Pashler, H. (2005). Attention capacity and task difficulty in visual search. *Cognition, 94(3),* B101–B111. 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. 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., 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. 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.

Koller, S., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2nd International Conference on Research in Air Transportation*, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 399-402.

Latorella, K., & Drury, C. G. (1992). A framework for human reliability in aircraft inspection. In: *Proceedings of the Seventh FAA Meeting on Human Factors Issues in Aircraft Maintenance and Inspection*. Federal Aviation Administration, Washington, DC.

Macmillan, N. A., & Creelman, C. D. (1992). *Detection theory: A user's guide.* New York: Cambridge University Press.

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

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.

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.

Megherbi, N., Breckon, T. P., Flitton, G. T., & Mouton, A. (2012, October). Fully Automatic 3D Threat Image Projection: Application to Densely Cluttered 3D Computed Tomography Baggage Images. In *Image Processing Theory, Tools and Applications (IPTA)*, 2012 3rd International Conference on (pp. 153-159). IEEE.

Mendes, M., Schwaninger, A., Strebel, N., & Michel, S. (2012). Why laptops should be screened separately when conventional x-ray screening is used. *Proceedings of the 46th IEEE International Carnahan Conference on Security Technology*, Boston MA, October 15-18, 2012.

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

Merks, S., Hättenschwiler, N., Melina, Z., & Schwaninger, A. (2018). X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging? In *Proceedings of the 52th IEEE International 52th Carnahan Conference on Security Technology*, Montreal, Canada, October, 2018.

Mery, D., Mondragon, G., Riffo, V., & Zuccar, I. (2013). Detection of regular objects in baggage using multiple X-ray views. *Insight: Non-Destructive Testing and Condition Monitoring, 55(1),* 16–20. https://doi.org/10.1784/insi.2012.55.1.16

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of Reliance and Compliance in Aided Visual Scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 56(5),* 840–849. https://doi.org/10.1177/0018720813512865

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., & Schwaninger, A. (2014). A multi-method approach towards identifying situational factors and their relevance for X-ray screening. *Proceedings of the 48th IEEE International Carnahan Conference on Security Technolog*y, Rome Italy, October 13-16, 2014, 208-213.

Miyazaki, Y. (2015). Influence of being videotaped on the prevalence effect during visual search. *Frontiers in Psychology, 6,* 583. https://doi.org/10.3389/fpsyg.2015.00583

Mitroff, S. R., Biggs, A. T., & Cain, M. S. (2015). Multiple-target visual search errors: Overview and implications for airport security. *Policy Insights from the Behavioral and Brain Sciences. 2(1),* 121–128.

Mouton, A., & Breckon, T. P. (2015). A review of automated image understanding within 3D baggage computed tomography security screening. *Journal of X-ray Science and Technology, 23(5),* 531–555. doi:10.3233/XST-150508

Nakashima, R., Kobayashi, K., Maeda, E., Yoshikawa, T., & Yokosawa, K. (2013). Visual search of experts in medical image reading: the effect of training, target prevalence, and expert knowledge. *Frontiers in psychology, 4,* 1-8.

Nakayama, K., & Martini, P. (2010). Situating visual search. *Vision Research*. doi: 10.1016/j.visres.2010.09.003.

Neeman, A. (2012). X-rays in focus: Past, present and future. *Aviation Security International, 18(6),* 18-21.

Nercessian, S., Panetta, K., & Agaian, S. (2008, May). Automatic detection of potential threat objects in X-ray luggage scan images. In *Technologies for Homeland Security, 2008 IEEE Conference* on (pp. 504-509). IEEE.

Novakoff, A.K. (1993). FAA bulk technology overview for explosives detection, *SPIE 1824,* 2–12.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In *Attention and Performance IX*, eds. J. Long and A. Baddeley (Hillside, NJ: Lawrence Erlbaum), 135-151.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors, 29*, 695–706.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39(2*), 320-253.

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.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3(1)*, 1–23. doi:10.1207/s15327108ijap0301_1

Parasuraman, R., Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50(3)*, 511–520. doi:10.1518/001872008X312198

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.

Rice, S., & McCarley, J. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied, 17(4*), 320–331.

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),* 1–17. doi: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*, 439. doi: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*, 1327–1344. https://doi.org/10.1080/17470218.2012.656662

Schuster, D., Rivera, J., Sellers, B.C., Fiore, S.M., & Jentsch, F. (2013). Perceptual training for visual search. *Ergonomics, 56(7)*, 1101-1115.

Schwaninger, A. (2003a). Reliable measurements of threat detection. *AIRPORT, 2003(1),* 22-23.

Schwaninger, A. (2003b). Evaluation and selection of airport security screeners. *AIRPORT, 2003(2),* 14-15.

Schwaninger, A. (2005). Increasing efficiency in airport security screening. *WIT Transactions on the Built Environment,* 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, D.C., USA, November 27 – December 1, 2006, 265-270Schwaninger, Hardmeier & Hofer, 2005

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., & Hay, L. (2008). The impact of image based factors and training on threat detection performance in X-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation*, ICRAT 2008, Fairfax, Virginia, USA, June 1-4, 2008, 317-324.

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

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38*, 258-264.

Schwaninger, A., Hardmeier, D., Riegelnig, J., & Martin, M. (2010). Use it and still lose it? The influence of age and job experience on detection performance in x-ray. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry, 23(3)*, 169-175.

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

Sheehan, J.J., & Drury, C.G. (1971). The analysis of industrial inspection. *Applied Ergonomics, 2.2*, 74–78.

Shepherd, W., & Parker, J. (1990). Human factors issues in aircraft maintenance and inspection "training issues. In *Final Report of the Third FAA Meeting,* Atlantic City, NJ

Sheridan, T. B., & Parasuraman, R. (2000). Human versus automation in responding to failures: An expected-value analysis. Human Factors: *The Journal of the Human Factors and Ergonomics Society, 42(3),* 403-407.

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

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*, Madrid Spain, October, 2017. doi: 10.1109/CCST.2017.8167812

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*, Taipei Taiwan, September 21–24. doi:10.1109/CCST.2015.7389657

Thapa, V., Gramopadhye, A.K., Melloy, B., and Grimes, L. (1996). Evaluation of different training strategies to improve

decision-making performance in inspection. *The International Journal of Human Factors in Manufacturing, 6(3),* 243–261.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlenbaum.

Tarr, M. J., & Vuong, Q. C. (2002). *Visual object recognition*. In H. Pashler (Series Ed.) & S. Santis (Ed.), Stevens' handbook of experimental psychology: Vol. 1. Sensation and perception (3rd ed., Vol. 1, pp. 287–314). New York, NY: Wiley. doi:10.1002 /0471214426.pas0107

Global Terrorism Database. https://www.start.umd.edu/gtd/ (accessed 15.11.2017).

Turner, S. (1994). *Terrorist explosive sourcebook countering terrorist use of improvised explosive devices*. Boulder CO: Paladin Press.

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? *Applied Ergonomics, 53*, 190–202. https://doi.org/10.1016/j.apergo.2015.09.013

Verghese, P. (2001). Visual search and attention: A signal detection approach. *Neuron. 31,* 523–535. doi:10.1016/S0896-6273(01)00392-0

Vuong, Q. C., & Tarr, J. T. (2004). Rotation direction affects object recognition. *Vision Research, 44*, 1717–1730. doi:10.1016/j.visres.2004.02.002

Wales, A., Anderson, C., Jones, K., Schwaninger, A., & Horne, J. (2009). Evaluating the two-component inspection model in a simplified luggage search task. *Behavior Research Methods, 41(3),* 937-943.

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.

Weiner, E.L. (1986). Vigilance and inspection. *In Sustained Attention and Human Performance*. By J.S. Warm (ed.) (John Wiley and Sons, London).

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.

Wetter, O. (2013). Imaging in airport security: Past, present, future and the link to forensic and clinical radiology. *Journal of Forensic Radiology and Imaging, 1*, 152-160.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. New York: Oxford University Press.

Wickens, C.D., & Colcombe, A. (2007). Performance consequences of imperfect alerting automation associated with a cockpit display of traffic information. *Human Factors, 49*, 839–850.

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. doi:10.1080/14639220500370105

Wiener, E.I. (1984). Vigilance and inspection. In J.S. Warm (Ed.), *Sustained Attention in Human Performance* (pp. 207-246). Chichester: Wiley.

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.

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.

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

Publication list

Hättenschwiler, N., Mendes, M., & Schwaninger, A. (2018). Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging. *Human Factors*. doi:10.1177/0018720818799215

Hättenschwiler, N., Merks, S., & Schwaninger, A. (2018). Airport security screening of hold baggage: 2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol. Proceedings of the 52th IEEE International Carnahan Conference on Security Technology, Montreal Canada, October, 2018.

Hättenschwiler, N., Merks, S., Sterchi, Y., & Schwaninger, A. (n.d.). Traditional visual search vs. X-ray image inspection in students and professionals: Are the same visual cognitive abilities needed? Under Review in *Frontiers in Psychology: Cognition*.

Hättenschwiler, 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, Taipei Taiwan, September 21-24, 2015, 25-30. doi: 10.1109/CCST.2015.7389652

Hättenschwiler, N., Michel, S., Kuhn, M., Ritzmann, S., & Schwaninger, A. (2016). Eine erste explorative Studie zur Relevanz von Wissen über das Aussehen von Alltagsgegenständen bei der Röntgenbildbeurteilung in der Luftsicherheit. 62. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft (GfA), Aachen, 02.-04. März, 2016, 1-5.

Hättenschwiler, N., Michel, S., Kuhn, M., Strebel, N., & Schwaninger, A. (2015). Relevanz situativer Einflussfaktoren auf die Arbeit von Luftsicherheitskontrollpersonal bei der Röntgenbildbeurteilung - eine Arbeitsanalyse. 61. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft (GfA), Karlsruhe, 25.-27. Februar, 2015, 1-5.

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.

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), Brugg AG, 15.-17. Februar, 2017, 1-5.

Merks, S., Hättenschwiler, N., Zeballos, M. & Schwaninger, A. (2018). X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging?. *Proceedings of the 52th IEEE International Carnahan Conference on Security Technology*, Montreal Canada, October, 2018

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., & Schwaninger, A. (2014). A multi-method approach towards identifying situational factors and their relevance for X-ray screening. *Proceedings of the 48th IEEE International Carnahan Conference on Security Technology, Rome Italy, October 13-16, 2014*, 208-213. doi: 10.1109/CCST.2014.6987001

Michel, S., Hättenschwiler, N., Zeballos, M., & Schwaninger, A. (2017). Comparing e-learning and blended learning for threat detection in airport security X-ray screening. Proceedings of the 51th IEEE International Carnahan Conference on Security Technology, Madrid, Spain, October 23-26, 2017

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, Madrid Spain, October, 2017. doi: 10.1109/CCST.2017.8167812

Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (n.d.). Detection Measures for Visual Inspection of X-ray Images of Passenger Baggage. Under Review in *Attention, Perception, & Psychophysics*.

Appendix

1) Hättenschwiler, 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,* Taipei Taiwan, September 21-24, 2015, 25-30. doi:10.1109/CCST.2015.7389652

2) 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*, Madrid Spain, October, 2017. doi: 10.1109/CCST.2017.8167812

3) 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.

4) Hättenschwiler, N. Mendes, M., & Schwaninger, A. (2018). Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging. *Human Factors,* doi:10.1177/0018720818799215.

5) Hättenschwiler, N., Merks, S., & Schwaninger, A. (2018). Airport security screening of hold baggage: 2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol. *Proceedings of the 52th IEEE International Carnahan Conference on Security Technology*, Montreal Canada, October, 2018.

6) Merks, S., Hättenschwiler, N., Zeballos, M. & Schwaninger, A. (2018). X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging?. *Proceedings of the 52th IEEE International Carnahan Conference on Security Technology*, Montreal Canada, October, 2018.

7) Hättenschwiler, N., Merks, S., Sterchi, Y., & Schwaninger, A. (n.d.). Traditional visual search vs. X-ray image inspection in students and professionals: Are the same visual cognitive abilities needed? Under Review in *Frontiers in Psychology: Cognition*.

8) Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (n.d.). Detection Measures for Visual Inspection of X-ray Images of Passenger Baggage. Under Review in *Attention, Perception, & Psychophysics*.

A first exploratory study on the relevance of everyday object knowledge and training for increasing efficiency in airport security X-ray screening

Nicole Hättenschwiler, Stefan Michel, Milena Kuhn, Sandrina Ritzmann 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*— Secure air transportation is vital for economy and society and it relies heavily on airport security screening. Passenger bags and other belongings are screened using X-ray machines to ensure that they do not contain prohibited items. Human operators (X-ray screeners) visually inspect X-ray images to decide whether they are harmless or whether they might contain a prohibited item and therefore require secondary search (typically using manual search and/or explosive trace detection technology). Several previous studies have shown that learning which items are prohibited and what they look like in X-ray images of passenger bags is important to achieve good detection performance. As passenger bags contain a large variety of harmless everyday objects, it could be assumed that knowing what such objects look like in X-ray images could help X-ray screeners to work more efficiently by reducing false alarms (i.e. sending a passenger bag to secondary search even though it does not contain a prohibited item). In the first experiment, the relationship between knowledge of harmless everyday objects and false alarm rate was investigated with 15 certified X-ray screeners of one large European airport. Statistical analyses revealed a good knowledge of harmless everyday objects on average with some variation between X-ray screeners and a negative correlation with false alarm rate. In the second experiment, the effectiveness of an e-learning course for acquiring knowledge of everyday objects in X-ray images was evaluated. Thirty novices conducted a test-retest experiment where half of the participants conducted an e-learning course about harmless everyday objects in X-ray images between the two tests. The results revealed that e-learning can be an effective and efficient method for increasing the knowledge of everyday objects in X-ray images. Based on the results of both studies, the relevance to learn everyday objects as part of initial and recurrent training of X-ray screeners is discussed.**

Keywords — aviation security; X-ray screening; detection performance; everyday object recognition; everyday object training

I. INTRODUCTION

Reviews of attacks against civil aviation since 11th September 2001 underline the importance of aviation security measures (e.g. [1]). At airport security checkpoints, passengers and their belongings are screened to ensure that prohibited items (guns, knives, improvised explosive devices (IED) and other threat items) are not carried. State-of-the-art X-ray screening equipment offers good image quality with high resolution, automated detection of explosives, several image enhancement functions and other features [2]; [3]; [4]. The main task of an X-ray screener is to visually inspect X-ray images of passenger bags and to decide whether a bag is harmless or whether it might contain a prohibited item and therefore needs further inspection by secondary search (typically by using manual search and/or explosive trace detection technology). As pointed out by [5]; [6]; [7] prohibited items are difficult to recognize without training because 1) objects often look very different in X-ray images than in reality, 2) certain prohibited items are not known from everyday experience (e.g. IEDs), 3) some prohibited items look similar to harmless objects (e.g. a switchblade knife can resemble a pen), and 4) when objects are depicted from unusual viewpoints, they become difficult to recognize. During initial classroom, computer-based[1] and on the job training, X-ray screeners learn how to interpret X-ray images in order to recognize everyday objects and prohibited items. In the last decade, several studies have shown that computer-based training (CBT[1]) is important, efficient and effective to achieve good detection performance in X-ray image interpretation ([8]; [9]; [10]; [11]; [12]; [13]; [14]; [15]). While these studies have provided converging on the importance to learn which items are prohibited and what they look like in X-ray images, the role of everyday object knowledge has not yet been addressed specifically. This is an interesting topic especially from an operational point of view. In particular, one could assume that the knowledge on what everyday objects look like in X-ray images could result in fewer cases where an everyday object is confused with a prohibited item (e.g. pen can resemble a switchblade knife). This would result in fewer false alarms, i.e. wrongly judging a

[1] In the literature, the term CBT is used in different ways. The definition used here refers to all forms of self-paced distance training and learning activities using computers [16] and therefore includes e-learning courses.

bag to contain a prohibited item. False alarms have to be resolved by secondary search which typically involves manual search and/or alarm resolution using explosive trace detection technology [17]. Due to the additional time needed for secondary search, high false alarm rates can have a strong negative impact on throughput [17] and they could also result in lower passenger satisfaction [18]. Therefore, it is worth investigating the role of everyday object knowledge and training because it could be very relevant for more efficient X-ray screening.

In this first exploratory study, two experiments were conducted. For both experiments, a new test was created to measure how well novices and X-ray screeners can categorize and name everyday objects in X-ray images. In Experiment 1 bivariate and partial correlation analyses were performed to examine whether there is a statistically significant and meaningful relationship between everyday object knowledge and false alarm rate in a simulated X-ray baggage screening task. In Experiment 2 it was investigated whether e-learning can be used to learn effectively and efficiently everyday object knowledge which could be relevant for more efficient X-ray screening.

II. EXPERIMENT 1

The main aim of Experiment 1 was to investigate the relationship between everyday object knowledge and false alarm rates. To this end, X-ray screeners were tested using an everyday object categorization and naming test and a simulated X-ray baggage screening task. Two X-ray screeners with experience in test development and teaching X-ray image interpretation assisted for creating the stimuli. They did not participate as subjects in the experiments and are referred to as X-ray screening experts from now on.

III. METHOD AND PROCEDURE

Fifteen X-ray screeners (7 female), with a mean age of 39.8 years (*SD* = 10.13) and a mean work experience of 6.03 years (*SD* = 4.44) in cabin baggage screening at a large international European airport were tested. The X-ray screeners first conducted an X-ray object categorization and naming test (X-Ray OCNT) and four weeks later a simulated X-ray baggage screening task (XBST).

A. X-Ray Object Categorization and Naming Test (X-Ray OCNT)

For the purpose of this study, 32 X-ray images of passenger bags were selected from a pool of 2800 images by the X-ray screening experts to be representative.

Fig.1. Screenshot showing an X-ray image of a passenger bag from the X-ray OCNT with framed harmless everyday objects.

X-ray images of 17 prohibited items were blended into the X-ray images of passenger bags using a validated X-ray image blending software provided by CASRA. Then, three objects per bag were marked by a red frame. 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 that had to be categorized and named.

For each object to be categorized, participants had to click one of three option buttons describing the categories: harmless everyday object, improvised explosive device (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. Simulated X-ray baggage screening task (XBST)

160 color X-ray images of passenger bags were selected by the two X-ray screening experts to be representative. In half of the X-ray images, one prohibited item was added using the validated X-ray image blending software mentioned above. Four categories of prohibited items were used (guns, knives, IEDs and other prohibited items). For each category there were ten exemplars. Each exemplar was displayed once in easy view (as defined by the two X-ray screening experts and the authors) and difficult view (rotated around the horizontal or vertical axis by 85 deg.). For each category, half of the prohibited items were part of the CBT system used at this airport. Since it cannot be assumed that terrorists would use a prohibited item that is contained in the CBT, half of the prohibited items were newly recorded and visual comparison was used to make sure that they are different from the prohibited items contained in the CBT. The X-ray images were displayed on the screen without time limit and X-ray screeners had to decide whether it contained a prohibited item or not by pressing a key. There was no feedback on the correctness of responses and the participants took about 30 minutes to complete the test.

IV. RESULTS AND DISCUSSION

On average, 96.19% of the objects were correctly categorized (*SD* = 1.67%) in the X-ray OCNT. Harmless

everyday objects were correctly named on average in 82.07% of the cases (*SD* = 11.19%)[2]. In the XBST, when X-ray screeners wrongly judged a bag to contain a prohibited item, this counted as a false alarm. False alarm rate was calculated by number of false alarms divided by number of X-ray images of bags not containing a prohibited item[3]. On average, screeners had a mean false alarm rate of *M* = .037 (*SD* = .031). As can be seen in Fig. 2 there is a negative linear relationship between % correctly named harmless everyday objects and false alarm rate. Even though a relatively small number of X-ray screeners participated in Experiment 1 (n=15), a one-tailed[4] Pearson correlation was significant, $r(13) = -.471$, $p=.038$, showing an effect size which is modest according to [19].

Two observations are worth mentioning when looking at Fig. 2. First, there was substantial variation between X-ray screeners regarding % correctly named harmless everyday objects ranging from 54% to 96%%. Two X-ray screeners had quite low values of 54%. Since correlations are very sensitive to outliers, it was examined whether the values of these two screeners on % correct naming are below the often used criterion of three standard deviations (11.19% *3 = 33.57%) from the mean (82.07%). Since this was not the case, the found correlation can be regarded as valid subject to verification with a larger sample size (for a review and discussion on different methods for outlier analysis, see [20]). Second, although all X-ray screeners had false alarm rates measured with the XBST that were below 11%, there was still enough variation between screeners so that a correlation with % correct naming could be revealed.

A key concept of signal detection theory [21] is that variation in false alarm rate can also result from differences in response bias, i.e. a tendency to report more or less often that a signal (in this case a prohibited item) is present. Response bias depends on a variety of factors including individual ones, stimuli used, test design and other situational factors ([21]; for a more recent detailed overview and discussion on detection theories, see [22]). An often used measure of response bias is the criterion which can be calculated as follows:

$$c = -0.5[z(H) + z(FA)] \quad (1)$$

H refers to hit rate which in the XBST corresponds to number of correct decisions for X-ray images containing a prohibited item divided by number of X-ray images of passenger bags containing a prohibited item.

FA refers to false alarm rate, which is calculated as defined above.

z refers to the z transformation which is used to convert hit and false-alarm rates to z scores (i.e. standard deviation units).

Fig. 3 shows the relationship between response bias (criterion) and false alarm rate. The high negative bivariate Pearson correlation, $r(13) = -.870$, $p<.001$, with a high effect size according to [19] is consistent with signal detection theory according to which response bias strongly influences the false alarm rate.

Although it seems not plausible to assume that X-ray screeners with good knowledge of harmless everyday objects would have systematically different response biases, this possibility should be ruled out by showing that the correlation between % correctly named harmless everyday objects in the X-ray OCNT and the false alarm rate from the XBST is not due to the background variable response bias. Therefore, a partial correlation (Pearson) between % correctly named harmless everyday objects in the X-ray OCNT and the false alarm rate from the XBST controlling for response bias (criterion c) was performed. The result $r(13) = -.492$, $p=.037$ (one-tailed) shows that variation in response bias cannot explain the correlation between everyday object knowledge and false alarm rate. In contrary, when controlling for response bias, the correlation between everyday object knowledge and false alarm rate becomes even slightly higher.

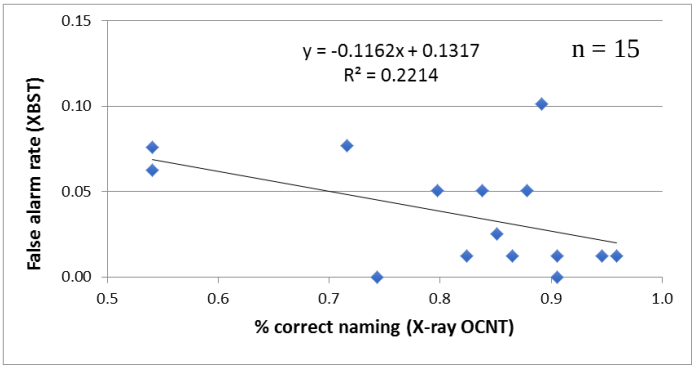

Fig.2. Correlation between % correctly named harmless everyday objects (X-ray OCNT) and false alarm rate (XBST). The explained variance is provided as R^2.

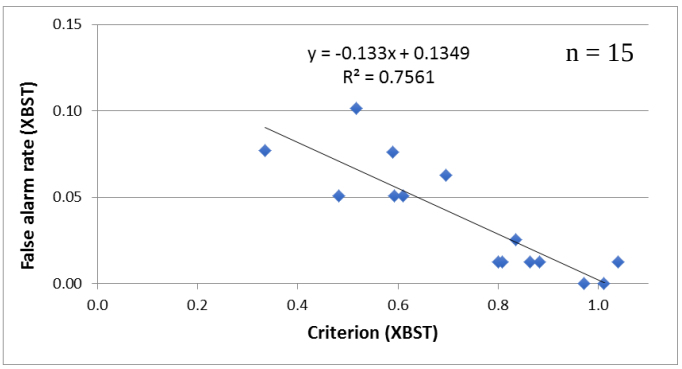

Fig.3. Correlation between response bias (criterion c) and the false alarm rate of the XBST. The explained variance is provided as R^2.

[2] The results for naming prohibited items in the X-ray OCNT are not reported because this information could be regarded as security sensitive and because these results are not relevant for testing the relationship between knowledge of harmless everyday objects and false alarm rates from the XBST.

[3] Due to recording failure of the software a few trial responses were missing. More specifically, for six of the 15 X-ray screeners, one response was missing. This resulted in a total of 0.25% missing responses which for the calculation of the false alarm rate and the criterion resulted in negligible deviations compared to when all responses would have been recorded.

[4] Since the hypothesis stated in the introduction would predict that with increasing everyday object knowledge, false alarm rates would decrease, one-tailed significance levels are reported.

These results are highly consistent with the hypothesis stated in the introduction, namely that everyday object knowledge helps reducing false alarm rates in X-ray screening tasks which is important for increasing efficiency.

It should be noted that causation cannot be inferred from correlational analyses even if partial correlation is used. However, the results of Experiment 1 are at least encouraging with regard to further studies in which the role of everyday object knowledge could be examined using an experimental design in which the effect of systematic variation of everyday object knowledge on X-ray screening performance is investigated, ideally including a control group.

V. EXPERIMENT 2

While several studies exist on the detection of prohibited items by X-ray screeners before and after several months of CBT ([8]; [10]; [11]; [12]; [13]; [14]; [15]; [23]), no study has been conducted yet to investigate how well novices can recognize harmless everyday objects in X-ray images and whether this can be trained efficiently and effectively. To address this topic, a test-training-test design was used with an experimental and a control group.

VI. METHOD AND PROCEDURE

The experimental design is illustrated in Fig.4. Thirty novices (15 female) with a mean age of 31.1 years (*SD* = 11.87) and no work experience in X-ray screening were first tested on their visual abilities using the X-Ray Object Recognition Test (X-Ray ORT; [24]) as described below. Afterwards, two equal groups (an experimental and a control group) were built based on the results from the X-Ray ORT so that the mean and standard deviation of *d'* (a measure of sensitivity, Green & Swets, 1966) did not differ significantly between both groups, $t(28) = 1.176$, $p = .25$. Afterwards, all participants conducted an adapted version of the X-Ray OCNT (X-Ray ONT) used in Experiment 1 (prohibited items were excluded) to assess their knowledge of harmless everyday objects. Afterwards, the experimental group conducted an e-learning course on everyday objects, which is described below. The control group did not receive any training. Both groups were tested again one week later using the X-Ray ONT.

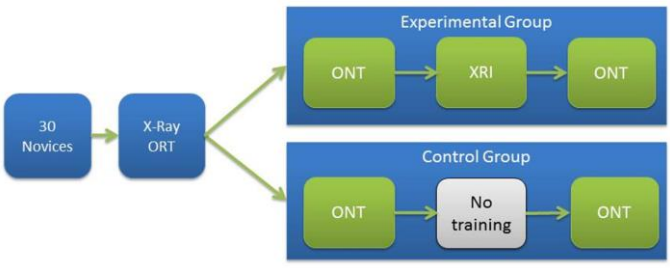

Fig.4.Experimental design used in Experiment 2.

A. X-Ray Objects Recognition Test (ORT)

The X-Ray ORT is an X-ray image interpretation test which was developed to measure the ability to cope with image-based factors in X-ray image interpretation (i.e. effects of viewpoint, superposition by other objects, and bag complexity). It consists of 256 grey-scale X-ray images of passenger bags. Half of them contain either a gun or a knife; the other 128 X-ray images are harmless bags. Each bag is displayed for 4 seconds on the screen and participants have to decide for each image whether the bag is OK (i.e. there is no prohibited item in the bag) or whether it is NOT OK (i.e. the bag contains a gun or a knife) by clicking on a button on the screen. The X-Ray ORT takes about 30 minutes to complete. Information on test construction, its reliability and validity measures can be found in [24]; [25].

B. X-Ray Introduction of Everyday Objects (XRI)

X-Ray Introduction (XRI) is an e-learning course for beginners teaching the look of harmless everyday objects in cabin baggage and air cargo, which was provided by CASRA. The XRI is recommended for people who have not worked with X-ray images before. In a short introduction, X-ray technology and the meaning of colors in X-ray images are explained. In the following nine sessions, 90 harmless everyday objects are introduced in three steps. First, harmless everyday objects are shown as X-ray images together with a photograph. As objects sometimes look very different in X-ray images compared to reality, this is an important step to learn what everyday objects look like in X-ray images. Second, X-ray images of a passenger bags containing these everyday objects are displayed one after the other, giving the trainees the opportunity to view the introduced objects within a realistic context. Then, trainees have to find each learned everyday object in X-ray images of passenger bags by clicking on the object indicated in the instruction. Participants can also monitor their learning progress in form of a self-evaluation exercise.

The XRI was conducted without interruption and took about 2.5h to complete.

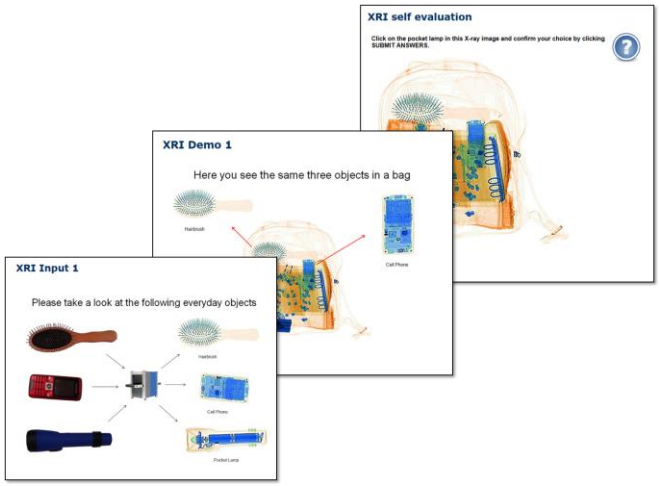

Fig.5. Screenshots of the XRI: Input, demo and self-evaluation.

C. Everyday Objects Naming Test (X-Ray ONT)

The X-Ray OCNT from Experiment 1 was used with two adaptations: 1) the red frames which marked the prohibited items were deleted which resulted in 19 X-ray images containing three harmless everyday objects, nine X-ray images containing two harmless objects and four X-ray images

containing one harmless object. 2) Objects had only to be named and not categorized[5]. Different X-ray images were used for the X-Ray ONT and the XRI. The participants completed the test in about 45 - 60 minutes.

VII. Results and Discussion Experiment 2

Fig.6 shows the results of the X-Ray ONT for the experimental and the control group. When conducting the X-ray ONT the first time, the groups did not differ significantly, $t(28) = -.377$, $p = .709$. The control group achieved 52.72% correct naming ($SD = 12.48$), and the experimental group 54.55% ($SD = 14.11$). This result indicates that novices could recognize about half of the everyday objects in X-ray images used in the X-Ray ONT before taking any training. While this result is encouraging, it also means that many everyday objects cannot be recognized in X-ray images without training. When conducting the X-Ray ONT the second time, the control group achieved 58.28% correct naming ($SD = 10.82$), whereas the experimental group reached 66.05% ($SD = 10.29$). Bonferroni-corrected post-hoc analyses revealed a significant difference between the X-ray ONT performance before and after training for the experimental group $t(28) = -2.549$, $p = .008$, but not for the control group $t(28) = -1.305$, $p = .101$. To examine main effects and interaction, a mixed 2x2 univariate ANOVA was conducted with the dependent variable % correct naming (X-Ray ONT), the within-subjects factor time (before and after taking the XRI) and the between-subject factor group (experimental vs. control group). There was no main effect of group, $F(1, 28) = 1.26$, $p = .272$, a main effect of time, $F(1, 28) = 77.89$, $p < .001$, and a significant interaction between time and group, $F(1, 28) = 9.42$, $p = .005$.

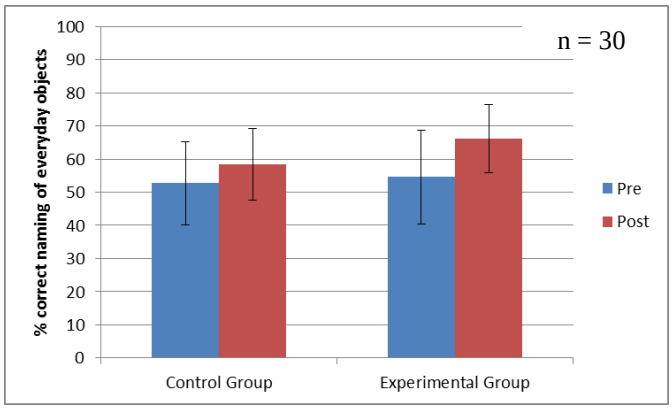

Fig.6 Means and standard deviations for % correct naming in the X-Ray ONT for the experimental and control group before and after the experimental group conducted the XRI.

These results can be summarized as follows: Novices could recognize about half of the everyday objects in the X-Ray ONT without training. Using a short e-learning module (XRI) during 2.5 hours did already increase % correct naming in the X-ray ONT by about 12%. This result is very encouraging regarding the question whether knowledge of harmless everyday objects in X-ray images can be trained effectively and efficiently.

[5] Categorization would not make sense since only one category ("harmless object") would be applicable.

VIII. Summary, Conclusions and Limitations

To get first insights into the relevance of knowledge of everyday objects in X-ray security screening, two experiments were conducted. Firstly, to explore the relationship between the knowledge about harmless everyday objects and the false alarm rate measured in a simulated X-ray baggage screening task, and secondly, to investigate the effectiveness of an e-learning course on harmless everyday objects in X-ray images.

From an operational perspective, a low false alarm rate is desirable to guarantee the efficiency of the security screening process. We assumed a negative relationship between the percentage of correct answers for the naming of harmless everyday objects in an object categorization and naming task and the false alarm rate in a simulated X-ray baggage screening task. This hypothesis could be confirmed while controlling for variation in response bias using partial correlation. As mentioned earlier, it should be noted that causation cannot be inferred from correlational analyses even if partial correlation is used. Moreover, Experiment 1 was conducted with a rather small sample size (n=15). Further studies with a bigger sample size using an experimental design with systematic variation of everyday object knowledge and a control group would lead to stronger causal conclusions. Also, the performance of X-ray screeners with CBT on prohibited item detection who either do or do not receive additional everyday object training could be compared.

Experiment 2 showed promising results on e-learning as an efficient and effective tool for building knowledge of harmless everyday objects in X-ray images. This is certainly important for initial training so that X-ray screeners can work efficiently as early as possible. Although on average, certified X-ray screeners reached high values in everyday object naming in Experiment 1, there was still substantial inter-individual variation. Therefore, it could makes sense to run additional studies in which cost benefit analyses are conducted to see whether everyday object recognition training should be implemented not only for initial but also for recurrent training.

To conclude, this report shed light on the role of knowledge of harmless everyday objects for X-ray screening. Such knowledge, in combination with sound visual knowledge of prohibited items, could have positive effects on the effectiveness and the efficiency of airport security screening.

References

[1] T. Hunter, "Islamist fundamentalist and separatist attacks against civil aviation since 11th September 2001," in F. Chau (Ed.), Aviation security challenges and solutions, Hong Kong: Avseco, pp. 35-54, 2011.

[2] S. Michel, and A. Schwaninger, A. "Human-machine interaction in x-ray screening," Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology, Zurich Switzerland, October 5-8, 2009.

[3] K. Wells, and D.A. Bradley, D. a. "A review of X-ray explosives detection techniques for checked baggage," Applied Radiation and Isotopes, 70(8), pp. 1729–1746, 2012. doi:10.1016/j.apradiso.2012.01.011

[4] O.E. Wetter, "Imaging in airport security: Past, present, future, and the link to forensice and clinical radiology," Journal of Forensic Radioloty and Imaging, 1, pp. 152-160, 2013.

[5] A. Schwaninger, "Training of airport security screeners," AIRPORT, 5, pp. 11-13, 2003.

[6] A. Schwaninger, "Increasing efficiency in airport security screening," WIT Transactions on the Built Environment, pp. 407-416, 2005.

[7] A. Schwaninger, " Airport security human factors: From the weakest to the strongest link in airport security screening," Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, pp. 265-270, November 27 – December 1 2006.

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

[9] D. Hardmeier, F. Hofer, and A. Schwaninger, "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, Serbia and Montenegro, pp. 393-397, June 2006.

[10] S. Michel,J. de Ruiter, M. Hogervorst, S. Koller, R. Moerland, and A. Schwaninger, A. "Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners," Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 8-11, 2007.

[11] A. Schwaninger, F. Hofer, and O. Wetter, "Adaptive computer-based training increases on the job performance of x-ray screeners," Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 8-11, 2007.

[12] S. Koller, D. Hardmeier, S. Michel, and A. Schwaninger, "Investigating training, transfer and viewpoint effects resulting from recurrent CBT of x-ray image interpretation," Journal of Transportation Security, 1(2), pp. 81-106, 2008.

[13] S. Koller, C. Drury, and A. Schwaninger, "Change of search time and non-search time in X-ray baggage screening due to training," Ergonomics, 52(6), pp. 644-656, 2009.

[14] T. Halbherr, A. Schwaninger, G. Budgell, and A. Wales, "Airport security screener competency: a cross-sectional and longitudinal analysis," International Journal of Aviation Psychology, 23(2),pp. 113-129, 2013.

[15] S. Michel, M. Mendes, J. de Ruiter, G. Koomen, and A. Schwaninger, "Increasing X-ray image interpretation competency of cargo security screeners," International Journal of Industrial Ergonomics , 44, pp. 551-560, 2014.

[16] A. Schwaninger, A. "Computer-based training: advantages and considerations," Aviation Security International, 17(6), pp. 18-23, 2011.

[17] Y. Sterchi, and A. Schwaninger, " Optimizing the introduction of EDS in cabin baggage screening: A first simulation of the effect on throughput," Proceedings of the 46th IEEE International Carnahan Conference on Security Technology, Taipei, September 21-24, 2015.

[18] K. Gkritza, D. Niemeier, and F. Mannering, "Airport security screening and changing passenger satisfaction: An exploratory assessment," Journal of Air Transport Management, 12(5), pp. 213–219, 2006. doi:10.1016/j.jairtraman.2006.03.001

[19] J. Cohen, "A power primer," Psychological Bulletin, 112, pp. 155-159, 1992.

[20] D. Cousineau, and S. Chartier, "Outliers detection and treatment: a review," International Journal of Psychological Research, 3(1), pp. 58-67, 2010.

[21] D.M. Green, and J.A. Swets, "Signal Detection Theory and Psychophysics," New York: Wiley, 1966.

[22] N.A. Macmillan, and C.D. Creelman, "Detection theory: A users guide (2nded.)," New York: Cambridge University Press, 2004.

[23] D. Hardmeier, F. Hofer, and A. Schwaninger, "The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening," Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, pp. 338-342, 2006.

[24] D. Hardmeier, F. Hofer, and A. Schwaninger, "The x-ray object recognition test (x-ray ort) – a reliable and valid instrument for measuring visual abilities needed in x-ray screening," IEEE ICCST Proceedings, 39, pp. 189-192, 2005.

[25] D. Hardmeier, F.Hofer, and A. Schwaninger, "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, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, pp. 393-397, 2006.

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).

978-1-5386-1585-0/17/$31.00 ©2017 IEEE

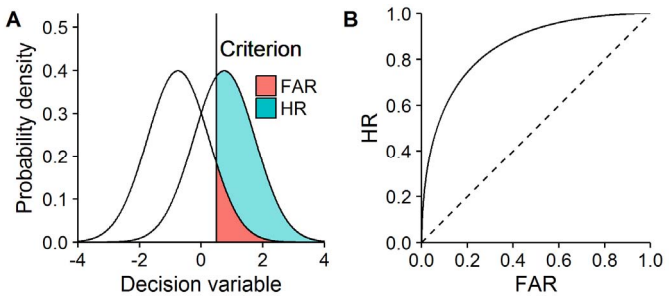

Fig. 1. Illustration of SDT. A: Noise and signal plus noise distribution, decision criterion and resulting hit rate (HR) and false alarm rate (FAR). B: Receiver operating characteristic curve resulting from shifting the criterion in Fig. 1A.

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).

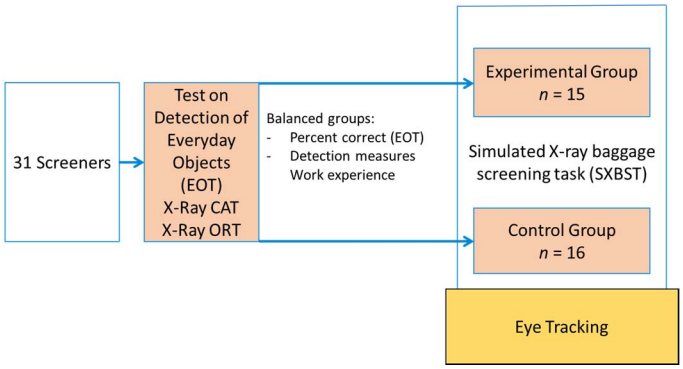

Fig. 2. Illustration of the study design.

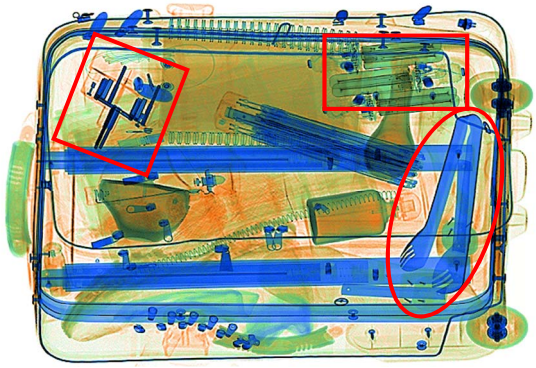

Fig. 3. Screenshot showing an X-ray image of a passenger bag from the everyday objects test with framed harmless everyday objects.

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]

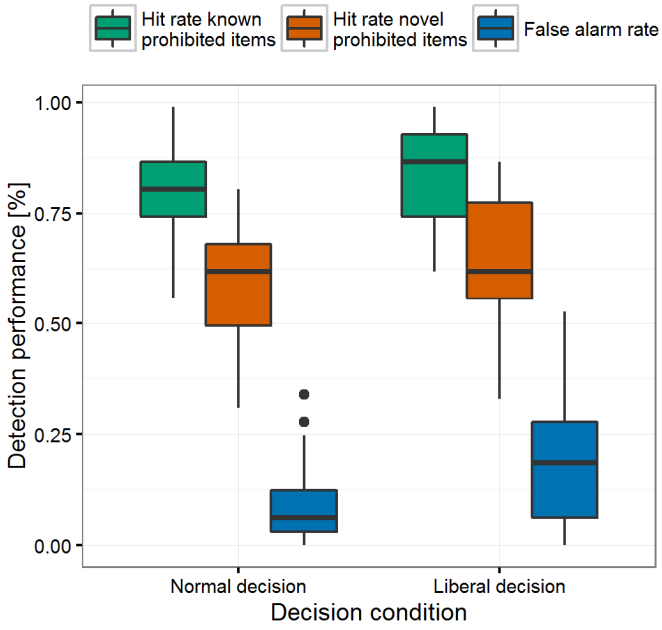

Fig. 4. Box plots of hit and false alarm rates depending on decision condition and prohibited item class (known vs. novel). (*Note: Performance values are multiplied by an arbitrary constant for security purposes.*)

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, $W = 38$, $p < .01$.

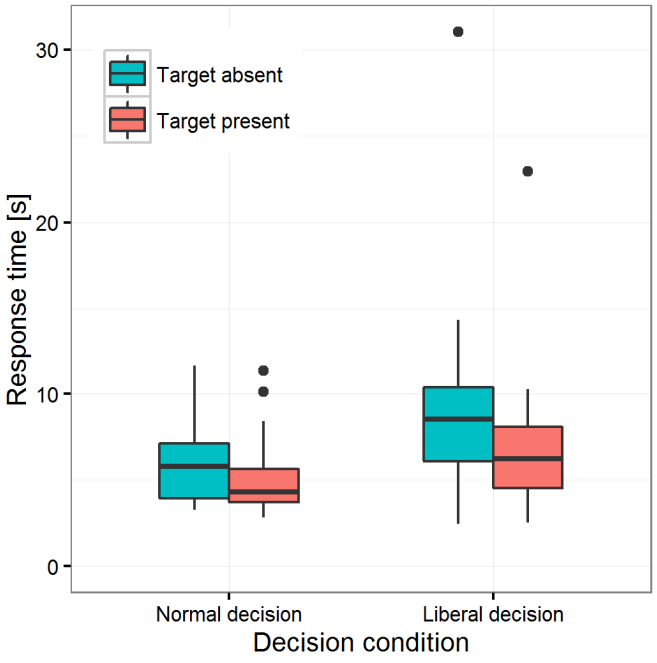

Fig. 5. Boxplot of individual median response times [s] depending on decision condition and separate for target absent and target present trials.

[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.

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.

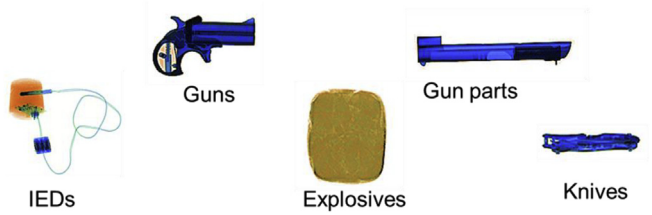

Fig. 1. Examples of the five threat categories.

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

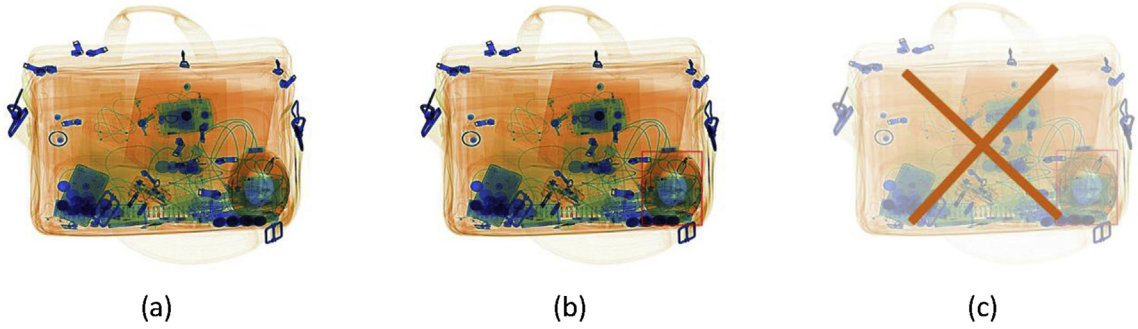

Fig. 2. Illustration of the three automation conditions: (a) baseline condition without automation, (b) OSAR, and (c) automated decision.

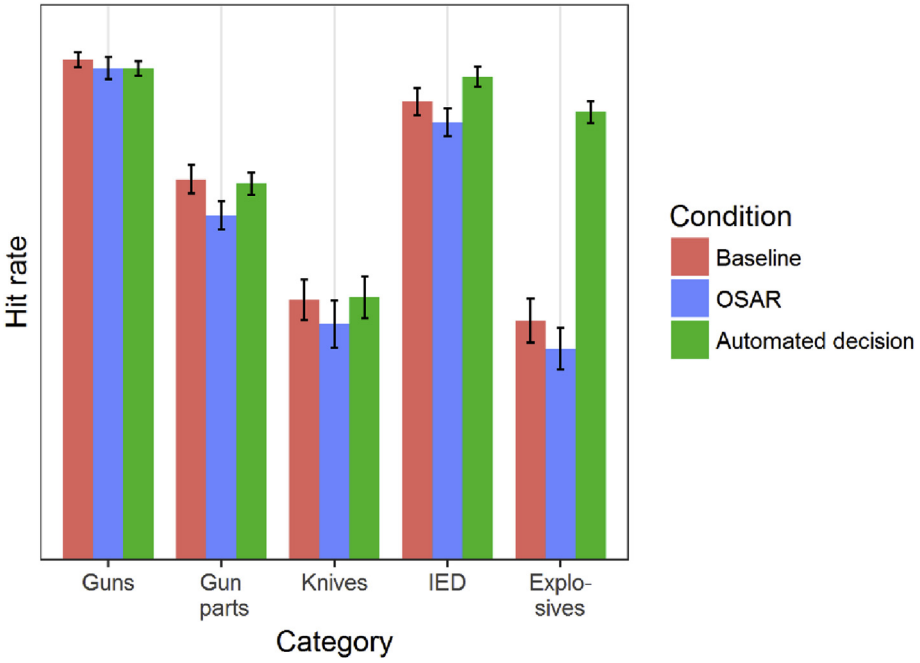

Fig. 3. Mean human–machine hit rates by condition (baseline, OSAR, automated decision) and prohibited item categories (guns, gun parts, knives, IEDs, and explosives). Absolute values of hit rate are not shown due to security restrictions in this project. Error bars are ± one standard error.

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.

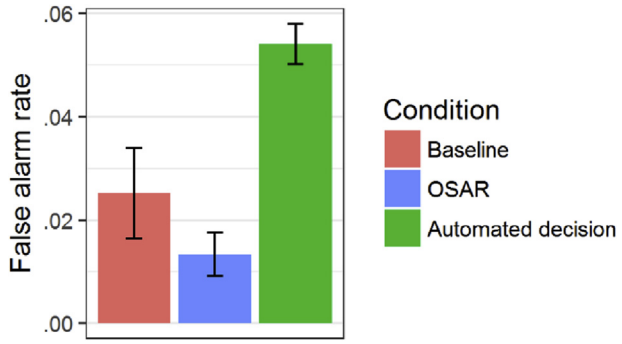

Fig. 4. Mean human–machine false alarm rates by condition (baseline, OSAR, and automated decision). Error bars are ± one standard error.

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

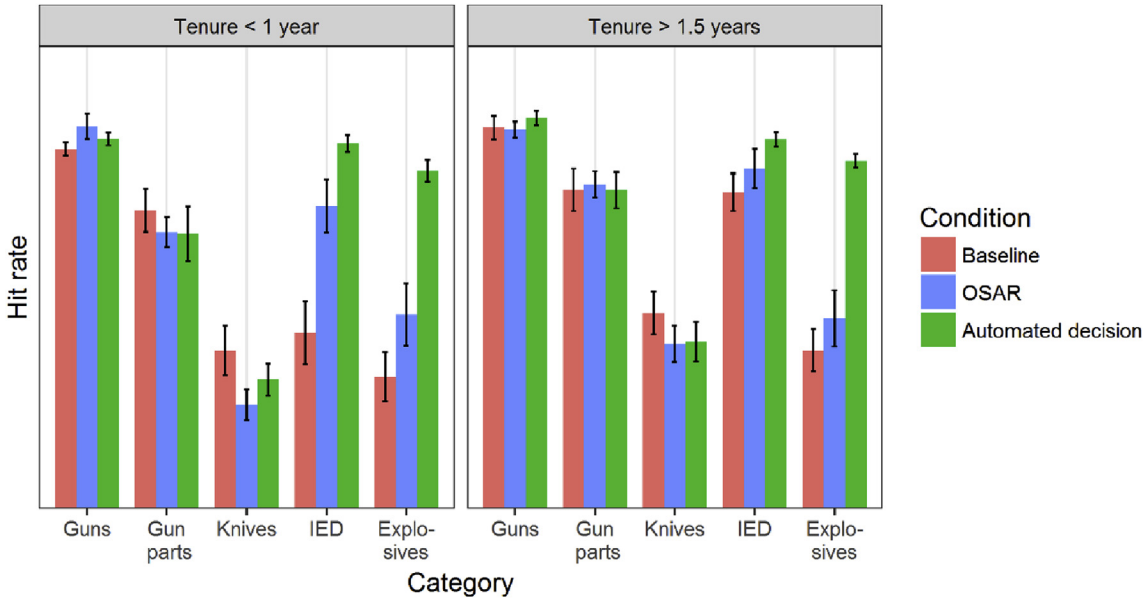

Fig. 5. Mean human–machine hit rates by condition (baseline, OSAR, automated decision), threat categories (gun, gun parts, and knives), and variation of work experience (tenure < 1 year and tenure > 1.5 years). Absolute hit rate values are not shown due to security restrictions in this project. Error 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.

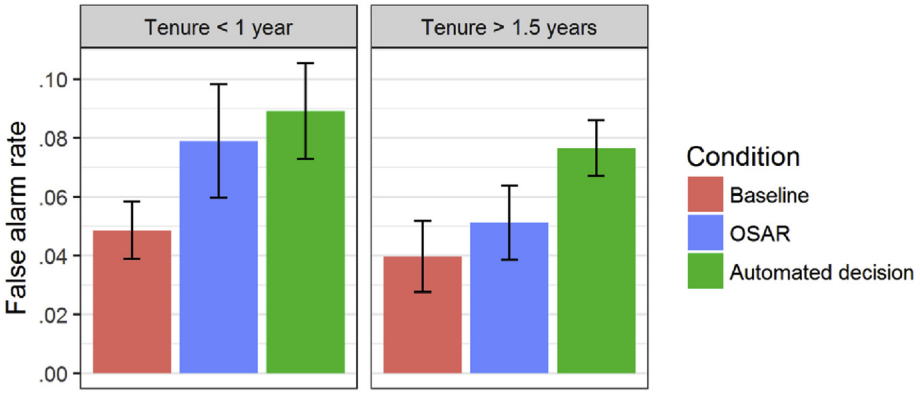

Fig. 6. Mean human–machine false alarm rates by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 1.5 years). Error bars are ± one standard error.

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

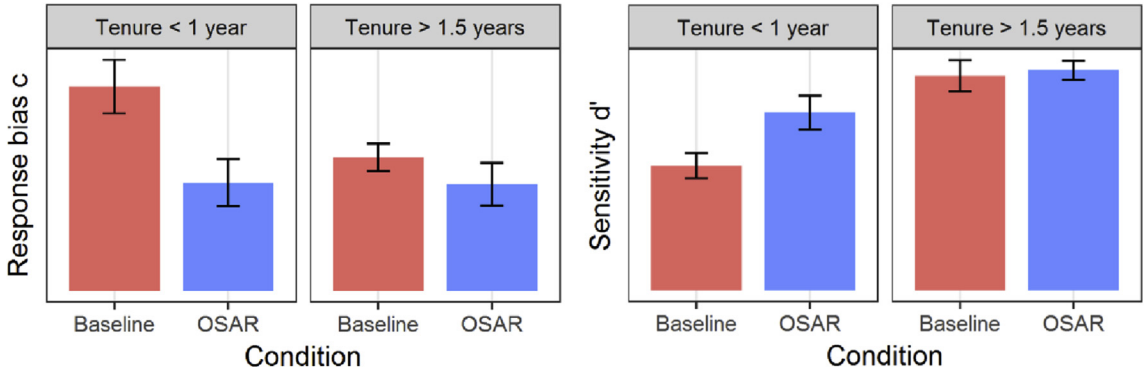

Fig. 7. (a) Mean response bias (*c*) of screeners by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 1.5 years). (b) Mean sensitivity measure *d*' of screeners by condition (baseline, OSAR, automated decision) and work experience (tenure < 1 year and tenure > 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.

Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging

Nicole Hättenschwiler, **Marcia Mendes**, and **Adrian Schwaninger**, University of Applied Sciences and Arts Northwestern Switzerland, Olten

Objective: This study compared the visual inspection performance of airport security officers (screeners) when screening hold baggage with state-of-the-art 3D versus older 2D imaging.

Background: 3D imaging based on computer tomography features better automated detection of explosives and higher baggage throughput than older 2D X-ray imaging technology. Nonetheless, some countries and airports hesitate to implement 3D systems due to their lower image quality and the concern that screeners will need extensive and specific training before they can be allowed to work with 3D imaging.

Method: Screeners working with 2D imaging (2D screeners) and screeners working with 3D imaging (3D screeners) conducted a simulated hold baggage screening task with both types of imaging. Differences in image quality of the imaging systems were assessed with the standard procedure for 2D imaging.

Results: Despite lower image quality, screeners' detection performance with 3D imaging was similar to that with 2D imaging. 3D screeners revealed higher detection performance with both types of imaging than 2D screeners.

Conclusion: Features of 3D imaging systems (3D image rotation and slicing) seem to compensate for lower image quality. Visual inspection competency acquired with one type of imaging seems to transfer to visual inspection with the other type of imaging.

Application: Replacing older 2D with newer 3D imaging systems can be recommended. 2D screeners do not need extensive and specific training to achieve comparable detection performance with 3D imaging. Current image quality standards for 2D imaging need revision before they can be applied to 3D imaging.

Keywords: human–automation interaction, visual search, graphical user interfaces (GUI), experience, transfer of training

Address correspondence to Nicole Hättenschwiler, School of Applied Psychology, Institute Humans in Complex Systems, University of Applied Sciences and Arts Northwestern Switzerland, Riggenbachstrasse 16, CH-4600 Olten, Switzerland; e-mail: nicole.haettenschwiler@fhnw.ch.

HUMAN FACTORS
Vol. XX, No. X, Month XXXX, pp. 1–17
DOI: 10.1177/0018720818799215

On December 21, 1988, Pan Am Flight 103 exploded over Lockerbie, Scotland, due to a bomb in a passenger bag transported in the hold of the aircraft (Strantz, 1990). Since then, many terrorist attacks have targeted airplanes (Baum, 2016; Singh & Singh, 2003). The most recent involving a bomb in hold baggage occurred on October 31, 2015, when Metrojet Flight 9268 was blown up during flight killing all 224 passengers (Baum, 2016). In response to such bomb threats, explosive detection systems (EDS) based on 2D imaging for hold baggage screening (HBS) were developed and introduced about 15 years ago (Caygill, Davis, & Higson, 2012; Harding, 2004; Singh & Singh, 2003). Such EDS-HBS assist airport security officers (screeners) who visually inspect X-ray images of passenger bags before they are loaded into the hold of an aircraft (Wells & Bradley, 2012). Newer 3D imaging technology uses computer tomography (CT). Technically, this has better automated explosive detection, higher baggage throughput, and 3D-rotatable images. Nonetheless, it also has lower image resolution and, therefore, poorer image quality than older 2D imaging technology (Flitton, Breckon, & Megherbi, 2013; Mouton & Breckon, 2015; Oftring, 2015; Wells & Bradley, 2012).

Human-machine system performance depends on technology and human factors. For instance, if lower image quality with 3D imaging would make it harder for screeners to decide whether a bag contains an improvised explosive device (IED), then 3D screening could in fact be inferior to 2D screening despite having better automated explosive detection. On the other hand, and this is a very important point to consider, if screeners would achieve at least similar detection performance with 3D imaging compared with 2D imaging, then the human-machine system as a whole would perform better with 3D imaging because this technology has better

automated explosive detection and higher baggage throughput. Investigating this issue is of major practical relevance: Although some countries introduced 3D imaging several years ago, other countries do not accept such technology due to their lower image quality compared with older 2D imaging EDS-HBS technology, even though 3D imaging has better automated explosive detection capability and higher baggage throughput (Flitton et al., 2013; Oftring, 2015). Moreover, there is a current debate on the international regulatory level regarding whether screeners working with 2D imaging need extensive and specific training before they can be allowed to work with 3D imaging technology. Our study addressed both issues by testing 2D and 3D screeners with 2D and 3D imaging in a simulated hold baggage screening task with the following research questions: (a) Can screeners achieve at least similar detection performance using 3D imaging compared with 2D imaging despite lower image resolution? (b) Does visual inspection competency acquired with one type of imaging transfer to the other type of imaging? These research questions are also interesting from a theoretical perspective—in particular, with regard to human-machine interaction, visual information processing and transfer of learning. Before discussing the relevant literature, it is important to clarify important terms and processes regarding the airport security screening of cabin and hold baggage.

Passengers store their carry-on bags in the cabin of airplanes. Because such cabin baggage can be accessed during flight, guns, knives, IEDs, and other items that could pose a threat (e.g., electric shock devices) are prohibited (Hancock & Hart, 2002; Harris, 2002; Schwaninger, 2005). As required by law (e.g., European Commission, 2015), screeners visually inspect every piece of cabin baggage at airport security checkpoints using X-ray machines. Larger baggage, in contrast, is stored in the hold of an aircraft and processed differently (Shanks & Bradley, 2004). Passengers have to register such hold baggage at check-in stations before going through airport security checkpoints. Hold baggage is then processed by a baggage handling system containing X-ray machines that have EDS-HBS (Level 1 of hold baggage screening) that highlights areas on the X-ray image that might contain explosive with colored rectangles (2D imaging systems) or by coloring the suspect area (3D imaging systems; see Figure 1 for illustrations). Whereas there are multiple target types (guns, knives, IEDs, explosives, other prohibited items) in cabin baggage screening, this is not the case in hold baggage screening. Because passengers cannot access items stored in the hold of an aircraft, guns or knives do not pose a threat, and hold baggage screening targets only fully functioning IEDs (Bretz, 2002). Only X-ray images of hold baggage on which an EDS-HBS has raised an alarm are sent to remote screening locations for on-screen alarm resolution by screeners (Level 2 of hold baggage screening). They visually inspect the X-ray images and decide whether the bag is harmless or contains a fully functioning IED with the following components: a triggering device, a power source, an explosive, and a detonator that need to be connected to each other by, for example, wires (Turner, 1994; Wells & Bradley, 2012). If screeners decide that an X-ray image is suspicious, more time-consuming investigations follow including rescreening with other X-ray technology, trace detection, explosive detection dogs, passenger reconciliation, and the opening of bags (Shanks & Bradley, 2004; Singh & Singh, 2003).

Since the terrorist attacks on September 11, 2001, there have been many studies on the visual inspection of X-ray images of cabin baggage, which consists of visual search and decision making (Koller, Drury, & Schwaninger, 2009; McCarley, Kramer, & Wickens, 2004; Wales, Anderson, Jones, Schwaninger, & Horne, 2009; Wolfe & Van Wert, 2010). Visual search challenges include low target prevalence, variations in target visibility, and the possible presence of multiple targets (Biggs & Mittroff, 2014; Clark, Cain, Adamo, & Mitroff, 2012; Godwin et al., 2010; Godwin, Menneer, Cave, Thaibsyah, & Donnelly, 2015; Mitroff, Biggs, & Cain, 2015). When it comes to decision making on whether a bag contains a prohibited item, screeners need to know which items are prohibited and what they look like in X-ray images (Schwaninger, 2005). Several studies have shown the importance of computer-based training in helping screeners to

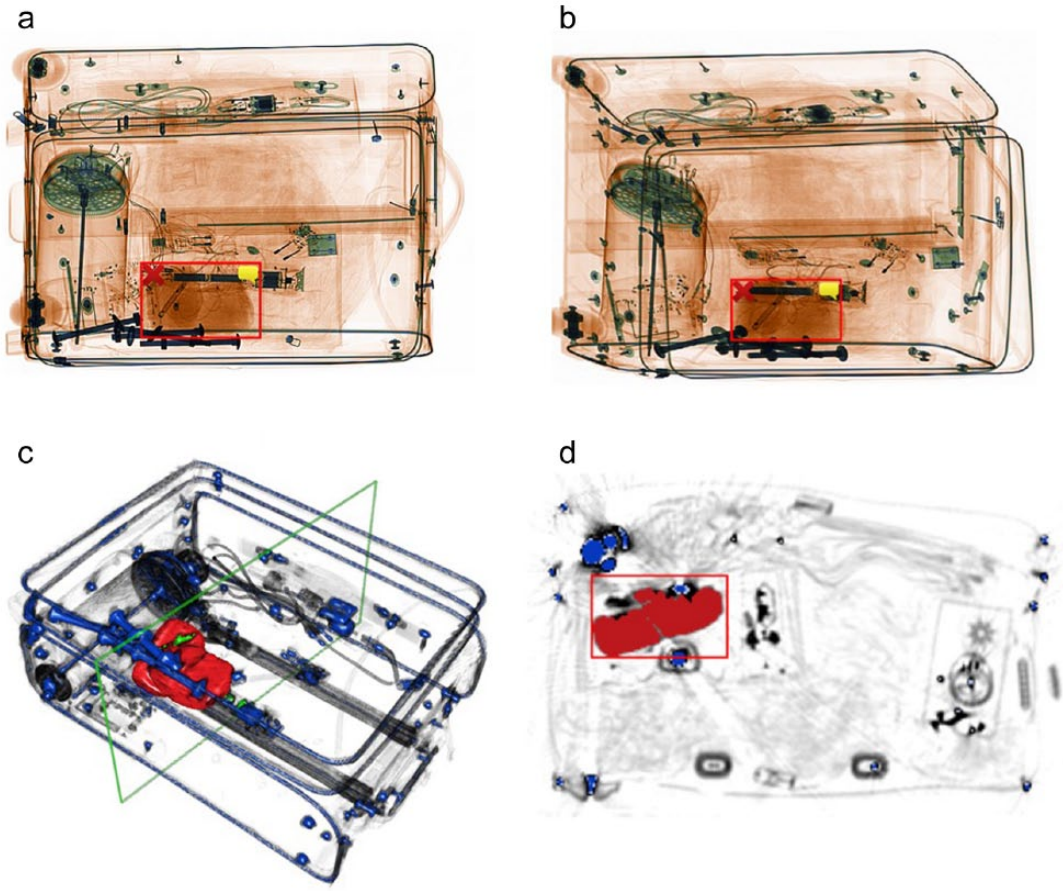

Figure 1. Target-present bag containing an IED recorded with a 2D multiview X-ray and a 3D CT imaging system currently used at airports: (a) 2D default image, (b) second 2D image with 30 degrees difference in perspective, (c) 3D-rotatable image, and (d) 3D-sliceable image. Explosive material is highlighted by the 2D imaging system with red rectangles (Figure 1a and 1b) and by the 3D imaging system with red coloring (Figure 1c and 1d). With 3D imaging, the detonator is visible in green (Figure 1c) and in blue (Figure 1d).

achieve and maintain high visual inspection performance (Fiore, Scielzo, Jentsch, & Howard, 2006; Halbherr, Schwaninger, Budgell, & Wales, 2013; Koller et al., 2009; Koller, Hardmeier, Michel, & Schwaninger, 2008; Schuster, Rivera, Sellers, Fiore, & Jentsch, 2013; Schwaninger & Hofer, 2004; Schwaninger, Hofer, & Wetter, 2007). International regulations take this into account by mandating initial and recurrent training of screeners. For example, European regulations mandate at least 6 hr of image recognition training and testing in every 6-month period for cabin- and hold-baggage screeners (European Commission, 2015).

Target prevalence in real-world baggage screening is about 2% because airports use threat image projection, a technology that projects X-ray images containing targets into the flow of images that are visually inspected by screeners (Hofer & Schwaninger, 2005; Schwaninger, 2006; Schwaninger et al., 2007; Schwaninger, Hardmeier, Riegelnig, & Martin, 2010). The challenge of low target prevalence in visual search refers to the finding that rare targets are frequently missed (Godwin et al., 2010; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013; Wolfe, Horowitz, & Kenner, 2005). This is consistent with signal detection theory (SDT, Green &

Swets, 1966) according to which the probability of signal occurrence (target prevalence) influences the probability of responding that a signal (target) is present. Using the SDT framework, the target prevalence effect can be explained as a shift in response bias (Fleck & Mitroff, 2007; Godwin et al., 2010; Lau & Huang, 2010; Wolfe et al., 2007; Wolfe & van Wert, 2010). SDT provides a measure of detection performance (d′) that is independent of response bias (and therefore also of target prevalence). This has been confirmed for different domains and tasks (Green & Swets, 1966; MacMillan & Creelman, 2005; Swets, 1996) including X-ray image inspection and visual search (Meneer, Donnelly, Godwin, & Cave, 2010; Verghese, 2001; Wolfe & Reynolds, 2008; Wolfe & Van Wert, 2010). Moreover, Schwaninger, Hofer, and Wetter (2007) found very similar detection performance (d′) in screeners when performing a computer-based test with a target prevalence of 50% compared with detection performance (d′) measured on the job using threat image projection data with a target prevalence of 2%.

Regarding target visibility, studies have shown how image-based factors impact on visual inspection performance (e.g., Bolfing, Halbherr, & Schwaninger, 2008; Schwaninger, Hardmeier, & Hofer, 2005; Schwaninger, Michel, & Bolfing, 2005, 2007). For example, objects depicted from unusual viewpoints are more difficult to recognize (effect of viewpoint). Moreover, in X-ray images, objects appear with overlay, and detecting prohibited items depends on how much they are superimposed by other objects (effect of superposition). Finally, prohibited items are more difficult to recognize in complex bags containing many other items and clutter (effect of bag complexity). These challenges can be reduced with 2D imaging that displays a passenger bag as two X-ray images from different perspectives (dual-view imaging). However, previous studies on cabin baggage screening have shown that although dual-view imaging leads to higher detection performance than single-view X-ray imaging, it also increases response time (von Bastian, Schwaninger, & Michel, 2008; Franzel, Schmidt, & Roth, 2012). Similar results have been found for motion imaging in which bags are displayed as an animated sequence of X-ray images depicting a bag from different viewpoints (Mendes, Schwaninger, & Michel, 2013).

Several years ago, advanced CT technology, which has been implemented highly successfully in medical imaging (Barrat, 2000), became available for hold baggage screening (Mouton & Breckon, 2015; Wetter, 2013). Compared with the older 2D imaging technology used in HBS, state-of-the art CT scanners feature better automated explosive detection, slicing, and 3D-rotatable images (Flitton et al., 2013; Mouton & Breckon, 2015; Oftring, 2015; Wells & Bradley, 2012). Slicing refers to the production of cross-sectional images or "slices" of a bag. From a series of image slices, a bag can be reconstructed as a 3D CT volume image and the bag can be displayed as a 3D-rotatable and 3D-sliceable image (Flitton, Breckon, & Megherbi, 2010, 2013). This could result in better detection performance among screeners for two reasons: First, it might be easier to recognize the different components of an IED that, in certain 2D views, would be displayed from a difficult viewpoint and/or superimposed by other items in a complex bag (Bolfing et al. 2008; Schwaninger, Michel, & Bolfing, 2005, 2007). Second, object recognition research has shown that exposure to 3D images results in richer visual object representations (Tarr & Vuong, 2002; Vuong & Tarr, 2004). This could improve screeners' detection performance not only in 3D but also in 2D images. On the other hand, CT systems have lower image resolution and therefore lower image quality compared with EDS-HBS 2D imaging (Flitton et al., 2010, 2013; Mouton & Breckon, 2015), and this could impair screeners' detection performance with 3D imaging. Regarding response times (RT), screeners might take more time to visually inspect 3D images because rotating X-ray images and slicing both require additional time.

This study extends previous research on cabin baggage screening by addressing questions of high practical and theoretical relevance for hold baggage screening. We wanted to know (a) whether screeners using 3D imaging can achieve at least similar detection performance to that when using 2D imaging despite lower image

TABLE 1: Description of Screeners Participating in the Study

Participants	*n*	% female	Age	Work experience with 2D imaging (months)	Work experience with 3D imaging (months)
2D Screeners	42	61%	*M* = 44.90 *SD* = 10.36	*M* = 138.31 *SD* = 78.35	
3D Screeners	42	35%	*M* = 36.76 *SD* = 9.22	*M* = 86.02 *SD* = 64.42	*M* = 19.12 *SD* = 5.07

resolution, and (b) whether the visual inspection competency acquired with one type of imaging transfers to the other type of imaging. We addressed these research questions by asking two screener groups that differed in their experience in working with the two imaging technologies to perform a simulated hold baggage screening task with both 2D and 3D imaging. In order to achieve high external validity, we used X-ray images that were recorded with 2D and 3D imaging systems that are currently operational at airports. It is important to note that the reason for comparing 2D and 3D imaging differing in image quality is that the two types of imaging tested in this study are from real-world systems; there is therefore a need to know whether 3D screening results in better human-machine system performance despite lower image quality.

Our main dependent variable was detection performance (d′), which has high external validity for real-world baggage screening because it is independent of target prevalence. Due to the fact that airports use threat image projection with a target prevalence of about 2% (Hofer & Schwaninger, 2005; Schwaninger, Hofer, & Wetter, 2007), target-absent RT were also important, because they account for about 98% of X-ray images in real-world hold baggage screening. Based on our results, we shall discuss whether replacing older 2D with newer 3D imaging technology improves the human-machine system performance in terms of efficiency and effectiveness of the hold baggage screening process as a whole. In addition, our results have important implications in light of current international discussions on whether extensive and specific training should be mandated for 2D screeners before allowing them to work with 3D imaging technology.

METHOD

Participants

Participants were professional hold baggage screeners from two international airports (see Table 1 for details). All screeners had been 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). Eighty-eight screeners consented to participate in the study (43 2D screeners and 45 3D screeners). Three screeners (one 2D and two 3D) who could not attend the main test due to illness were excluded. One further 3D screener had to be excluded due to a malfunction of the simulator. This left a total of 84 screeners (42 2D screeners [21 tested with 2D imaging and 21 tested with 3D imaging] and 42 3D screeners [23 tested with 2D imaging and 19 tested with 3D imaging]). 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.

Design

All participants attended the airport test facilities twice. First, they completed a pretest to familiarize themselves with the 2D and 3D simulators and the testing procedure. For the main test 2 weeks later, screeners were randomly assigned to be tested with either 2D or 3D imaging. The experiment (main test) used a between-subjects design with X-ray imaging technology (2D vs. 3D imaging) and screener group (3D vs. 2D screeners) as independent variables

and visual inspection performance measures as dependent variables (detection performance [d′], target-absent RT, and target-present RT).

Materials

Aviation security experts from a specialized police organization running one of the test centers responsible for airport security equipment testing and certification in Europe created 64 different IEDs (32 for the pretest and 32 for the main test, IEDs were randomly assigned to be used in the pretest or the main test). X-ray images of hold baggage were recorded at this test center by five aviation security experts and the first and second author using 2D multiview X-ray and 3D CT imaging systems that are currently being used at airports (see Figure 1 for examples of images and further information).

Thirty-two different bags were used repeatedly by repacking them to create unique stimuli for the pretest and the main test. All bags were of medium complexity as defined by the aviation security experts. Target-present images contained one IED. Target-absent images contained EDS-HBS false alarms (e.g., cheese, certain liquids, etc.). To ensure that the 3D imaging condition had the same system reliability (e.g., Rice & McCarley, 2011) as the 2D imaging condition, we used EDS-HBS alarms from 3D imaging as a reference when setting red frames manually around the same objects of interest in 2D imaging stimuli.

The pretest consisted of 64 2D X-ray images and 64 3D CT images of different bags. Target prevalence was 50%. Each IED was used twice in different bags: once recorded from a more frontal perspective displaying more surface area, and once from a horizontally or vertically rotated perspective using medium superposition. The main test consisted of 256 bags that were recorded with 2D and 3D imaging. Target prevalence was 50%. Each of the 32 IEDs was used four times in four different bags by varying viewpoint and superposition.

As described in the introduction, 3D imaging systems have lower image quality than 2D imaging systems. To assess such differences, we used the standard test piece (STP) and protocol, which is currently the most widely used international standard for the assessment of image quality of 2D imaging systems (see the Appendix for details).

Procedure

Tests were conducted without giving performance feedback using simulators provided by the manufacturer of the 2D and 3D imaging systems. Six computer workstations with 19” TFT monitors were set up in a normally lit room. Each screener sat approximately 50 cm away from the monitor. The X-ray images covered about two thirds of the screen. Four to six participants performed the test in each session while working individually, quietly, and under supervision. This is a typical scenario in hold baggage screening (Kuhn, 2017). Screeners received instructions before the start of each test informing them about the imaging systems, the number of images, and that the target items were IEDs. To prevent a criterion shift (change of response bias) during the experiment, we informed the screeners beforehand about the target prevalence in the experiment (see also McCarley, 2009; Rich et al., 2008).

Screeners were instructed to visually inspect each X-ray image as if they were working at the airport and to decide as accurately and quickly as possible whether or not the image contained a target by clicking on a target-present or a target-absent button on the simulator interface (a yes–no task in signal detection theory; see MacMillan & Creelman, 2005). After receiving their instructions, all participants started the experiment with 10 practice trials (5 target-absent and 5 target-present images in random order). A time limit of 90 s was set for viewing an X-ray image; afterwards, the image disappeared, but the screeners still had to make a decision.

European regulations mandate that screeners have to take a break of at least 10 min after 20 min of continuous visual inspection of X-ray images (European Commission, 2015). Therefore, tests were divided into four blocks, and screeners were asked to take breaks of 10 to 15 min after completing each block. Block order was counterbalanced across participants. Images appeared in random order within a block. All participants completed the pretest in less than 40 min and the main test in less than 1.5 hr including breaks.

TABLE 2: Definition of Hit, False Alarm, Miss, and Correct Rejection According to SDT (Green & Swets, 1966)

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

Note. SDT = signal detection theory (Green & Swets, 1966).

Analyses

We computed analyses of covariance (ANCOVA) with detection performance (d′), target-absent RT, and target-present RT as dependent variables and age and 2D work experience as covariates (using SPSS version 22 and an alpha level of .05). Age was used as covariate because 3D screeners were, on average, younger than 2D screeners (see Table 1) and because previous research showed a negative correlation between age and the visual inspection performance of screeners (Ghylin, Drury & Schwaninger, 2006; Schwaninger et al., 2010). 2D work experience was used as covariate because 2D screeners had on average more 2D work experience than 3D screeners (see Table 1). We conducted post hoc comparisons with R version 3.22 (R Core Team, 2015) and applied Holm–Bonferroni corrections (Holm, 1979). We report ANCOVA effect sizes with η_p^2; effect sizes of *t* tests, with Cohen's *d*.

According to SDT (Green & Swets, 1966), there are four possible outcomes depending on stimuli and participant responses (Table 2). Detection performance (d′) was calculated using the following SDT formulae, whereby *z* refers to the inverse of the cumulative distribution function of the standard normal distribution (Green & Swets, 1966; MacMillan & Creelman, 2005):

$$Hit\ Rate\left(HR\right) = Hits / \left(Hits + Misses\right) \quad (1)$$

$$False\ Alarm\ rate\left(FAR\right) = False\ Alarms / \left(False\ Alarms + Correct\ Rejections\right) \quad (2)$$

$$\mathrm{d}' = z\left(HR\right) - z\left(FAR\right) \quad (3)$$

RESULTS

Image Quality

Detailed results on image quality assessment with six tests of the STP are reported in the Appendix. In summary, results confirmed that the 2D imaging system passed all image quality tests. The 3D imaging system did not pass two of the six tests: The spatial resolution and useful penetration tests could not be solved using either the 3D-rotatable or the 3D-sliceable image. Nonetheless, taking all test results into account, it should be possible to recognize main IED components (triggering devices, power sources, explosives, and detonators) to a similar degree with 2D and 3D imaging. However, recognizing thin wires when they are hidden behind aluminum of a thickness of 7.9 mm or more was not possible with 3D imaging.

Visual Inspection Performance

We first present the results on detection performance (d′) because this is the main dependent variable for addressing our research questions. We then present RT, whereby target-absent RTs are more important due to the fact that they account for about 98% of all X-ray images in real-world hold baggage screening when using threat image projection (Hofer & Schwaninger, 2005; Schwaninger, 2006; Schwaninger, Hofer, & Wetter, 2007; Schwaninger et al., 2010). Figure 2 shows detection performance d′ depending on X-ray imaging technology (2D vs. 3D imaging) and screener group (2D vs. 3D screeners).

A 2 (2D vs. 3D imaging) × 2 (2D vs. 3D screeners) ANCOVA with d′ as dependent variable while controlling for age and 2D work experience revealed a trend toward better detection performance (d′) with 3D imaging (mean values of main effect: 2D imaging d′ = 1.80; 3D

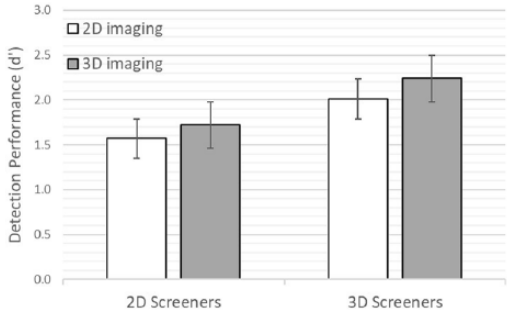

Figure 2. Detection performance (d′) by X-ray imaging technology (2D vs. 3D imaging) and screener group (2D vs. 3D screeners). Error bars are ± one standard error.

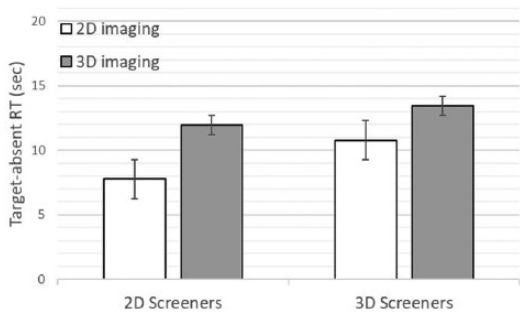

Figure 3. Target-absent RT by X-ray imaging technology (2D vs. 3D imaging) and screener group (3D vs. 2D screeners). Error bars are ± one standard error.

imaging d′ = 1.97). However, this effect did not attain statistical significance, $F(1, 78) = 3.56$, $p = .065$, $\eta_p^2 = .04$. There was a significant effect of screener group with 3D screeners performing better with both types of imaging than 2D screeners (mean values of main effect: 2D screeners d′ = 1.72; 3D screeners d′ = 2.05), $F(1, 78) = 10.18$, $p = .002$, $\eta_p^2 = .12$. The interaction between imaging and screener group was not significant. There was a significant effect of the covariate age, $F(1, 79) = 2.86$, $p < .001$, $\eta_p^2 = .16$ but not of the covariate 2D work experience.

Figure 3 shows target-absent RT by X-ray imaging technology and screener group.

We calculated a 2 (2D vs. 3D imaging) × 2 (2D vs. 3D screeners) ANCOVA for target-absent trials with RT as dependent variable while controlling for age and 2D work experience. We found a main effect of imaging $F(1, 78) = 12.12$, $p < .001$, $\eta_p^2 = .13$, and screener group, $F(1, 78) = 11.22$, $p < .001$, $\eta_p^2 = .13$, but no significant effect for their interaction. Further, there was a significant effect of the covariate age, $F(1, 78) = 7.75$, $p = .007$, $\eta_p^2 = .09$, but not of the covariate 2D work experience. To examine whether speed–accuracy trade-offs can explain why 3D screeners had higher detection performance (d′) than 2D screeners with both imaging systems, we used two-tailed independent samples *t* tests to examine accuracy in target-absent trials (percent correct rejections, PCR). For 2D imaging, 2D screeners had significantly higher PCR than 3D screeners, $t(42) = -3.88$, $p < .001$. For 3D imaging, we did not find a difference between the screener groups for PCR, $t(38) = -.00$, $p = .997$. This means that we found no evidence that the better detection performance (d′) of 3D screeners compared with 2D screeners could be explained by a speed–accuracy trade-off in target-absent trials.

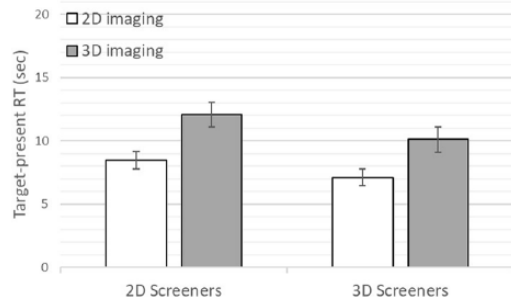

Figure 4. Target-present RT by X-ray imaging technology (2D vs. 3D imaging) and screener group (2D vs. 3D screeners). Error bars are ± one standard error.

Figure 4 shows target-present RT dependent on X-ray imaging technology and screener group.

We calculated a 2 (2D vs. 3D imaging) × 2 (2D vs. 3D screeners) ANCOVA for target-present trials with RT as dependent variable

TABLE 3: Correlations Between Speed (Target Absent and Target Present RT) and Detection Performance (d′) Controlling for Age and Work Experience

	3D Screener Group		2D Screener Group	
Trial type	2D imaging	3D imaging	2D imaging	3D imaging
Target-present trials	$r = –.01$	$r = –.33$	$r = .03$	$r = .22$
	$p = .97$	$p = .10$	$p = .92$	$p = .36$
Target-absent trails	$r = .28$	$r = .25$	$r = .16$	$r = .19$
	$p = .20$	$p = .31$	$p = .49$	$p = .40$

while controlling for age and 2D work experience. We found a main effect of imaging, $F(1, 78) = 20.32$, $p < .001$, $\eta_p^2 = .21$, and a significant effect of the covariate age, $F(1, 78) = 25.52$, $p < .001$, $\eta_p^2 = .25$ but not of the covariate 2D work experience. Neither the main effect of screener group nor the interaction was significant, making a speed–accuracy trade-off an implausible explanation for the better detection performance (d′) of 3D screeners compared with 2D screeners.

To examine speed-accuracy trade-offs within the screener groups, we also calculated two-tailed partial correlations between response times and detection performance (d′) while controlling for age and work experience (Table 3). A speed-accuracy trade-off would have been supported if at least one significant positive correlation (longer reaction times and higher detection performance [d′]) would have been found. This was not the case, which makes speed-accuracy trade-offs very unlikely.

DISCUSSION

This study addressed two questions of high practical and theoretical relevance for the airport security screening of hold baggage: (a) Can screeners achieve at least similar detection performance using 3D imaging compared with 2D imaging despite the lower image quality of 3D imaging? (b) Does visual inspection competency acquired with one type of imaging transfer to the other type of imaging? We addressed these questions by asking 2D screeners and 3D screeners to perform a simulated hold baggage screening task with both types of imaging. We first discuss the results on detection performance (d′), the main dependent variable for our research questions. We then discuss the results on response times (RT) whereby target-absent RTs are more meaningful for real-world baggage screening. We conclude by discussing implications of our results for the efficiency and effectiveness of hold baggage screening using 2D versus 3D imaging systems.

Despite lower image quality (see the Appendix for these results and their discussion), 3D imaging resulted in a similar detection performance (d′) of screeners compared with that for 2D imaging. Benefits of 3D imaging allowing three-dimensional rotation and slicing seem to compensate for the potentially negative effects of lower image quality. This is consistent with earlier research on cabin baggage screening that showed better detection performance for motion imaging compared with static 2D imaging (Mendes et al., 2013). 2D screeners achieved a similar detection performance (d′) with 3D imaging to that with 2D imaging. This indicates a very large transfer effect and has important practical implications in light of the current international discussions on whether specific training should be mandated for 2D screeners before allowing them to work with 3D imaging systems. Our results suggest that 2D screeners do not need extensive and specific training to achieve similar detection performance with 3D imaging compared with that attained with 2D imaging.

3D screeners also achieved similar detection performance (d′) with both imaging systems, but they performed better than 2D screeners with both types of imaging. As explained in the introduction, object recognition research has shown that exposure to 3D images results in richer visual representations that could therefore also increase detection performance in 2D images

(Tarr & Vuong, 2002; Vuong & Tarr, 2004). This is a plausible explanation for our finding that 3D screeners performed better than 2D screeners not only with 3D imaging but also with 2D imaging. Alternative explanations might be based on group differences in age, cognitive abilities, training, or work experience along with speed–accuracy trade-offs. Because we used age as covariate, age differences are an unlikely explanation for performance differences between 2D and 3D screeners. Visual-cognitive abilities have also been shown to impact on screener performance (Hardmeier & Schwaninger, 2008; Rusconi, Ferri, Viding, & Mitchener-Nissen, 2015; Rusconi, McCrory, & Viding, 2012; Schwaninger, Hardmeier, & Hofer, 2005). However, it is also unlikely that differences in these abilities can explain the detection performance differences between 3D and 2D screeners in our study. The organization providing the 2D screeners had implemented a very selective preemployment screening procedure including a visual-cognitive test battery and an X-ray object recognition test (Hardmeier, Hofer, & Schwaninger, 2006; Schwaninger, Hardmeier, & Hofer, 2005). Moreover, it is difficult to explain differences between 2D and 3D screeners by amount of training because both screener groups were qualified, trained, and certified according to the same European standards including a 6-hr mandatory image recognition training and testing every 6 months (European Commission, 2015). Finally, differences in 2D work experience cannot explain why 3D screeners were better with 2D imaging than 2D screeners, because the latter had more work experience with 2D imaging, and 2D work experience was used as covariate. Thus, the most plausible explanation based on results from object-recognition research (Tarr & Vuong, 2002; Vuong & Tarr, 2004) would seem to be that extensive exposure to 3D imaging during work and training resulted in richer visual representations and therefore better performance of 3D screeners than 2D screeners for both types of imaging.

The target-absent RT of 2D screeners when using 2D imaging was 8 s. Threat image projection data from experienced 2D screeners working with a similar 2D imaging system revealed target-absent RTs of about 7 s (Schwaninger, Hofer, & Wetter, 2007). This suggests that the target-absent RT found in our study would generalize quite well to real-world conditions (at least for 2D screeners when using 2D imaging) despite large differences in target prevalence. Both screener groups needed more time (about 2 s [3D screeners] and 4 s [2D screeners]) when using 3D imaging compared with 2D imaging. This result was anticipated, because rotating and slicing 3D images takes longer to process than the visual inspection of static 2D X-ray images. When no target was present, 3D screeners took longer for visual inspection than 2D screeners. For 3D imaging, the difference was small (about 1 s). For 2D imaging, 3D screeners took 3 s longer than 2D screeners. Although speculative, one possible explanation could be that 3D screeners were used to rotating and slicing images but were unable to do this when using 2D imaging. This may have resulted in longer target-absent RT. However, the important result is that the higher detection performance (d′) of 3D screeners with both imaging systems compared with 2D screeners could not be explained by a speed–accuracy trade-off.

The target-present RT of 2D screeners when using 2D imaging was 8 s. This was similar to the real-world target-present RT of 9 s for experienced 2D screeners when using 2D imaging for hold baggage screening (Schwaninger, Hofer, & Wetter, 2007). This provides further support for the view that the RT found in our study would generalize to real-world conditions despite large differences in target prevalence. As for target-absent RT, both screener groups needed more time: 3 s (3D screeners) and 4 s (2D screeners) when using 3D imaging. Differences between screener groups were not significant for target-present RT, making a speed–accuracy trade-off an extremely implausible explanation for the better detection performance (d′) of 3D screeners compared with 2D screeners with both imaging systems.

Whereas 2D work experience did not have an impact, age had an influence on all dependent variables: Older screeners had lower detection performance (d′) and longer response times. This result is consistent with previous research showing a negative correlation between age and visual inspection performance of screeners

TABLE 4: Estimation of Efficiency Increase (Throughput) When Using 3D Imaging Compared With 2D Imaging Based on Target-Absent RT Results

Scenario	Bags per hour	EDS-HBS FAR	Approval capacity	Efficiency increase Level 1	Bags sent to visual inspection	Target absent RT [sec]	Visual inspection time [hr]	Efficiency increase Level 2
2D screeners / 2D imaging	1,500	35%	975		525	8	1.2	
2D screeners / 3D imaging	1,500	15%	1,275	31%	225	12	0.8	36%
3D screeners / 2D imaging	1,500	35%	975		525	11	1.6	
3D screeners / 3D imaging	1,500	15%	1,275	31%	225	13	0.8	49%

Note. EDS = explosive detection systems; HBS = hold baggage screening; FAR = false alarm rate; RT = response time.

(Ghylin et al., 2006; Schwaninger et al., 2010). Because we used 2D work experience and age as covariates, the observed screener group differences in detection performance (d′) and response times cannot be explained by preexisting differences in the covariates.

To summarize, the results on detection performance (d′) answered our two research questions: (a) Screeners achieved a similar detection performance (d′) using 3D imaging compared with 2D imaging despite lower image resolution of 3D imaging. (b) Visual inspection competency acquired with one type of imaging transferred to visual inspection with the other type of imaging. However, both screener groups needed more time (2–4 s) when using 3D imaging compared with 2D imaging.

What do our results on screeners' visual inspection performance mean for the efficiency (throughput) of 2D versus 3D hold baggage screening at airports? According to Oftring (2015), 2D and 3D imaging systems can process about 1,500 bags per hour, but 2D imaging systems have false alarm rates of at least 35%, whereas 3D imaging systems achieve much lower false alarm rates (15%). The installation of 3D imaging (Level 1 in hold baggage screening) should therefore already result in a 31% increase in efficiency. Based on the amount of bags sent to visual inspection and the target-absent RTs found in our study, an efficiency increase from 36% to 49% on Level 2 of hold baggage screening (alarm resolution of screeners) can be achieved (see Table 4 for the calculation). As explained in the introduction, if screeners decide that an X-ray image is suspicious, more time-consuming investigations follow including rescreening with other X-ray technology, trace detection, explosive detection, dogs, passenger reconciliation, and the opening of bags (Shanks & Bradley, 2004; Singh & Singh, 2003). Therefore, efficiency gains will be even higher in practice because 3D imaging results in less hold baggage being sent to Level 2.

Estimating the increase in effectiveness (detection of IEDs) is more difficult, because the detection rates of 2D and 3D imaging systems are not publicly available for security reasons. However, it is clear that 3D imaging systems achieve substantially higher detection of explosives than 2D imaging systems (e.g., Oftring, 2015; Singh & Singh, 2003; Wells & Bradley, 2012). Moreover, in Europe, EDS-HBS have to meet European detection standards and be approved by test centers of the European Civil Aviation Conference (ECAC). So far, only 3D imaging systems have met ECAC Standard 3, whereas 2D imaging systems achieve only Standard 2 (European Civil Aviation Conference, 2018). ECAC Standard 3 requires higher hit rates and lower false alarm rates and therefore higher detection performance (d′) of EDS-HBS.

Taking together the results of our study on screeners' visual inspection performance with the performance advantages of 3D imaging technology, it is reasonable to infer that the whole human-machine system performance when using 3D imaging technology is superior to 2D imaging not only in terms of efficiency (throughput) but also in terms of effectiveness (detection of IEDs) of the HBS process as a whole. The results of our study further suggest that extensive and specific training is not needed for 2D screeners before allowing them to work with 3D imaging systems. Nonetheless, some limitations do call for further research: Screener performance was tested with only one 2D and 3D imaging system. It would be interesting to see whether different results would be obtained with other 2D systems using a larger angular difference between the two views of a bag (e.g., 60–90 deg), with 3D systems that have higher image resolution, and with hybrid systems that show four views (3D-rotatable, 3D sliceable, and two different STP-compliant 2D views). Although it is not possible to conclude from our study that higher image resolution of 3D imaging systems would result in better visual inspection performance among screeners, it would be worth investigating this in future studies. Second, it would be interesting to see whether the results of our study can be replicated with screeners from other airports using a within-subjects design to investigate transfer effects from 2D to 3D imaging and vice versa over several months (although this might be rather difficult to achieve in practice). Conducting such a study with student participants is not an option for reasons of external validity as well as the security-sensitive nature of the image material and on-screen alarm resolution protocols.

Despite these limitations, we believe that our study is robust enough to make a significant contribution to the theory, practice, and knowledge base of human factors and ergonomics—particularly with regard to its practical relevance. First, we can recommend a wide-scale implementation of 3D imaging systems with an image quality equal to or higher than that of the 3D imaging system tested in this study, because it can be expected to result in better human–machine system performance in terms of efficiency and effectiveness of the hold baggage screening process as a whole. Second, due to large transfer effects, 2D screeners do not require extensive and specific training to achieve similar detection performance with state-of-the-art 3D imaging. Third, image quality standards and procedures need revision before they can be applied to 3D imaging systems.

APPENDIX

Assessment of Image Quality

The most widely used international standard for assessing the image quality of X-ray imaging systems is the standard test piece (STP) and a procedure developed about 30 years ago for 2D imaging systems (WG Standard Test Piece, n.d.). Whereas 2D imaging systems comply with the STP standard, this is not the case for many 3D imaging systems. This is not surprising given the fact that the STP was developed for 2D imaging and that there is no specific image quality assessment procedure available yet for 3D imaging. Some countries have implemented 3D imaging systems at many of their airports (Oftring, 2015) because they have better automated explosive detection capability and higher baggage throughput (Flitton et al., 2013; Mouton & Breckon, 2015; Wells & Bradley, 2012). Other countries hesitate to change from 2D imaging to 3D imaging because it is unclear whether screeners achieve a similar detection performance using 3D imaging compared with 2D imaging due to the lower image quality of 3D imaging.

Our study evaluated screener performance using 2D and 3D imaging while also evaluating image quality differences between the two imaging systems using the STP. We used the STP to provide a quantitative measure of the differences of the 2D and 3D imaging systems tested in this study. We first explained the STP and the image quality assessment procedure and then presented the results for the 2D and 3D imaging systems used in our study. Based on the results, we discussed whether current image quality standards for 2D imaging need to be revised before they can be applied to 3D imaging systems.

The STP contains samples of materials of varying density and needs to be X-rayed with the tested machine. Based on the X-ray image of the STP, six tests are carried out to assess single wire resolution, useful penetration, spatial resolution, simple penetration, and material

TABLE A1: Results of Image Quality Tests for the 2D and 3D Imaging Systems Used in This Study

Test	Requirement	3D imaging system		
		3D-rotatable image	3D-sliceable image	2D imaging system
Test 1 STP: Single wire resolution	Ability to display a single thin wire (30 American Wire Gauge = 0.254 mm) when not covered by the aluminum step wedge.	Yes	Yes	Yes
Test 2 STP: Useful penetration	Wire (24 American Wire Gauge = 0.5105 mm) needs to be visible behind different thickness of aluminum (4.8 mm, 7.9 mm, and 11.1 mm).	4.8mm: No 7.9mm: No 11.1mm: No	4.8mm: Yes 7.9mm: No 11.1mm: No	4.8 mm: Yes 7.9 mm: Yes 11.1 mm: Yes
Test 3 STP: Spatial resolution	Ability to distinguish and display objects that are close together; gaps between the relevant vertical and horizontal gratings can be seen (2.0 mm slots on a 4.0 mm pitch).	No	No	Yes
Tests 4 & 5 STP: Simple penetration	Thin materials: The relevant steel plate (0.10 mm thick) can be seen.	No	Yes	Yes
	Thick materials: The lead bar (1.5 mm thick) can be seen behind 14 mm of steel.	No	Yes	Yes
Test 6 STP: Material discrimination	Different colors are allocated to the sample of organic and inorganic substances (sugar and salt discrimination).	Yes	Yes	Yes

Note. "Yes" means the requirement of the STP is fulfilled. For a machine to be STP compliant, it must pass all tests. STP = standard test piece.

discrimination. For each measure, certain requirements need to be fulfilled for the machine to pass the test. Test 1: Single Wire Resolution. This defines the ability to display a single thin wire. Test 2: Useful Penetration. This determines what level of detail should be seen behind a thickness of known material. Test 3: Spatial Resolution. This defines the ability to distinguish and display objects that are close together. Tests 4 and 5: Simple Penetration. These test the X-ray machine's ability to image thin and thick material as well as the thickness of steel the X-ray machine should be able to penetrate. Test 6: Material Discrimination. This ensures that different colors are allocated to organic and inorganic substances.

In our study, we used one 2D multiview X-ray and one 3D CT imaging system; both are operational at airports and representative of their category (the names of the systems cannot be revealed for this publication, but we can state that the 3D imaging system belongs to the most widely used in the world). A certified European test center conducted the image quality assessment using the STP. The results are shown in Table A1.

As expected, the 2D imaging system was STP compliant; that is, it passed all tests. The 3D imaging system did not pass two of the six tests: The spatial resolution and useful penetration tests could not be solved using either the 3D-rotatable or the 3D-sliceable image. However, when comparing the STP results of both systems, it can be

assumed that it should be possible to recognize main IED components (triggering devices, power sources, explosives, and detonators) to a similar degree with 2D and 3D imaging. Nonetheless, recognizing thin wires when they are hidden behind aluminum of a thickness of 7.9 mm or more is not possible with 3D imaging.

To summarize, 3D imaging systems have lower image quality than 2D imaging systems according to tests using the STP and protocol, which is currently the most widely used international standard for assessing the image quality of 2D imaging systems. Despite this, our study could show that 2D and 3D screeners attained the same detection performance with 2D and 3D imaging. Based on the fact that newer 3D imaging technology has better automated explosive detection and therefore higher baggage throughput (Flitton et al., 2013; Mouton & Breckon, 2015; Oftring, 2015; Wells & Bradley, 2012), we argue that 3D imaging is superior to 2D imaging despite its lower image quality. Whether 3D screening is superior or inferior to 2D screening also depends on the visual inspection performance of X-ray screeners working with such systems. Therefore, when installing new technology at airports, it is important to consider not only technical features but also human factors.

Regarding image quality testing, no specific image quality assessment procedure is yet available for 3D imaging. Based on the results of our study, we argue that image quality standards and procedures need revision before they can be applied to 3D imaging systems. Regulatory bodies should not only evaluate technical aspects when testing image quality but also take human factors into account using experiments with highly realistic images, simulators, and screeners as participants as we did in our study.

ACKNOWLEDGMENTS

We thank the German Federal Police Technology Center for the valuable expertise and support for creating the stimulus material.

KEY POINTS

- This study compared the performance of airport security officers (screeners) using state-of-the-art 3D imaging and older 2D imaging for airport security screening of hold baggage.
- Despite lower image quality, screeners achieved a similar detection performance with 3D imaging to that for 2D imaging.
- 3D screeners revealed higher detection performance with both types of imaging than 2D screeners.
- Features of 3D imaging systems (3D rotation and slicing) seem to compensate for the lower image quality.
- Visual inspection competency acquired with one type of imaging seems to transfer to the other type of imaging.
- 2D and 3D screeners required more time for visual inspection of 3D versus 2D images. However, baggage throughput would still be substantially higher with 3D imaging systems for hold baggage screening due to lower EDS alarm rates than those for older 2D imaging systems.
- Replacing older 2D with newer 3D imaging systems for hold baggage screening can be recommended to increase the efficiency and effectiveness of hold baggage screening.
- Extensive and specific training of 2D screeners before allowing them to work with 3D imaging is not needed to achieve a similar performance to that with 2D imaging.
- Current image quality standards for 2D imaging need to be revised before they can be applied to 3D imaging systems for hold baggage screening.

REFERENCES

Barrat, H. H. (2000). *Handbook of medical imaging*. Bellingham, WA: SPIE Press.

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

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. *Applied Cognitive Psychology*, *29*(1), 142–148. doi: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. *HCI and Usability for Education and Work, Lecture Notes in Computer Science*, *5298*, 419–438. doi:10.1007/978-3-540-89350-9_30

Bretz, E. A. (2002). Slow takeoff. *IEEE Spectrum*, *39*(9), 37–39.

Caygill, J. S., Davis, F., & Higson, S. P. (2012). Current trends in explosive detection techniques. *Talanta*, *88*, 14–29. doi:10.1016/j.talanta.2011.11.043

Clark, K., Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2012). Overcoming hurdles in translating visual search between the lab and the field. *Nebraska Symposium on Motivation*, 59, 147–181.

European Commission (2015, November 5). 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*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R1998&from=EN

European Civil Aviation Conference (ECAC). (2018, April). Retrieved from https://www.ecac-ceac.org/cep

Fiore, S. M., Scielzo, S., Jentsch, F., & Howard, M. L. (2006). Understanding performance and cognitive efficiency when training for X-ray security screening. *In Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 2610–2614). Santa Monica, CA: Human Factors and Ergonomics Society.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*, 943–947.

Flitton, G., Breckon, T., & Megherbi, N. (2010). Object recognition using 3D SIFT in complex CT volumes. In *British Machine Vision Conference* (pp. 11.1–11.12). Aberystwyth, Wales: BMVA Press. doi:10.5244/C.24.11

Flitton, G., Breckon, T., & Megherbi, N. (2013). A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery. *Pattern Recognition*, *46*(9), 2420–2436. doi:10.1016/j.patcog.2013.02.008

Franzel, T., Schmidt, U., & Roth, S. (2012). Object detection in multi-view X-ray images. In A. Pinz, T. Pock, H. Bischof, & F. Leberl (Eds.), *Pattern recognition: Joint 34th DAGM and 36th OAGM Symposium, Graz, Austria* (pp. 144–154). Berlin, Germany: Springer.

Ghylin, K. M., Drury, C. G., & Schwaninger, A. (2006). *Two-component model of security inspection: Application and findings*. 16th World Congress of Ergonomics, IEA 2006, Maastricht, The Netherlands, July, 10–14, 2006. doi:10.13140/RG.2.1.2216.8567

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. doi: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: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. *International Journal of Aviation Psychology*, *23*(2), 113–129. doi:10.1080/10508414.2011.582455

Hancock, P. A., & Hart, S. G. (2002). Defeating terrorism: What can human factors/ergonomics offer? *Ergonomics in Design*, *10*(1), 6–16.

Harding, G. (2004). X-ray scatter tomography for explosives detection. *Radiation Physics and Chemistry*, *71*, 869–881. doi:10.1016/j.radphyschem.2004.04.111.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006, June). 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, ICRAT 2006, Belgrade, Serbia and Montenegro* (pp. 393–397). doi:10.5167/uzh-97986

Hardmeier, D., & Schwaninger, A. (2008, June). Visual cognition abilities in X-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008* (pp. 311–316). Fairfax, VA. doi:10.13140/RG.2.1.4335.7924

Harris, D. H. (2002). How to really improve airport security. *Ergonomics in Design*, *10*(1), 17–22.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

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. doi:10.1080/00140130802526935

Koller, S., 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:10.1007/s12198-007-0006-4

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

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

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

McCarley, J. S. (2009). Response criterion placement modulates the benefits of graded alerting systems in a simulated baggage screening task. *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (pp. 1106–1110). Santa Monica, CA: Human Factors and Ergonomics Society.

McCarley, S., Kramer, A. F., & Wickens, C. D. (2004). Visual skills in airport-security screening. *Psychological Science*, *15*(5), 302–306. doi:10.1111/j.0956-7976.2004 .00673.x

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*, 1–10. doi: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:10.1037/a0019569

Mitroff, S. R., Biggs, A. T., & Cain, M. S. (2015). Multiple-target visual search errors: Overview and implications for airport security. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 121–128. doi:10.1177/237273221560

Mouton, A., & Breckon, T. P. (2015). A review of automated image understanding within 3D baggage computed tomography security screening. *Journal of X-ray Science and Technology*, *23*(5), 531–555. doi:10.3233/XST-150508

Oftring, C. (2015). Assessing the impact of ECAC3 on baggage handling systems – considerations for upgrading existing ECAC2 systems. Retrieved from https://www.copybook.com/media/airport/profiles/beumer/documents/1464176454_ECAC%20Standard%203.pdf

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rice, S., & McCarley, J. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*(4), 320–331.

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), 1–17. doi: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*, 439. doi:10.3389/fnhum.2015.00439

Rusconi, E., McCrory, E., & Viding, E. (2012). Self-rated attention to detail predicts threat detection performance in security X-ray screening. *Security Journal*, *25*, 356–371. doi:10.1057/sj.2011.27

Schuster, D., Rivera, J., Sellers, B. C., Fiore, S. M., & Jentsch, F. (2013). Perceptual training for visual search. *Ergonomics*, *56*(7), 1101–1115. doi:10.1080/00140139.2013.790481.

Schwaninger, A. (2005). Increasing efficiency in airport security screening. *WIT Transactions on the Built Environment*, *82*, 407–416. doi:10.2495/SAFE050401

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

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

Schwaninger, A., Hardmeier, D., Riegelnig, J., & Martin, M. (2010). Use it and still lose it? The influence of age and job experience on detection performance in X-ray. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *23*(3), 169–175. doi:10.1024 /1662-9647/a000020

Schwaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In K. Morgan & M. J. Spector (Eds.), *The Internet society: Advances in learning, commerce and security* (pp. 147–156). Southampton, England: WIT Press. doi:10.13140/RG.2.1.4051.8649

Schwaninger, A., Hofer, F., & Wetter, O. E. (2007). Adaptive computer-based training increases on the job performance of x-ray screeners. *Proceedings of the 41st Carnahan Conference on Security Technology*, *Ottawa, October 8–11, 2007*. doi:10.1109/CCST.2007.4373478

Schwaninger, A., Michel, S., & Bolfing, A. (2005). Towards a model for estimating image difficulty in X-ray screening. *Proceedings of the 39th Carnahan Conference on Security Technology*, 39, 185–188. doi:10.1109/CCST.2005.1594875

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in X-ray screening using image measurements. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization* (pp. 123–130). New York, NY: ACM Press. doi:10.1145/1272582.1272606

Shanks, N. E. L., & Bradley, A. L. W. (2004). *Handbook of checked baggage screening: Advanced airport security operation*. London, England: Professional Engineering Publishing.

Singh, S., & Singh, M. (2003). Explosives detection systems (EDS) for aviation security. *Signal Processing*, 83(1), 31–55. doi:10.1016/S0165-1684(02)00391-2

Strantz, N. J. (1990) Aviation security and Pan Am Flight 103: What have we learned. *Journal of Air Law and Commerce*, *56*, 413.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlbaum.

Tarr, M. J., & Vuong, Q. C. (2002). Visual object recognition. In H. Pashler (Series Ed.) & S. Santis (Ed.), *Stevens' handbook of experimental psychology: Vol. 1. Sensation and perception* (3rd ed., Vol. 1, pp. 287–314). New York, NY: Wiley. doi:10.1002 /0471214426.pas0107

Turner, S. (1994). *Terrorist explosive sourcebook countering terrorist use of improvised explosive devices*. Boulder, CO: Paladin Press.

Verghese, P. (2001). Visual search and attention: A signal detection approach. *Neuron*. *31*, 523–535. doi:10.1016/S0896-6273(01)00392-0

Vuong, Q. C., & Tarr, J. T. (2004). Rotation direction affects object recognition. *Vision Research*, *44*, 1717–1730. doi:10.1016/j .visres.2004.02.002

von Bastian, C. C., Schwaninger, A., & Michel, S. (2008). Do multi-view X-ray systems improve X-ray image interpretation in airport security screening? *Zeitschrift für Arbeitswissenschaft*, *3*, 166–173. doi:10.3239/9783640684991

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

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. doi:10.1016/j.apradiso.2012.01.011

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

WG Standard Test Piece (n.d.). Retrieved April 4, 2018 from https://www.wi-ltd.com/wp-content/uploads/2016/03/WG-X-Ray-Machine-Standard-Test-Piece-STP.pdf

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*(3), 33. doi:10.1167/13.3.33

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, *435*, 439–440. doi: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:10.1037/0096-3445.136.4.623

Wolfe. J. M., & Reynolds, J. H. (2008). Visual search. In A. I. Basbaum, A. Kaneko, G. M. Shepherd, & G. Westheimer (Eds.), *The senses: A comprehensive reference* (Vol. 2, pp. 275–280). San Diego, CA: Academic Press.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology*, *20*: 121–124. doi:10.1016/j .cub.2009 .11.066.

Nicole Hättenschwiler is a PhD student working at the University of Applied Sciences and Arts, Northwestern Switzerland, in the field of human factors in aviation security. She obtained her Master of Science in Psychology from the University of Bern in 2014.

Marcia Mendes earned her PhD in the field of human factors in aviation security from the University of Basel in 2016.

Adrian Schwaninger received his PhD in psychology from the University of Zurich in 2003. Since 2008, he has been Professor of Psychology at the Institute Humans in Complex Systems of the School of Applied Psychology of the University of Applied Sciences and Arts Northwestern Switzerland. Since 2009, he has been the Head of this Institute.

Date received: September 20, 2017
Date accepted: August 15, 2018

Airport security X-ray screening of hold baggage: 2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol

Nicole Hättenschwiler, Sarah Merks, 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

Email: nicole.haettenschwiler@fhnw.ch

***Abstract*— In airport security screening, passenger baggage that is transported in the hold of an aircraft (hold baggage) is screened using X-ray machines with explosive detection technology. Older systems are based on 2D multi-view imaging whereas newer systems are based on computer tomography (CT) that features 3D rotatable images (3D imaging). Regulators, airport operators and security providers currently discuss whether extensive and specific training is necessary for screeners who are used to 2D multi-view imaging before they start working with 3D imaging. Moreover, to facilitate the decision making of screeners, so called on-screen alarm resolution protocols (OSARP) are available for 3D imaging. However, their effectiveness has not been investigated yet. To address these issues, we compared the visual inspection performance of screeners using state-of-the-art 2D multi-view imaging versus 3D imaging versus 3D imaging following a specific on-screen alarm resolution protocol (OSARP). In a simulated hold baggage screening task, screeners had to decide whether X-ray images contained an improvised explosive device (IED) or not. Results showed that there was no difference in detection performance (d') between 2D and 3D imaging. Visual inspection with 3D imaging following an OSARP resulted in higher detection performance (d') compared to 2D and 3D imaging. In conclusion, screeners currently working with 2D multi-view technology do not need extensive and specific training to achieve comparable detection performance with 3D imaging. The application of an OSARP has the potential to further increase screeners' detection performance (d') with 3D imaging.**

Keywords*— *Airport security screening, 3D imaging, X-ray imaging technology, on-screen alarm resolution, visual search

I. Introduction

Compared to cabin baggage screening, where multiple target types (guns, knives, IEDs, explosives, other threats) pose a threat, there is only one threat category in hold baggage screening. As passengers cannot access items stored in the hold of an aircraft, guns and knives in hold baggage do not pose a threat. Therefore, hold baggage screening targets only fully functioning improvised explosive devices (IEDs) [1]. At airports, all hold baggage is screened by X-ray machines that are explosive detection systems (EDS). They indicate areas in X-ray images that might be explosive by colored frames or a specific surface color [2]. X-ray images on which an EDS has raised an alarm are sent to remote screening locations for *on-screen alarm resolution* (OSAR) by airport security officers (screeners). The task of hold baggage screeners is to visually inspect alarmed areas in X-ray images and decide whether such EDS alarms are harmless (false alarms of the EDS) or whether the hold baggage needs further inspection because it might contain an IED. Therefore, in HBS, the screeners' task is mainly a decision task whereas in cabin baggage screening, visual inspection consists of search and decision making [3], [4].

In order to decide whether a bag contains a fully functioning IED or not, screeners need to identify the following necessary components of an IED: a triggering device, a power source, explosive mass, and a detonator that need to be connected to each other by, for example, wires [2], [5]. Through computer-based training, screeners can learn to recognize these components, and they can achieve and maintain a high detection performance for IEDs [6] – [11]. International regulations consider this by mandating extensive and recurring training of screeners. For example, European regulations mandate at least 6 hours of image recognition training and testing in every 6-month period [12].

To screen hold baggage at airports, 2D multi-view X-ray imaging systems are currently the most widely used technology. However, 2D multi-view systems are not able to unambiguously reveal the exact bag content for complex packed and cluttered bags [13]. Newer technology is based on computer tomography (CT). Such systems generate a stack of contiguous slice images that are used to calculate and visualize a volumetric view of the bag [14]. Screeners have then the possibility to inspect bags as a 3D-rotatable images that are enhanced by using depth cues with the additional option to slice through them [2], [14], [15]. Due to these additional functions, 3D-rotatable images should facilitate the recognition of threats when they appear from an unusual viewpoint or if they are superimposed by other objects [16], [17]. Further, continuous exposure to 3D objects could

978-1-5386-7931-9/18/$31.00 ©2018 IEEE

result in richer visual object representations [18], [19], that could improve screeners' visual inspection performance in general. However, CT systems have lower image quality compared to 2D imaging [13], [15], [20], which could impair detection performance with 3D imaging. It is therefore interesting to compare 2D and 3D imaging because it is unclear whether the benefits of 3D imaging outweigh the potential negative effects of lower image quality.

Comparing 2D and 3D imaging is also of high practical relevance: Regulators, airport operators and security providers currently discuss whether extensive and specific training is necessary for screeners who are used to 2D multi-view imaging before they can be allowed to work with 3D imaging. Therefore, the first goal of this study was to compare the visual inspection performance of 2D screeners using 2D versus 3D imaging. Moreover, to facilitate decision making of screeners, so called on-screen alarm resolution protocols (OSARP) are available for 3D imaging that are assumed to improve screeners inspection performance. As their effectiveness has not been investigated yet, the second goal of this study was to compare 2D screeners' visual inspection performance using 3D imaging with and without following a specific on-screen alarm resolution protocol (OSARP).

To address these issues, we tested 2D screeners using a simulated hold baggage screening task under three conditions: 2D multi-view imaging versus 3D rotatable imaging versus 3D rotatable imaging following OSARP.

II. METHODS

A. Participants

Participants were $n = 62$ professional hold baggage screeners from an international airport with work experience using 2D multi-view imaging (for descriptive statistics see Table 1). All screeners had been 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 [12]. The current research complied with the American Psychological Association Code of Ethics and was approved by the Review Board of the School of Applied Psychology of the University of Applied Sciences and Arts Northwestern Switzerland. Informed written consent was obtained from all participants.

TABLE I.

Group	Description of Screeners Participating in the Study			
	n[a]	*Age in years*	*2D working experience in years*	*% female*
2D	20	45.7 (12.8)	9.8 (6.6)	60
3D	22	42.5 (9.4)	9.4 (9.0)	52
OSARP	20	45.6 (11.6)	9.8 (5.8)	45

[a] Two screeners dropped out due to illness after the first test date. Another screener had to be excluded prior to data analyses due to a system error of a simulator.

B. Design

Each screener first performed a pretest to familiarize with the 2D and 3D simulators and the testing procedure[1]. Screeners were assigned to balanced groups to be tested with either 2D imaging, 3D imaging (Fig. 1) or 3D imaging following an OSARP training. All three groups conducted the main test two weeks after the pretest using a between-subjects design with condition (2D vs. 3D vs. OSARP) as independent variable and visual inspection measures as dependent variables.

C. Apparatus

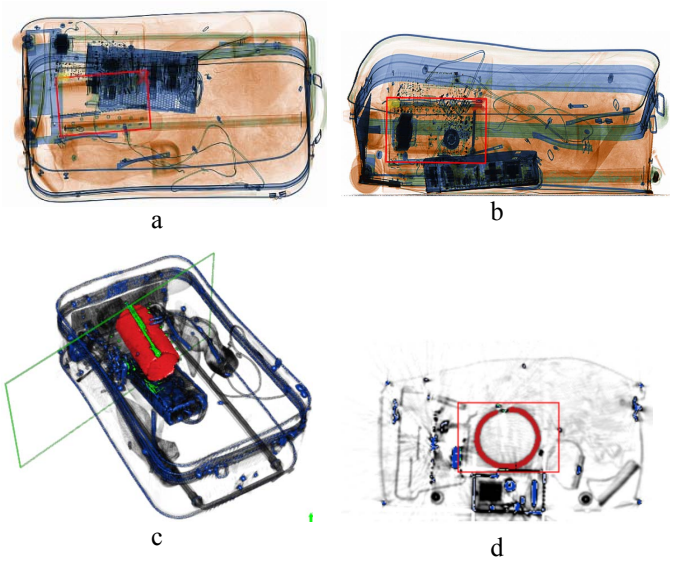

Fig. 1. Target-present bag containing an IED recorded with 2D and 3D imaging: (a) 2D default image, (b) 2D image with 90 degrees difference in perspective, (c) 3D-rotatable CT image, and (d) CT slice image. With 3D imaging, the detonator is visible in green (Fig. 1c) and in blue (Fig. 1d).

Screeners were tested using simulators provided by the manufacturer of the imaging systems. There were six individual testing stations with 19" TFT monitors and the room was dimly lit for testing. Four to six participants were tested at a time performing the test individually, quietly, and under supervision. This is a typical scenario in hold baggage screening [22].

D. Stimuli

X-ray image recording of baggage was conducted at a test center of a national transportation security organization. IEDs were prepared by an IED expert serving as consultant for this study. Thirty-two different bags were repeatedly used by repacking them to create unique stimuli for the pretest and the main test. All bags were packed in a way that resulted in medium X-ray image complexity as judged by the IED expert and the authors. Target-present images contained one IED and target-absent images one EDS false alarm (e.g., cheese, certain liquids, etc.).

The pretest consisted of 64 bag X-ray images recorded with 2D imaging and 64 different bag X-ray images recorded with 3D imaging. Target prevalence was 50% using 32 IEDs that

[1] Screeners also conducted different visual-cognitive tests; these results are reported elsewhere [21].

were shown twice in different bags using medium superposition: once recorded from a more frontal perspective, and once from a horizontally or vertically rotated perspective. The main test consisted of 256 bags that were recorded with both the 2D and 3D imaging system. To ensure the same system reliability (e.g., [23]), we used EDS alarms from the 3D imaging system as a reference when setting red frames manually around the objects of interest in the 2D imaging stimuli. Target prevalence was 50% using 32 different IEDs than in the pretest. Each IED was used four times in four different bags by varying viewpoint and superposition.

E. Procedure

For the pretest and the main test, screeners were instructed to visually inspect each X-ray image as if they were working at the airport and decide as accurately and quickly as possible whether or not the image contained a target (IED) by clicking on a target-present or a target-absent button on the simulator interface (a yes–no task in signal detection theory; see [24]). After the instructions, all participants started with 10 practice trials (5 target-absent and 5 target-present images in random order). A time limit of 60 s was set for viewing an X-ray image.

Screeners in the OSARP condition received a short training before conducting the main test. The original protocol developed by the IED expert is usually taught in two days and is customized to the images and interface of the 3D CT machine used in this study. This OSARP was adapted and shortened for our study so that it could be taught in 40 min. It included two heuristic steps on how to decide whether the bag is harmless or whether it contains an IED. After the training, the screeners took a break of 15 min before conducting the main test.

As the European regulation mandates that screeners have to take a break of at least 10 min after continuous visual inspection of X-ray images [12], the main test was divided into two blocks and screeners took a break of 10 to 15 min after completing the first block. Block order was counterbalanced across participants. Images appeared in random order within a block. All participants completed the pretest in less than 40 min and the main test in less than 1.5 hr including breaks.

F. Statistical Analyses

We computed analyses of variance (ANOVA) with detection performance (d'), response bias (c), target-absent RT, and target-present RT as dependent variables and test condition (2D, 3D and 3D with OSARP) as independent variable. All ANOVAs were conducted with SPSS version 22 and alpha was set at 0.05. Post hoc comparisons were conducted with R version 3.22 [25] and Holm–Bonferroni corrections were applied [26]. Effect sizes of ANOVAs are reported with η_p^2 (partial eta-squared); effect sizes of *t* tests, with Cohen's *d*. Our dependent variables detection performance (1) and response bias (2) were calculated using the following signal detection theory (SDT) formulae, whereby z refers to the inverse of the cumulative distribution function of the standard normal distribution [24], [27]:

$$d' = z(HR) - z(FAR) \quad (1)$$

$$c = -0.5 * z(FAR) - z(HR) \quad (2)$$

III. RESULTS

Fig. 2 shows detection performance d' depending on the three test conditions.

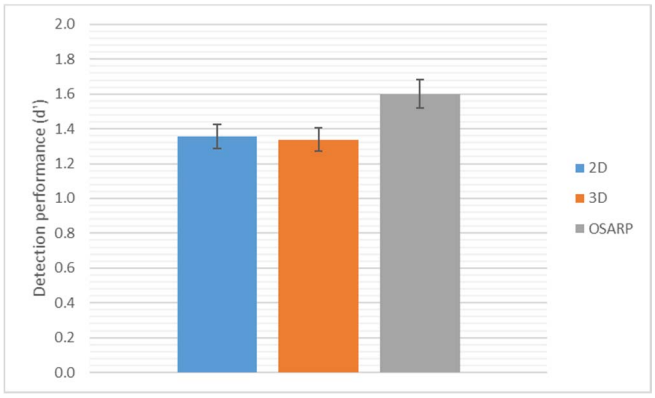

Fig. 2. Detection performance (d') by condition (2D vs. 3D vs. OSARP). Error bars are ± one standard error.

A one-way between subjects ANOVA with detection performance (d') as dependent variable [2] and condition as independent variable showed a significant main effect, $F(2, 62) = 3.28$, $p = 0.045$, $\eta_p^2 = 0.100$. To investigate whether detection performance (d') was different between the 2D and 3D imaging condition, a two-tailed *t*-test was performed. No statistical difference was found, $t(40) = 0.145$, $p = 0.886$, $d = 0.05$. To test the hypothesis that OSARP results in better detection performance (d') than 3D imaging, a one-tailed *t*-test was conducted. This showed that OSARP increased detection performance of screeners, $t(40) = 3.05$, $p = 0.002$, $d = 1.05$.

Fig. 3 shows response bias c by test conditions.

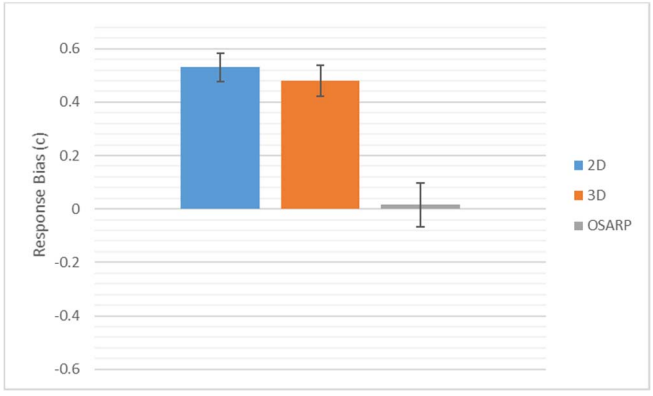

Fig. 3. Response bias c by condition (2D vs. 3D vs. OSARP). Error bars are ± one standard error.

[2] Sensitivity was recalculated under the assumption of an unequal variance distribution of threat- and false-alarm images [24]. A one-way between subjects ANOVA with da as sensitivity measure (with a slope parameter of 0.6) revealed a significant effect for condition $F(2, 62) = 19.01$, $p < 0.001$ and significant post-hoc comparisons between 2D and OSARP ($p < 0.001$) and 3D and OSARP ($p < 0.001$).

A one-way between subjects ANOVA with response bias (c) as the dependent variable showed a significant effect for condition, $F(2, 62) = 15.69$, $p < 0.001$, $\eta_p^2 = 0.35$. Posthoc tests using Holm-Bonferroni corrections showed that response bias of the OSARP condition was significantly lower and therefore more neutral than of the 2D ($p < 0.001$) and 3D condition ($p < 0.001$). The difference between the 2D and 3D condition did not reach statistical significance ($p = 0.610$).

Fig. 4 shows target-present and target-absent response times depending on the three test conditions.

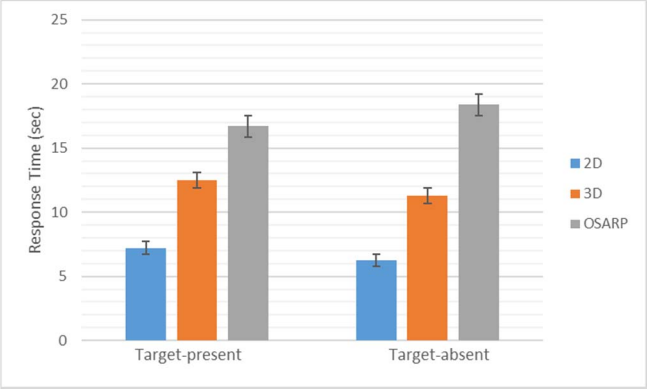

Fig. 4. Target-present and target-absent response times by condition (2D vs. 3D vs. OSARP). Error bars are ± one standard error.

A mixed design ANOVA with condition (2D vs 3D vs OSARP) as between-subjects variable, trial type (target present vs target absent) as within-subjects variable and response time (in seconds) as the dependent variable showed a significant effect of condition, $F(2, 59) = 48.32$, $p < 0.001$, $\eta_p^2 = 0.62$, and an interaction between condition and trial type, $F(2, 59) = 9.50$, $p < 0.001$, $\eta_p^2 = 0.24$. Posthoc comparisons using Holm-Bonferroni corrections showed a significant difference between all three conditions (2D, 3D, OSARP) for target present trials (all $p < 0.001$) as well as for target-absent trials (all $p < 0.001$). Screeners needed more time with 3D imaging and OSARP than with 3D imaging without OSARP while 2D imaging resulted in the fastest response times. The comparisons between target-present and target-absent trials showed longer response times for target-present in the 2D condition ($p = 0.022$) and 3D condition ($p < 0.001$), but longer target-absent response times for the OSARP condition ($p = 0.041$).

These results are very interesting as both the 2D and 3D condition showed longer response times for target-present trials, while in the OSARP condition screeners needed more time for target-absent trials. Longer response times for the conditions using 3D imaging were expected as rotating and slicing obviously needs more time. However, to test whether the longer response times were also a result of the inexpertness with the new 3D system, response times were compared between the two image blocks of the main test, to see whether screeners got faster over time. A mixed-design ANOVA with condition (3D vs. OSARP) as between-subjects variable, trial type (target-presents vs. target-absent) and block (1 vs 2) as within-subjects variables and response time as dependent variable showed a significant main effect of condition $F(1, 41) = 29.04$, $p < 0.001$, $\eta_p^2 = 0.42$, block $F(1, 41) = 53.59$, $p < 0.001$, $\eta_p^2 = 0.57$, and the interaction between condition and trail type, $F(1, 41) = 14.28$, $p < 0.001$, $\eta_p^2 = 0.26$. Post hoc comparisons were carried out for both conditions separately and showed significantly different response times between the first and second block for both target-absent and target-present trials ($p < 0.001$) for both the 3D imaging and OSARP condition with shorter response times for the second block.

IV. SUMMARY, DISCUSSION AND CONCLUSION

In this study, we addressed two issues of theoretical and practical relevance. First, with the implementation of advanced 3D CT technology at airports currently using 2D multi-view X-ray systems, the question arises whether extensive and specific training is necessary for screeners who have never worked with 3D imaging before. Second, it was investigated whether a specific on-screen alarm resolution protocol (OSARP) would increase detection performance of screeners when using 3D imaging. We compared visual inspection performance of 2D hold baggage screeners by using three different test conditions: state-of-the-art 2D multi-view imaging versus newer 3D rotatable imaging versus 3D rotatable imaging following OSARP.

Screeners achieved similar detection performance with 2D compared to 3D imaging despite lower image quality of 3D imaging. The possibility of 360° rotation allowing visual inspection from all angles eliminates, or at least drastically reduces, the challenges resulting from low target visibility due to viewpoint or superposition effects in 2D imaging. This seems to compensate potential negative effects of lower image quality of 3D imaging. Moreover, our results suggest that screeners experienced with 2D multi-view X-ray imaging can transfer their visual inspection competency to 3D imaging. Considering this result, extensive training of 2D screeners before they can be allowed to work with 3D imaging is not needed.

Our study further showed that screener's detection performance could be increased when they followed a specific OSARP to visually inspect 3D rotatable images. This result is especially interesting as screeners only received a short 40 min training of the protocol. Due to the OSARP training, screeners also shifted their response bias to be more neutral (more biased toward judgement of target present). This implies that screeners were very compliant with the protocol resulting in more hits but also more false alarm decisions in the OSARP condition. However, the relevance of response bias results should be interpreted with caution, as they are dependent on target prevalence [24], [27]. With a significantly lower target prevalence in operation, a different response bias can be expected. Further, in operation, a longer training would be required for the use of such a protocol and therefore even larger improvements in detection performance might be expected.

Screeners needed more time when hold baggage was displayed with 3D imaging compared to 2D imaging and even more when OSARP was applied. It was anticipated that visual inspection with 3D imaging takes more time, because rotating and slicing 3D images obviously takes longer to process than a visual inspection of static 2D X-ray images. However, due to lower EDS alarm rates of 3D imaging systems compared to 2D

imaging systems in operation [2], [13], [15], [20], baggage throughput would still be substantially higher in hold baggage screening. Further, by instructing the screeners to explicitly follow the OSARP it is not surprising that screeners took more time. However, we found a significant decrease of response times in the second block of testing, which suggests that longer response times with OSARP decrease with practice.

As mentioned in the introduction, in HBS, the screeners' task is mainly a decision task whereas in cabin baggage screening, visual inspection consists of search and decision making [3], [4]. Screeners in the 2D and 3D condition showed longer response times for target-present trials, whereas screeners in the OSARP condition needed more time for target-absent trials. A possible explanation is that in the 2D and 3D condition, screeners were able to quickly recognize EDS false alarms (i.e. when not all components of an IED are available), but took more time to confirm actual IEDs. This would also be consistent with the finding of a more conservative response bias.

Screeners in the OSARP condition took their decisions using a specific protocol following heuristic steps. They showed longer response times in target absent trials, which indicates that they took more time when an IED could not be identified. The more neutral response bias (more hits and more false alarms) indicates that the OSARP training could be improved further by adding heuristics to reduce false alarm decisions of screeners. However, it could also be possible that a longer training durations of the OSARP (compared to only 40 min like in this study) could already mitigate this problem. In summary, 2D screeners do not need extensive training to achieve comparable detection performance with 3D imaging. Training screeners with an OSARP increases detection performance but further research should be conducted to enhance OSARP training.

ACKNOWLEDGMENT

The authors particularly acknowledge the valuable contribution of Yotam Margalit during the whole process of the study. We also thank Melina Zeballos and Myrta Isenschmid for their help when conducting the data collection.

REFERENCES

[1] E. A. Bretz, "Slow takeoff," IEEE Spectrum, 39(9), pp. 37–39, 2002.

[2] K. Wells, and D. A. Bradley, "A review of X-ray explosives detection techniques for checked baggage," Applied Radiation and Isotopes, 70(8), pp. 1729–1746, 2012.

[3] S. Koller, C. Drury, and A. Schwaninger, "Change of search time and non-search time in X-ray baggage screening due to training," Ergonomics, 52(6), pp. 644–656, 2009.

[4] 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," Behavior Research Methods, 41(3), pp. 937–943, 2009.

[5] S. Turner, Terrorist explosive sourcebook countering terrorist use of improvised explosive devices. Boulder CO: Paladin Press, 1994.

[6] S. M. Fiore, S. Scielzo, F. Jentsch, and M. L. Howard, "Understanding performance and cognitive efficiency when training for X-ray security screening," Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, Santa Monica, CA: Sage, 2006, pp. 2610-2614.

[7] T. Halbherr, A. Schwaninger, G. R. Budgell, and A. W. J. Wales, "Airport security screener competency: A cross-sectional and longitudinal analysis," International Journal of Aviation Psychology, 23(2), pp. 113–129, 2013.

[8] S. Koller, D. Hardmeier, S. Michel, and A. Schwaninger, "Investigating training, transfer and viewpoint effects resulting from recurrent CBT of x-ray image interpretation," Journal of Transportation Security, 1(2), pp. 81–106, 2008.

[9] D. Schuster, J. Rivera, B. C. Sellers, S. M. Fiore, and F. Jentsch, "Perceptual training for visual search," Ergonomics, 56(7), pp. 1101–1115, 2013.

[10] A. Schwaninger, F. Hofer, and O. Wetter, "Adaptive computer-based training increases on the job performance of x-ray screeners," Proceedings of the 41st Carnahan Conference on Security Technology, Ottawa, October 2007.

[11] A. Schwaninger, and F. Hofer, "Evaluation of CBT for increasing threat detection performance in Xray screening," in The Internet society: Advances in learning, commerce and security, K. Morgan and M. J. Spector, Eds. Southampton, England: WIT Press, 2004, pp. 147–156.

[12] European Commission, "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. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri =CELEX:32015R1998&from=EN

[13] N. Megherbi, G. T. Flitton, and T. P. Breckon, "A classifier based approach for the detection of potential threats in CT based baggage screening," Proceedings of the International Conference on Image Processing ICIP, Hong Kong, pp. 1833–1836, September 2010.

[14] S. Singh, and M. Singh, "Explosives detection systems (EDS) for aviation security," Signal Processing, 83(1), pp. 31–55, 2003.

[15] A. Mouton, and T. P. Breckon, "A review of automated image understanding within 3D baggage computed tomography security screening," Journal of X-ray Science and Technology, 23(5), pp. 531–555, 2015.

[16] 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 Education and Work, Lecture Notes in Computer Science, 5298, pp. 419–438, 2008.

[17] A. Schwaninger, S. Michel, and A. Bolfing, "Towards a model for estimating image difficulty in x-ray screening," IEEE ICCST Proceedings, 39, pp. 185-188, 2005.

[18] M. J. Tarr, and Q. C. Vuong, "Visual object recognition," in Stevens' handbook of experimental psychology, Vol. I Sensation and perception, 3rd ed., H. Pashler, Series Ed., & S. Santis, Ed. New York: Wiley, 2002, pp. 287–314.

[19] Q. C. Vuong, and J. T. Tarr, "Rotation direction affects object recognition," Vision Research, 44, pp. 1717–1730, 2004.

[20] G. Flitton, T. Breckon, and N. Megherbi, "A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery," Pattern Recognition, 46(9), pp. 2420–2436, 2013.

[21] S. Merks, N. Hättenschwiler, M. Zeballos, and A. Schwaninger, "X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging?" Proceedings of the 52nd Carnahan Conference on Security Technology, Montreal Canada, October 2018.

[22] M. Kuhn, "Centralised image processing: The impact on security checkpoints," Aviation Security International, 23(5), pp. 28-30, 2017.

[23] S. Rice, and J. McCarley, "Effects of response bias and judgment framing on operator use of an automated aid in a target detection task," Journal of Experimental Psychology: Applied, 17(4), pp. 320–331, 2011.

[24] N. A. Macmillan, and C. D. Creelman, Detection theory: A user's guide, 2nd ed., Mahwah, NJ: Lawrence Erlbaum, 2005.

[25] R Core Team, R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2015.

[26] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, 6, pp. 65–70, 1979.

[27] D. M. Green, and J. A. Swets, Signal detection theory and psychophysics, New York, NY: Wiley, 1966

X-ray screening of hold baggage: Are the same visual-cognitive abilities needed for 2D and 3D imaging?

Sarah Merks, Nicole Hättenschwiler, Melina Zeballos, 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

Email: sarah.merks@fhnw.ch

Abstract— **2D multi-view X-ray imaging technology is widely used for security screening of hold baggage at airports. Newer technology is based on 3D CT imaging. Such systems offer the possibility to rotate a bag around 360 degrees. With the transition from 2D multi-view to advanced CT imaging, the question arises whether airport security officers (screeners) need the same visual-cognitive abilities when visually inspecting X-ray images of hold baggage. This study investigated the relationship between visual-cognitive abilities and visual inspection performance of screeners. Screeners conducted a computer-based visual cognitive test battery (VCTB) and a simulated hold baggage screening task with 2D and 3D imaging. We found that aspects of processing speed and visual processing correlated significantly with visual inspection performance of screeners using 2D imaging technology. In comparison, performance of screeners that visually inspected 3D images showed less correlations with the VCTB. These results indicate that with the expected change from 2D to 3D imaging technology in airport security, visual-cognitive requirements of the screeners might change. Therefore, further studies need to elucidate in more detail what visual-cognitive skills future 3D screeners need as it could affect personnel selection and development.**

Keywords— ***airport security, hold baggage screening, 2D multi-view imaging, 3D imaging, operator performance, visual inspection, visual cognitive abilities, X-ray imaging technology***

I. INTRODUCTION

X-ray screening of cabin baggage is an essential component of airport security. It takes place at security checkpoints before boarding an aircraft. Airport security officers (screeners) visually inspect X-ray images of cabin baggage, which consists of visual search and decision making [1]-[4]. In cabin baggage screening (CBS), prohibited items are guns, knives, improvised explosive devices (IEDs) and other items such as for example a self defense gas spray [5]. While the prohibited item categories are limited, there is a large variety of different exemplars and shapes of prohibited items [6], [7]. Individually adaptive computer-based training has been shown to be very important and effective to achieve and maintain a good detection performance in visual inspection of X-ray images [1], [8]-[10]. Besides the knowledge about prohibited items and their appearance in X-ray images, so-called image-based factors, have been shown to be important as well: Prohibited items in X-ray images are more difficult to recognize when depicted from unusual viewpoints, when superimposed by other objects and when placed in visually complex bags [11]-[13]. Screeners who can better cope with such image-based factors have better detection performance in X-ray image inspection [14], [15]. This could be related to certain visual-cognitive abilities, like logical thinking, figure-ground segregation and spatial imagination, that have been shown to correlate with detection performance in X-ray image inspection of *cabin baggage* [16]. As visual-cognitive abilities are assumed to be relatively stable (but vary substantially between people), screeners who have been selected based on an X-ray object recognition test perform better on the job than screeners who did not have to take such a test in the pre-employment assessment process [17]. To improve personnel selection, it is therefore important to investigate visual cognitive abilities as determinants of visual inspection performance.

The Cattell Horn Carrol (CHC) theory of intelligence [18] is one of the empirically most supported models that structures cognitive abilities [19]. The CHC theory includes nine factors underlying general intelligence. It offers a good integration of visual-cognitive abilities that have been shown to correlate with detection performance in 2D imaging: visual processing (G_v), processing speed (G_s) and fluid reasoning (G_f). Visual processing particularly refers to the screener's ability to mentally rotate an object, which matters when objects are depicted from unusual viewpoints in an X-ray image [16]. Further, the ability to distinguish between a shape and irrelevant features of the background [20], and spatial imaging have been shown to predict screeners' detection performance with 2D imaging [16]. In addition, processing speed might be relevant for the visual inspection task in terms of visual search efficiency. Fluid reasoning, which corresponds to an individual's logical thinking, is measured by verbal or figural deductive reasoning

978-1-5386-7931-9/18/$31.00 ©2018 IEEE

and visual discrimination. Previous research has suggested that fluid reasoning is an important predictor for detection performance in cabin baggage screening with the state-of-the-art 2D X-ray imaging.

It should be noted that previous research only examined visual cognitive abilities as predictors for *cabin* baggage screening. It is unclear whether the results from previous studies would apply also to *hold* baggage screening (HBS) because the task differs substantially from cabin baggage screening. During the flight, passengers cannot access items stored in the hold of an aircraft, so guns or knives do not pose a threat, and hold baggage screening targets only fully functioning IEDs [21]. Furthermore, hold baggage screeners are assisted by explosive detection systems (EDS) which mark potentially harmful material within a bag in X-ray images [22]. Therefore, screeners in HBS only analyze images that have been alarmed by the EDS and decide whether the specific alarmed object is harmless or whether it might be an IED and therefore additional security checks must be performed [21]. Thus, the task in HBS mainly consist of deciding whether an X-ray image contains an IED or not, whereas visual inspection in CBS consist of visual search and decision making [1]-[4].

Although 2D imaging technology is still the most widely used technology for security screening of hold baggage at airports, newer technology is based on 3D CT imaging. Such systems offer the possibility to rotate a bag around 360 degrees to inspect an object from different angles and viewpoints. 3D CT imaging also allows screeners to look through an alarmed object by using a slice view. Moreover, 3D rotatable images might also facilitate the recognition of prohibited items that, in certain 2D views, would be superimposed by other items in a complex bag. The relevance of image-based factors could change when inspecting static 2D versus 3D rotatable images, as the viewpoint and superposition effect may disappear and the need for mental rotation would then become less important. The possibilities, which advanced 3D CT systems offer in terms of image processing could therefore have an impact on the requirement of visual-cognitive abilities needed for reaching high detection performance. With the implementation of this advanced 3D technology at airports that are currently using state-of-the-art 2D imaging technology, it is important to examine whether the same visual-cognitive abilities correlate with a high detection performance using 2D and 3D CT imaging technology.

In order to investigate visual-cognitive requirements of HBS screeners and to compare their relationship with detection performance using 2D versus 3D CT imaging technology, we assessed visual-cognitive abilities of 2D HBS screeners with a computer-based test battery of 10 standardized test scales. Furthermore, we measured detection performance in a simulated HBS screening task either conducted with 2D or 3D CT imaging technology. The correlations of these measurements give a first idea of personnel requirements for 2D vs 3D HBS screeners.

II. METHODS

A. Participants

Participants were $n = 42$ professional hold baggage screeners (55% female) from a European international airport currently working with state-of-the-art 2D multi-view imaging technology. Average age of participants was $M = 44.1$ years ($SD = 11.1$ years). Their average working experience was M = 9.6 years (SD = 7.9 years). All screeners had been 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 [23]. Informed written consent was obtained from all participants.

B. Materials

1) Visual Cognitive Test Battery (VCTB). The VCTB consists of 10 standardized test scales that assesses a wide variety of narrow abilities underlying *visual processing*, *processing speed* and *fluid intelligence*. Four scales came from a major German intelligence test, the Leistungsprüfsystem 2 (LPS-2) [24]. Three tests were taken from a cognitive development test, that assesses visual perceptual weaknesses and strengths - the Test of Visual Perceptual Skills (TVPS-3) [25]. Another three scales were used from a Swiss online assessment test for students (WSI; http://www.was-studiere-ich.ch/). Finally, one scale consisted of the Raven's standardized progressive matrices (SPM) [26]. Most scales were originally in paper-pencil form. For this study, computer-based versions were created.

a) Processing Speed. Processing Speed was measured with the subtest 8 and 10 of the LPS-2. They both measure the ability to quickly and accurately perceive visual details, similarities and differences.

In *subtest 8 (**Perceptual Speed**)*, the task of participants was to recognize one out of five shapes that was embedded in a more complex pattern. The scale consisted of 40 patterns with increasing complexity. The score was calculated by the number of correct responses reached whithin two minutes.

In *subtest 10* (***Scan and Search***), participants had to compare two lists of characters shown next to each other and mark characters that were different in the second list. Scored was the amount of correct markings within two minutes.

b) Visual Processing. Visual processing was assessed with subtests 6 and 7 of LPS-2, three scales of TVPS-3 and three scales of WSI. Visual processing according to the CHC-theory describes a broader ability that consists of mental image processing, visualization and pattern recognition. The scales used in this study were the following:

Spatial relation was measured with subtest 6 from the LPS-2. Participants have to search for the one mirror-inverted number or letter in a list and mark it. The scale consists of 40 trials. Scored are the correct responses reached within two minutes.

Visualization, the ability to visualize a three-dimensional object, was measured with subtest 7 of LPS-2. Participants' task was to determine the number of surfaces of the geometrical figures. For this, participants need to visualize the figures in a three-dimensional space. There are 40 trials. The score was build by counting the number of correct responses reached within three minutes.

Visual memory was measured with a scale from the TVPS-3. In this task, participants have to memorize a design for five seconds and then recognize this pattern out of 4 alternatives on the next slide. The scale consists of 16 tasks and scored is the sum of correct responses.

Form Constancy was measured with another scale from TVPS-3. Participants are instructed to find a target shape within five alternative, more complex patterns, although it can

be rotated, increased or decreased in size. There are 16 trials and scored is the amount of correct responses.

Figure-ground segregation is defined as the abiltity to recognize a target shape within a very cluttered, busy background, and was measured with another scale of the TVPS-3. Participants have to choose one out of four complex patterns that included the target shape. There were 16 trials, scored were the amount of correct responses.

Slicing, another form of three-dimensional visualization was measured with the scale Slicing of the WSI. 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 these dividers slice the full objects and then choose all these pieces from a series of alternatives. Scored is each correctly chosen piece.

Another scale of the WSI, ***Spatial Rotation***, was conducted to have another measure of the ability to mentally rotate objects. Participants have to choose objects that are rotated but otherwise exactly the same as the target out of six alternatives. Scored are the amount of correct responses.

Unfold, another scale of the WSI, was another measure of visualization. Participants see a three-dimensional object and a series of folding templates. Participants then have to visualize the template that forms the original three-dimensional object. Scored are the amount of correct responses.

c) Fluid Intelligence. The Raven Standard Progressive Matrices were used in order to measure fluid intelligence. Participants see a matrix of logical patterns and have to choose the missing piece. The tests consists of 48 items that increase in complexity. Scored is the amount of correct responses reached within 10 minutes.

2) X-Ray Image Interpretation Test. To measure detection performance in hold-baggage screening, a realistic X-ray image interpretation test was created (for further details, see [27]). Recording of baggage was conducted at a test center of a national transportation security organisation where the same bags were recorded with a 2D multi-view and with a 3D CT system. Thirty-two different bags were used repeatedly by repacking them to create unique stimuli. All bags were of medium complexity. Target-absent images contained one EDS false alarm (e.g., cheese, certain liquids, etc.) and target-present images one IED. Preparation of the IEDs was done by an IED expert serving as a consultant for this study. Each IED was used twice in different bags: once recorded from a more frontal perspective displaying more surface area, and once from a horizontally or vertically rotated perspective using medium superposition. The main test consisted of 256 bags that were recorded with both the 2D and 3D imaging system. Target prevalence was 50%. Each of the 32 IEDs was used four times in four different bags by varying viewpoint and superposition.

To ensure that the 3D imaging condition had the same EDS system reliability, e.g. [28], as the 2D imaging condition, we used EDS alarms from the 3D imaging system as a reference when setting red frames manually around the objects of interest in the 2D imaging stimuli.

C. Apparatus

Screeners were tested at the Center for Adaptive Security Research and Applications (CASRA) in Zurich with simulators provided by the manufacturer of the 2D and 3D imaging systems. Six computer work stations with 19" TFT monitors were available for testing. Each screener sat approximately 50 cm away from the monitor. The X-ray images covered about two-thirds of the screen. Four to six participants were tested at a time performing the test individually, quietly, and under supervision. This is a typical scenario in hold baggage screening [29].

D. Procedure

Screeners were invited to the test facilities on two different days. First for the pretest and the visual-cognitive test battery and on the second day for the main test. Screeners were then assigned to balanced groups to be tested with either 2D imaging or 3D imaging by conducting a simulated hold baggage screening task using 2D multi-view (n = 20) or 3D rotatable images (n = 22). They were instructed to visually inspect 256 images of hold baggage and decide whether an EDS alarm was an IED or a false alarm.

E. Statistical Analyses

Two-sided Pearson correlations were calculated between the raw scores of each VCTB scale and X-ray detection performance (d'), as well as the response time of target-present and target-absent stimuli. Detection performance (d') was calculated using the following formula from signal detection theory, whereby z refers to the inverse of the cumulative distribution function of the standard normal distribution [30], [31]:

$$d' = z(HR) - z(FAR) \quad (1)$$

III. RESULTS

Correlations were calculated between the visual-cognitive abilities and d', for the 2D and 3D imaging condition separately. A summary of the correlation coefficients is presented in Table 1.

TABLE I. CORRELATIONS BETWEEN VISUAL-COGNITIVE ABILITIES AND DETECTION PERFORMANCE USING 2D VS 3D IMAGING

		2D Imaging	3D imaging
Processing Speed	**Perceptual Speed**	**0.39***	0.35
	Scan and Search	0.18	-0.11
Visual Processing	**Spatial relation**	**0.51***	0.26
	Visualization	0.32	0.09
	Visual Memory	**0.45***	0.07
	Form Constancy	0.32	0.12
	Figure-Ground	0.26	0.01
	Slices	-0.20	-0.17
	Spatial Rotation	-0.23	0.00
	Unfold	0.30	**0.45***
Gf	**Ravens Matrices**	0.23	0.08

Note: significance levels: * means $p < .05$

One aspect of processing speed (perceptual speed) and two aspects of visual processing (spatial relation and visual memory) correlated with detection performance with 2D imaging, while only one aspect of visual processing (unfold) correlated significantly with detection performance using 3D imaging technology.

Two-sided Pearson correlations were also calculated between the visual- cognitive abilities and response times for target-present stimuli, for both conditions separately. A summary of the correlation coefficients is given in Table 2.

TABLE II. CORRELATIONS BETWEEN VISUAL-COGNITIVE ABILITIES AND TARGET-PRESENT AND TARGET-ABSENT RESPONSE TIMES USING 2D VS 3D IMAGING

		Target-present		**Target-absent**	
		2D Imaging	**3D imaging**	**2D Imaging**	**3D imaging**
Processing Speed	**Perceptual Speed**	-0.01	0.18	-0.13	0.08
	Scan and Search	0.21	0.15	0.18	0.09
Visual Processing	**Spatial relation**	0.23	0.2	0.09	0.22
	Visualization	-0.05	-0.02	-0.2	-0.09
	Visual Memory	0.31	0.3	0.32	0.16
	Form Constancy	0.15	-0.21	0.03	-0.3
	Figure-Ground	-0.03	-0.1	-0.14	-0.14
	Slices	-0.16	-0.26	-0.12	-0.27
	Spatial Rotation	**-0.54***	-0.19	**-0.44***	-0.27
	Unfold	0.31	0.01	0.14	0.02
Gf	**Ravens Matrices**	0.23	0.18	0.1	0.03

There was only one significant correlation. High scores in spatial relations, an aspect of visual processing, were related to faster response times with 2D imaging.

Last, two-sided Pearson correlations were calculated between the visual-cognitive abilities and response times for target-absent stimuli, for each condition separately. A summary of the correlations is given in Table 2. Again, only one correlation was significant. High scores in spatial relations, an aspect of visual processing, were related with a faster response time in 2D imaging.

IV. SUMMARY, DISCUSSION AND CONCLUSION

Previous studies on *cabin* baggage screening have found that several visual cognitive abilities, like figure-ground segregation, form constancy, logical thinking and spatial imagination are related to a high detection performance with 2D imaging technology [7], [16], [32]. The first aim of our study was to examine whether results would be different for *hold* baggage screening. This was the case: We found that a high detection performance with 2D imaging technology for *hold* baggage screening was related to perceptual speed, spatial relation and visual memory. As mentioned in the introduction, visual inspection in HBS is mainly a decision task whereas visual inspection in CBS consists of search and decision making [1], [3]. Therefore, it is not surprising that the search and scan aspect of processing speed becomes less relevant in HBS. In regards of visual processing, our results suggest that certain aspects like figure-ground segregation or form constancy might be less relevant for HBS compared to previous findings on CBS , while spatial relation is an important aspect for both CBS and HBS. Furthermore, visual memory seems to be important for the decision making in HBS screening.

A second aim of this study was to compare visual cognitive abilities related to HBS detection performance with 2D imaging versus 3D imaging. The results suggest that that *processing speed* loses relevance for 3D imaging compared to 2D imaging. While perceptual speed correlated with detection performance with 2D imaging, there was no relation to detection performance with 3D imaging. One reason for this could be that the screeners that participated in the current study generally took more time to analyze 3D images compared to when visually inspecting 2D images, as they were trained 2D screeners with no experience with 3D imaging [27]. Furthermore, it is obvious that 3D screening in general is slower than 2D screening due to additional functionalites like rotating and slicing. It is therefore possible that visual-cognitive abilities related to processing speed might lose their relevance for detection performance with 3D imaging, especially if screeners are new to the 3D technology. In regards of *visual processing*, again, different aspects were related to detection performance with 2D imaging compared to 3D imaging. For 2D screening, spatial relation and visual memory were related to high detection performance, while only the aspect unfold, the ability to visualize a 3D object, was relevant for 3D screening. It makes sense that spatial relation, the ability to mentally rotate an object, becomes less important with 3D rotatable images, since the screener can physically rotate the image. Visual memory, which is mostly needed to match a mental representation with the objects in the X-ray image seems to only be relevant for 2D screening, but not for 3D screening. A reason for this could be that with 3D rotatable images, screener no longer match mental representations of an IED with objects in the X-ray image. Instead, they might look for components like explosive material and detonators, which are accordingly coloured. Last, *fluid intelligence* was not related to HBS detection performance at all, neither for 2D imaging nor for 3D imaging. This is not consistent with previous research on CBS [16]. Fluid intelligence is very close related to the concept of working memory [33]. Working memory allows for temporary storing and processing information and to perform several tasks at the same time. Therefore, working memory is very important for the CBS inspection task, which consists of search and decision making for several different threat categories. However, working memory capacity might lose relevance in HBS, which consists in decision makinf for only one threat category.

In regard of response times, our results suggest that screeners using 2D imaging technology who decided faster on target-present, as well as target-absent images also had higher skills in spatial relations, an aspect of visual processing that describes the ability to perceive the positions of objects in relation to oneself and/or other objects [34]. For 3D screening, none of the visual-cognitive abilities showed a significant relationship with response times what might be again due to the fact that screeners were not yet used to the new technology.

Overall, the results of the current study suggest that different visual cognitive abilities are relevant for 2D HBS compared to 2D CBS. Furthermore, the study suggests that screening with 3D imaging systems might also require different visual cognitive

abilites. The possibility of rotating the X-ray image of a bag and its content around 360 degrees seems to facilitate the recognition of prohibited items when depicted from unusual viewpoints, when superimposed by other items and when placed in visually complex bags. This might explain why fewer visual congitive abilities become relevant for 3D screening compared to 2D screening. However, the screeners participating in this study were already selected using a visual-cognitive test battery in the pre-employment assessment process. In a next study it should be investigated whether novices to 2D and 3D X-ray screening show a different relationship between certain visual-cognitive abilities and detection performance. It should also be elucidated in more detail what visual-cognitive skills future 3D screeners need and if there are differences between cabin and hold baggage screening.

REFERENCES

[1] S. Koller, C. Drury, and A. Schwaninger, "Change of search time and non-search time in X-ray baggage screening due to training," Ergonomics, 52(6), pp. 644–656, 2009.

[2] S. McCarley, A. F. Kramer and C. D. Wickens, «Visual skills in airport-security screening," Psychological Science, 15(5), pp. 302–306, 2004.

[3] A. Wales, C. Anderson, K. Jones, A. Schwaninger and J. Horne, "Evaluating the two-component inspection model in a simplified luggage search task," Behavior Research Methods, 41(3), pp. 937–943, 2009.

[4] J. M. Wolfe, and M. J. Van Wert, «Varying target prevalence reveals two, dissociable decision criteria in visual search," Current Biology, 20, pp. 121–124, 2010.

[5] A. Schwaninger, "Increasing efficiency in airport security screening," WIT Transactions on the Built Environment, pp. 407-416, 2005.

[6] A. Schwaninger, "Computer based training: a powerful tool to the enhancement of human factors," Aviation Security International, pp. 31-36, 2004.

[7] A. T. Biggs, M. R. Kramer and S. R. Mitroff, "Using cognitive psychology research to inform professional visual search operations," Journal of Applied Research in Memory and Cognition, 7(2), pp. 189-198, 2018.

[8] S. Koller, D. Hardmeier, S. Michel and A. Schwaninger, «Investigating training, transfer and viewpoint effects resulting from recurrent CBT of x-ray image interpretation," Journal of Transportation Security, 1(2), pp. 81-106, 2008.

[9] T. Halbherr, A. Schwaninger, G. Budgell and A. Wales, "Airport security screener competency: a cross-sectional and longitudinal analysis," International Journal of Aviation Psychology, 23(2), pp. 113-129, 2013.

[10] D. Schuster, J. Rivera, B. C. Sellers, S. M. Fiore and F. Jentsch, «Perceptual training for visual search," Ergonomics, 56(7), pp. 1101–1115, 2013.

[11] A. Schwaninger, D. Hardmeier and F. Hofer, «Aviation security screeners visual abilities & visual knowledge measurement," IEEE Aerospace and Electronic Systems, 20(6), pp. 29-35, 2005.

[12] A. Schwaninger, S. Michel and A. Bolfing, «A statistical approach for image difficulty esti-mation in x-ray screening using image measurements," Proceedings of the 4th Symposi-um on Applied Perception in Graphics and Visualization, pp. 123-130, 2007.

[13] 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 Education and Work, Lecture Notes in Computer Science, 5298, pp. 419-438, 2008.

[14] D. Hardmeier, F. Hofer, and A. Schwaninger, "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 Carnahan Conference on Security Technology, Las Palmas, Oc-tober 2005.

[15] F. Hofer, D. Hardmeier, and A. Schwaninger, "Increasing airport security using the X-Ray ORT as effective pre-employment assessment tool," Proceedings of the 4th International Aviation Security Technology Symposium, Washington D.C., November 2006.

[16] D. Hardmeier, and A. Schwaninger, "Visual cognition abilities in x-ray screening," Pro-ceedings of the 3rd International Conference on Research in Air Transportation ICRAT, Fairfax, June 2008.

[17] D. Hardmeier, F. Hofer, and A. Schwaninger, "The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening," Proceedings of the 4th International Aviation Security Technology Symposium, Washing-ton D.C., November 2006.

[18] J. B. Carroll, "The higher-stratum structure of cognitive abilities: Current evidence sup-ports g and about ten broad factors," The scientific study of general intelligence, pp. 5-21, 2003.

[19] D. P. Flanagan, The Cattell-Horn-Carroll theory of cognitive abilities. Encyclopedia of special education, pp. 368-382, 2008.

[20] J. M. Wolfe, A. Oliva, T. S. Horowitz, S. J. Butcher, and A. Bompas, "Segmentation of objects from backgrounds in visual search tasks," Vision Research, 42(28), pp. 2985-3004, 2002.

[21] E. A. Bretz, "Slow takeoff," IEEE Spectrum, 39(9), pp. 37–39, 2002.

[22] K. Wells, and D. A. Bradley, "A review of X-ray explosives detection techniques for checked baggage," Applied Radiation and Isotopes, 70(8), pp. 1729–1746, 2012.

[23] European Commission, "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. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri =CELEX:32015R1998&from=EN

[24] L. Kreuzpointner, H. Lukesch, and W. Horn, Leistungsprüfsystem 2. LPS-2. Göttingen: Hogrefe, 2013.

[25] N. A. Martin, Test of visual perceptual skills (TVPS-3), 3rd ed. Novato, CA: Academy Publishers, 2006.

[26] R. Horn, Standard Progressive Matrices (SPM). Deutsche Bearbeitung und Normierung nach J. C. Raven, 2nd ed. Frankfurt: Pearson Assessment, 2009.

[27] N. Hättenschwiler, S. Merks and A. Schwaninger, «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 Carnahan Conference on Security Technology, Montreal Canada, October 2018.

[28] S. Rice, and J. McCarley, "Effects of response bias and judgment framing on operator use of an automated aid in a target detection task," Journal of Experimental Psychology: Applied, 17(4), pp. 320–331, 2011.

[29] M. Kuhn, "Centralised image processing: The impact on security checkpoints," Aviation Security International, 23(5), pp. 28-30, 2017.

[30] D. M. Green, and J. A. Swets, Signal detection theory and psychophysics. New York, NY: Wiley, 1966.

[31] N. A. Macmillan, and C. D. Creelman, Detection theory: A user's guide, 2nd ed. Mahwah, NJ: Lawrence Erlbaum, 2005.

[32] E. Rusconi, F. Ferri, E. Viding and T. Mitchener-Nissen, "XRIndex: a brief screening tool for individual differences in security threat detection in x-ray images," Frontiers in human neuroscience, 9, p. 439, 2015.

[33] A. D. Baddeley, "Working memory: theories, models, and controversies," Annual review of psychology, 63, pp. 1-29, 2012.

[34] J. C. Chalfant, and M. A. Scheffelin, Central processing dysfunctions in children: A review of research. Washington, D.C.: U.S. Government Printing Office, 1969.

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

Number of words: 8733

Number of figures and tables: 10

Author Note

Correspondence concerning this article should be addressed to Nicole Hättenschwiler, University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Riggenbachstrasse 16, CH-4600 Olten, Switzerland. Email: nicole.haettenschwiler@fhnw.ch.

*Joint first authorship between Nicole Hättenschwiler and Sarah Merks

Abstract

The act of looking for targets amongst an array of distractors is a demanding cognitive task that has many real-world applications. 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, where professionals search for targets that are more complex, ambiguous and less salient. More concrete, there are several potential challenges when interpreting traditional visual search results in terms of their implications for the X-ray image inspection task at security checkpoints. In this study, it was tested whether a theoretical model with known facets of visual-cognitive abilities (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, testing students and professionals. Results show that the same visual-cognitive abilities predict performance in the conducted tasks in both populations. Further, results from a traditional visual search task can be transferred to an X-ray image inspection task and vice versa if no domain-specific knowledge is needed. Based on the results, similarities and differences will be discussed and to what degree studies conducted with students can be generalized to professionals and vice versa.

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

1. Introduction

Visual search, the act of looking for targets amongst an array of distractors, is a demanding cognitive task (e.g. Treisman & Gelade, 1980) that has many real-world applications. Some individuals conduct visual searches professionally, for example airport security officers (screeners), who visually inspect X-ray images of luggage to search for prohibited items or radiologists, who are looking for cancer in mammograms. In such professional applications, search errors can have huge, even fatal consequences and research can provide a valuable contribution to reduce these errors. The ability to locate a target amongst an array of distractors has been extensively studied for many decades (see Carrasco, 2011; Eckstein, 2011; Nakayama & Martini 2010, for recent reviews). But many of the studies on visual search used traditional, simplified tasks with salient stimuli and have been conducted with non-professional searchers (mostly students). These studies with a high experimental control have provided vital insights into the cognitive mechanisms underlying visual search. It is, however, not always clear whether the results from such traditional, simplified visual search tasks do extrapolate to real-world inspection tasks, where people search for targets that are more complex, ambiguous and/or less salient (e.g., Radvansky & Ashcraft, 2016). It is also unclear to what extent the findings based on student samples can be transferred to professionals who often rely on extensive training and experience. To address this issue, we will first briefly discuss insights from research based on traditional visual search tasks and based on a real-world application, namely X-ray image inspection. There are potential challenges when interpreting traditional visual search results in terms of their implications for the X-ray image inspection task at security checkpoints, like the different nature of stimuli or the different experience and characteristics of the searchers (Clark, Cain, Adamo, & Mitroff, 2012). Therefore, investigating to what extent results can be transferred, requires that both, the tasks and the searchers conducting these tasks, are examined.

Visual search is a type of perceptual task requiring attention and typically involves an active scan of the visual environment for a particular target among many distractors (Treisman & Gelade, 1980). Search 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 (storing a representation of the target item or items); for recent reviews see for example Carrasco (2011), Nakayama and Martini (2011), or Eckstein (2011). The ability to consciously locate a target amongst an array of stimuli has been extensively studied since the experiments by Helmholtz in the 19th century. Helmholtz first demonstrated that covert fixations - shifts of attention over a scene without eye movements - were real (Helmholtz 1896). Since then, models for visual search have been proposed, criticized and optimized for almost 40 years. Today, researchers mostly agree on a two-stage model for visual search, originally proposed by Neisser (1967), which is divided into a pre-attentive (subconscious accumulation of information from the environment) and an attentive stage (information analysis by conscious processing), differing in the amount of visual information processed at the same time. Hoffman (1975; 1978) further suggested the pre-attentive stage to be a parallel search process that guides the operation of the slower attentive serial discrimination stage. The pre-attentive guidance allows the searcher to find the target much more quickly. Influenced by these earlier studies, Treisman and Gelade (1980) and Treisman (1988) introduced the feature integration theory (FIT) of attention. In their model, they propose that single features of objects, e.g. a certain shape or color, are registered early and automatically in a parallel process across the whole visual field (bottom-up process).

By contrast, the combination of two or more features – so called *conjunctions* - are identified serially and with focused attention in a later stage (top-down process). Based on research to the feature integration theory, Wolfe (1994; 1998; 2003; 2007) introduced the model of guided search. The guided search model additionally proposes the production of activation maps through the combination of bottom-up and top-down activation during the parallel stage (Wolfe, 1994). The activation maps guide the second, serial stage and make it possible to guide attention based on information from more than one feature. The guided search model can explain why experimental data indicate that triple conjunctions (the combination of three features) are found more efficiently than standard conjunctions. This is because three parallel processes can guide attention more effectively than two can (Wolfe, Cave & Franzel, 1989).

Over the past several decades, psychological research has made tremendous headway in understanding the processes responsible when performing visual search tasks and the mechanisms that allow for the successful identification of target items (Clark et al., 2012). Most of the research that led to the development of the theories and models described above was conducted in labs with artificial stimuli to allow for maximum experimental control. These traditional visual search tasks are the so-called *feature- and conjunction search tasks*, where participants have to identify a previously known target amongst distractors (Treisman & Gelade, 1980). For feature search, the target differs from the distractors by a unique visual feature such as color, shape, orientation, or size, (McElree & Carrasco, 1999). An example of a feature search task is to find a red dot (target) in a set of yellow dots (distractors). Unlike feature search, conjunction search involves a distractor (or a group of distractors) that may differ from each other but exhibit at least one common feature with the target and therefore requires a combination of features to distinguish them (Shen, Reingold & Pomplun, 2003). A known example of a conjunction search is the *letter search task* which has been studied in many variations. The letters T and L share exactly the same features, differing only in their spatial arrangement (*LT*-letter search task: Treisman & Gelade, 1980). In one variation of this letter search task, participants are asked to identify the perfectly shaped letter *T* (target) surrounded by distractors of many distractor letters, including symmetrical and asymmetrical *Ts* and *Ls*. The efficiency of conjunction search in regards to accuracy and reaction time is dependent on the distractor-ratio and the number of distractors present (McElree & Carrasco, 1999) while reaction time restraints of conjunction search tend to show improvement through training (Reavis, Frank, Greenlee & Tsu, 2016). Such traditional search tasks offer ideal experimental control and statistical power and have been studied thoroughly using mostly students as participants (for reviews see e.g. Duncan & Humphreys, 1989; Eckstein, 2011; Wolfe, 1994, 1998).

While bottom-up processes may dominate the search process when searching for objects that are artificial or not familiar to a searcher, top-down processing becomes much more influential in everyday situations, which often rely on object recognition (Wolfe, 1998). In everyday life, people most commonly search their visual fields for targets that are familiar to them. In such scenarios, searchers must use their prior knowledge in order to accurately and efficiently locate targets such as phones, keys, etc. among distractors with much more complex features compared to a traditional feature or conjunction search task (Radvansky & Ashcraft, 2016). When it comes to searching for familiar stimuli, top-down processing allows searchers to more efficiently identify targets with greater complexity (Radvansky & Ashcraft, 2016). Still, visual search with complex objects appears to rely on the same active scanning processes as conjunction search

(e.g. LT-letter search) with less complex, contrived laboratory stimuli (Alexander & Zelinsky, 2011; 2012). Even more complex real-world search scenarios are that some humans conduct visual inspection tasks as their profession; for example radiologist inspecting mammograms for cancer (e.g. Krupinski, 1996; Nodine & Kundel, 1987) or screeners inspecting X-ray images for prohibited items (Drury, 1978; Koller, Drury, & Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, & Horne, 2009). Screeners visually inspect X-ray images by searching for prohibited items and deciding whether a bag is harmless or not, usually within seconds (Koller et al., 2009). Compared to traditional visual search tasks, the accuracy of such inspection tasks can have life-or-death implications. Besides some similarities to the traditional visual search task and its underlying processes, differences in regard to the nature of stimuli and characteristics of searchers need to be evaluated when translating results from a traditional visual search task to X-ray image inspection and vice versa.

Traditional visual search tasks normally include searching for only one target at a time, which is often known beforehand and occurs with high prevalence (e.g. 50%). In the case of security X-ray image inspection, a screener's task is to stay alert even though targets are hardly ever present. However, missing a threat can have severe consequences, whereas falsely rejecting too many bags will result in inefficiency and long waiting lines. Differences in stimuli between a traditional visual search task and X-ray image inspection tasks are known in regard to target and distractor complexity as well as the requirement of domain specific knowledge of the searcher. Targets in the traditional visual search task are often commonly known with salient shapes and colors whereas targets in X-ray image inspection tasks are not well specified, not salient and not predictable by the context (Bravo & Farid, 2004), while distractors may additionally produce clutter and superposition. The large variety of potential threat items and distracting objects in passenger bags make X-ray image inspection a difficult task. Therefore, domain specific knowledge is needed as screeners must know which items are prohibited and what they look like in X-ray images (Schwaninger, 2005, 2006).

To get an indication of performance in a real-world scenario like X-ray image inspection, experiments are mostly conducted with realistic stimuli (X-ray images) while testing both, students and professionals. 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 successfully identified (Koller et al., 2009). Features of guns and knives, for example, are known from everyday life experience and can therefore be detected without specific experience and training. Other prohibited items which are rather uncommon or have never been seen before (e.g. improvised explosive devices, IEDs) thus become very difficult to recognize for students if they have not been trained to recognize certain features of these threats (Schwaninger, 2004a; Schwaninger, 2005a). As mentioned above, feature integration theory (FIT) proposes that in order to perceive an object correctly, features have to be combined (Treisman, 1980). Training for threat detection has the goal of creating internal visual representations of objects and storing them in memory. Detection of objects, especially if unknown, should therefore improve with training because features are known and recognized better with repeated exposure. Several studies have shown that computer-based training (CBT)

can be a very effective and efficient tool for increasing X-ray screeners' performance (Halbherr et al., 2013; Koller et al. 2008; Koller et al., 2009; McCarley et al., 2004; Schwaninger & Hofer, 2004) why international regulations mandate initial and recurrent training of screeners (e.g., European Commission, 2015). Professionals like screeners conducting these X-ray image inspections have received years of training and therefore have a lot of experience on this specific tasks. Further differences between traditional visual search and X-ray image inspection tasks could therefore be due to experience, training and characteristics of the searchers themselves (Clark et al., 2012). Often, university students are the first choice as participants for traditional visual search research as they are easily accessible. Traditional visual search tasks often reveal a great deal of variability in performance in student participants. While some of this variability may be tied to differences in the underlying search ability, some variability may also result from differences in motives and motivation to conduct visual search tasks compared to professionals whose decisions have life or death implications (Clark et al., 2012).

While these tasks differ and are often conducted by different populations, the underlying cognitive processes are comparable to a certain extent. Traditional visual search and X-ray image inspection can be characterized as a basic, core cognitive task. As by the definition of Caroll (1993), a cognitive task is any task in which correct processing of mental information is critical to successful performance. Therefore, specific cognitive abilities are needed to successfully perform such a cognitive task. These abilities can be measured by specific correlated measures, which then describe performance. To conduct visual search and inspection, certain visual-cognitive abilities have been found to correlate with a higher performance like attention, memory, visual processing or processing speed (Bolfing & Schwaninger, 2009; Wolfe, Oliva, Horowitz, Butcher & Bompas, 2002). If there are individual differences in performance on visual search or inspection tasks, the individual differences can be seen as the direct manifestation of differences in an underlying ability or latent trait (Caroll, 1993; 2003).

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, informing a substantial body of research and the ongoing development of intelligence tests (McGrew, 2005). There is a large number of distinct individual differences in cognitive abilities. 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, consisting of a single general ability also called *g* (Flanagan & Harrison, 2005). The factors describe stable and observable differences among individuals. However, the structure of the three strata is hierarchical. This means that the abilities within one stratum, e.g. the narrow abilities of stratum I are positively intercorrelating, which allows the estimation of stratum II, the broad abilities. Likewise, the abilities of stratum II have nonzero intercorrelations, which allows the estimation of stratum III. So while the abilities within one stratum are related, a large amount of evidence shows that they are unique and reliably distinguishable (see e.g., Keith & Reynolds, 2012). Thereby, visual processing (*Gv*), short-term memory (*Gsm*) and processing speed (*Gs*) are broad abilities of stratum II that are accepted components with a known influence on visual search and inspection performance and therefore included in most commonly used measures of intelligence (e.g., Stanford-Binet: Roid, 2003a, 2003b; Wechsler Intelligence Scale: Wechsler, 1997). Visual processing (*Gv*) describes a broader ability to perceive, analyze, synthesize, and think with 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. Since analyzing, synthesizing and thinking in visual patterns are cognitive operations as well, Gv and Gsm are closely related, but can be distinguished by the limited capacity of the short-term memory. Processing speed (*Gs*) describes the ability to quickly and accurately perceive visual details, similarities and differences. Several studies proved the influence of higher *Gv*, *Gsm* and *Gs* on better performance in traditional visual search tasks (Alvarez & Cavanagh, 2004; Eriksen & Schultz, 1979; Luck & Vogel, 1997).

For the visual inspection task of X-ray screening, authors have discovered so called image-based factors such as bag complexity, superposition and viewpoint of the threat item that influence screeners' performance (Bolfing, Halbherr & Schwaninger, 2008; Schwaninger et al. 2005). Items which are presented from unusual or rotated viewpoints become more difficult to detect (effect of viewpoint) (Palmer, Rosch & Chase, 1981). Similarly, the position of a prohibited item in a bag and its superimposition by other objects (effect of superposition), or the number and types of items in a bag which could attract attention (effect of bag complexity) also affect the difficulty of recognizing prohibited items. Bag complexity comprises the factors clutter (disarrangement, textural noise, chaos, etc.) and opacity (X-ray penetration of objects) (Schwaninger et al., 2008). Visual-cognitive abilities such as figure-ground segregation (related to superposition) or mental rotation (related to view difficulty) can therefore play an important role for visual inspection performance in X-ray screening. Therefore, it is not surprising that cognitive abilities have also been linked to inspection performance in many studies on X-ray image inspection with professionals (Hardmeier et al., 2005; Hardmeier & Schwaninger, 2008). More precisely, studies found an influence of mental rotation and figure-ground segregation on higher performance (Bolfing & Schwaninger, 2009; Wolfe, Oliva, Horowitz, Butcher & Bompas, 2002) which are narrow abilities of visual processing (*Gv*). Memory capacity, what can be classified as short-term memory (*Gsm*), is strongly associated with visual inspection in general (e.g., Lavie & DeFockert, 2005; Poole & Kane, 2009; Roper, Cosman, & Vecera, 2013). In addition, processing speed (*Gs*) might be relevant for the visual inspection task in terms of efficiency (Salthouse, 1996). Professionals should therefore ideally possess the visual-cognitive abilities required to conduct visual inspection of X-ray images before starting the job as a screener. Thus, when recruiting new personnel, some airports conduct pre-employment assessments that test for these visual abilities and aptitudes (e.g. X-Ray Object Recognition Test; see Hardmeier et al., 2005; Hardmeier & Schwaninger, 2008).

Taking all this into account, it would be very valuable to compare the influence of specific visual-cognitive abilities (Gv, Gsm and Gs) between a traditional visual search task and an X-ray image inspection task, in order to gain a deeper understanding of the determinants of performance. Comparing students and professionals can further give indications on whether visual-cognitive abilities can be compensated with experience and training of the described tasks. Research on visual cognitive abilities and X-ray image interpretation is scarce. To this date, no study exists comparing students and professionals using a traditional visual search task and an X-ray image inspection task. However, such a comparison can only be regarded as fair, if an X-ray image inspection task without domain-specific knowledge is used. To address this research gap, we conducted the same experiment once with students and once with professionals while addressing the following research questions: (1) Do the same 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?

Based on the literature mentioned above we postulate a theoretical model where several known facets of visual-cognitive abilities (visual processing *Gv*, short-term memory *Gsm* and processing speed *Gs*) can predict performance in the examined task. This model will be tested on two populations (students and professionals) using the same experimental stimuli —including both traditional and realistic real-world tasks and stimuli. Based on our results, insights shall be provided on similarities and differences between these tasks and searchers conducting the tasks, which could provide important information on how well studies conducted with students can be generalized to professionals and vice versa.

2. Methods

2.1. Participants

Descriptives of participants in this study can be seen in Table 1. N = 128 were *students* from the University of Applied Sciences and Arts Northwestern Switzerland. N = 112 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.

2.2. 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 and professional searchers were tested at the test facilities of the Center for Adaptive Security Research and Applications (CASRA) using the same computers and monitors.

2.3. Stimuli

2.3.1 Visual Cognitive Test Battery

A Visual-cognitive test battery (VCTB) was developed to measure a broad spectrum of visual-cognitive abilities that assesses a wide variety of narrow abilities underlying *visual processing (Gv), short-term memory (Gsm)* and *processing speed (Gs)* to make predictions on performance. The VCTB consists of 10 standardized tests scales which are mostly part of well established intelligence tests based on the Cattell-Horn-Carroll theory of intelligence (Carroll, 1993; 2003; Cattell, 1941; Horn, 1965). Four scales came from a major German intelligence test, the Leistungsprüfsystem 2 (LPS-2; Kreuzpointer, Lukesch & Horn, 2013). Three tests were taken from a cognitive development test, that assesses 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; http://www.was-studiere-ich.ch/). In addition, we included the Raven's standardized progressive matrices (SPM; Horn, 2009) as a general measure of fluid intelligence. Most scales were originally in paper-pencil form why computer-based versions were created. Psychometrical criteria of the tests scales are reported in Table 2.

2.3.1.1 Visual Processing (Gv).

Visual processing was assessed with three scales of TVPS-3 (visual memory, form constancy and figure-ground segregation). For *Visual memory*, participants have to memorize a design for five seconds and then recognize this pattern out of four alternatives on the next slide. The scale consists of 16 tasks and scored is the sum of correct responses. To measure *Form Constancy*, participants are instructed to find a target shape within five alternative, more complex patterns, although it can be rotated, increased or decreased in size. There are 16 trials and scored is the amount 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 included the target shape. There were 16 trials, scored were the amount of correct responses.

2.3.1.2 Short-term memory (Gsm).

Short-term memory was measured using three scales from the WSI (slicing, spatial rotation and unfurl). *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 these dividers slice the full objects and then choose all these pieces from a series of alternatives. Scored is each correctly chosen piece. *Spatial Rotation* was conducted to have another measure of the ability to mentally rotate objects. Participants have to choose objects that are rotated but otherwise exactly the same as the target out of six alternatives. Scored are the amount of correct responses. *Unfold* is another measure of visualization where 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. Scored are the amount of correct responses.

2.3.1.3 Processing Speed (Gs).

Processing speed was measured with the subtest 6, 7, 8 and 10 of the LPS-2 (spatial relation, visualization, perceptual speed and scan/search). All scales measure the ability to quickly and accurately perceive visual details, similarities and differences. *Spatial relation* was measured with subtest 6 where participants have to search for the one mirror-inverted number or letter in a list and mark it. The scale consists of 40 trials. Scored are the correct responses reached within two minutes. *Visualization,* the ability to visualize a three-dimensional object, was measured with subtest 7. Participants' task was to determine the number of surfaces of the geometrical figures. For this, participants need to visualize the figures in a three-dimensional space. There are 40 trials. The score was build by counting the number of correct responses reached within three minutes. In subtest 8, *Perceptual Speed*, the task of participants was to recognize one out of five shapes that was embedded in a more complex pattern. The scale consisted of 40 patterns with increasing complexity. The score was calculated by the number of correct responses reached within two minutes. In subtest 10, *Scan and Search*, participants had to compare two lists of characters shown next to each other and mark characters that were different in the second list. Scored was the amount of correct markings within two minutes.

2.3.1.4 Fluid Intelligence.

The Raven Standard Progressive Matrices Plus (SPM) is a language-independent test that was used in order to measure fluid intelligence. Participants see a matrix of logical patterns and have

to choose the missing piece out of six to eight abstract figures (Raven, Raven, & Court, 2003). The tests consists of 48 items that increase in complexity. Scored is the amount of correct responses reached within 10 minutes.

2.3.2 Simulated Baggage Screening Task

The simulated baggage screening task (SBST) was created based on the X-Ray Object Recognition Test (X-Ray ORT, Schwaninger, Hardmeier & Hofer, 2005; Hardmeier, Hofer & Schwaninger, 2006b). The original ORT (single-view) has been designed to measure how well professional and non-professional searchers can cope with image-based factors that have an 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, i.e., object shapes that can be assumed to be known by most people. All X-ray images are in black-and-white, as colors are mainly diagnostic for the material of objects in the bag, and thus, could be primarily helpful for experts. In addition, all guns and knives are shown for 10 seconds before the test starts, which further reduces the role of knowledge-based factors in this test.

The SBST created for this experiment included 256 X-ray images, half of them containing a 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. 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 closed packed bags with a higher degree of bag complexity. In addition, each bag was presented once with and once without the 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.

2.3.3 L/T Letter Search Task

Comparable to previous research using laboratory visual search tasks, an L/T-letter search task was created to evaluate visual search abilities that are independent of a specific domain (in accordance to Biggs et al., 2013). The test consisted of 96 trials. Each image comprised 25 pseudo-Ls as distractors, and half of the images contained one target T against a grey background. The items were randomly located in a 8x7 grid. Each item comprised of two perpendicular black lines that varied in six levels of transparency (70%, 67%, 65%, 40%, 35%, 30%) and four levels of rotation. Target Ts had a crossbar directly in the middle, whereas distractor Ls had a crossbar slid to variable distances away from the center. The distractor stimuli were variable in shape with some very close to the target Ts. This was done to increase the tasks 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.

2.4. Procedure

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

For the VCTB, all tests were conducted computer-based and not in the original paper-and-pencil form. Each of the 10 subtests started with general instructions followed by an example. The same accounted for the SPM that followed the VCTB scales. The test was divided into three blocks and participants were asked to take a break of 10-15 minutes in between. For the SBST participants came to the testing facilities again, approximately two weeks later. 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 luggage was presented for maximally 4 seconds. This duration was chosen to match the demands of high passenger flow where average X-ray image inspection time at checkpoints is in the range of 3-5 seconds. Participants' task was to decide as accurately and as quickly as possible whether the bag was OK (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 at test were presented for ten seconds, respectively. Feedback was provided after each trial but only in the introductory phase. For the L/T- letter search task, each trial started with a fixing cross in the middle of the screen. After 0.5 seconds, a grid with 25 stimuli was presented for maximally 15 seconds. 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 seconds, the next trial started.

2.5. Analyses

Both tasks used in this experiment can be described as a visual inspection consisting of visual search and decision (Koller, Drury, & Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, & Horne, 2009; Spitz and Drury, 1978). The outcome of this task is based on the searchers decisions on whether a target is present or absent. According to SDT (Green & Swets, 1966), there are four possible outcomes depending on stimuli and participant responses (Table 3).

Our dependent variable of the LT-letter search task (detection performance d') was calculated using the following SDT formulae, whereby z refers to the inverse of the cumulative distribution function of the standard normal distribution (Green & Swets, 1966; Macmillan & Creelman, 2005):

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

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 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 signal-plus-noise (target-present) and noise (target-absent) distribution, the resulting *z*ROC has slope *s*. For this SDT model, Macmillan and Creelman (2005) propose using Simpson and Fitter's (1973) detection measure:

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

Concerning the task of X-ray screening, many studies have brought up doubts on the equal variance Gaussian model. This makes sense as the prohibited items that have to be detected vary and therefore bring additional variation into the X-ray image. Therefore, 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 (Godwin, Menneer, Cave, & Donnelly, 2010; Van Wert, Horowitz, & Wolfe, 2009) reported zROC slopes similar to those reported by Wolfe et al. (2007) while a study reported by Wolfe & Van Wert (2010) found a slope of 0.56 and a study by Sterchi, Hättenschwiler & Schwaninger (n.d.) a slope of 0.5 to fit the data more accurately. In our study, data from the basic visual search task (L/T search task) were analyzed under the assumption of an equal variance model using *d*' whereas data from the X-ray screening task SBST was analyzed under the assumption of an unequal variance model with a zROC slope of 0.5 using *da*.

In a first step, we examined descriptive statistics (mean and standard deviation) as well as correlations (Spearman correlations; Spearman, 1927) with basic functions of R Statistics version 3.4.4 (R Core Team, 2015). 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, 2015). We report factor loadings of CFA, which should be minimally 0.50 and optimally be higher than 0.70. For the estimation of the goodness of fit for the models we report Chi2 value, 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 & 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 we report R^2, F value and p-value to evaluate the overall model fit. Furthermore we report β, se, t-value and p-value for each predictor.

3. Results

First, descriptive statistics and spearman correlations are reported. In accordance with the CHC-model of intelligence (e.g. Flanagan & Dixon, 2014) we then computed a confirmatory factor analysis 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. Last, we performed multiple regression analyses to test whether the z-standardisized summarized scale scores of Gs, Gms and Gv could predict performance in the tradition LT visual search task and the X-ray image inspection task.

3.1. Descriptive statistics and correlations

Table 4 shows means and standard deviations of all independent (Gs, Gv and Gsm) and dependent variables (da ORT, RT ORT, d' LT, RT LT) for students and professionals. Spearman correlations between all variables separated for students and professionals are shown in Table 5. Correlations with SPM scores served as a control variable 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 LT search task and ORT with the VCTB measures *Gv*

and *Gsm* were all statistically significant within both populations. *Gs* correlated with detection performance of the LT search task for professionals and with the X-ray image inspection task for students. The inter-correlations 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 ORT da. These are expected results, since it is known that fluid intelligence, processing speed as well as performance in ORT are decreasing with age. In the student population, we did not find these relations, which could be due to the much younger age and lower range of age in that population.

3.2. 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*). Confirmatory factor analysis (CFA) showed that the theoretical model fits the data well. All factor loadings reached statistical significance ($p < .001$), 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, $p = .005$, CFI = 0.961, TLI = 0.946 and RMSEA = 0.0359. As postulated by the CHC-model, the broad abilities of stratum II are 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, while 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 showed that the construct validity of the VCTB was given. For the 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.

3.3. Multiple-linear regression analyses

In a next step, multiple linear regression analyses were calculated to predict detection performance of the L/T search task and the ORT based on the z-standartisized summarized scale scores of Gs, Gsm and Gv and the group (students vs. professionals). A significant regression equation was found $F(4,235) = 9.64$, $p < .001$, with an adjusted R^2 of .13. zGv was the only significant predictor of detection performance (Table 6). The same analyses was calculated again with group as moderator variable however the moderation did not improve the model fit (adjusted $R^2 = .12$) and the comparison of the two models using the Wald test did not reach statistical significance [$F(3, 232) = 0.14$, $p = .939$].

A significant regression equation was found $F(4,235) = 159.3$, $p < .001$, with an adjusted R^2 of .73. Group, zGsm and zGv were significant predictors of detection performance (Table 7). The same analysis was calculated again with group as moderator variable however the moderation did not improve the model fit (adjusted $R^2 = .73$) and the comparison of the two models using the Wald test did not reach statistical significance [$F(3, 232) = 1.83$, $p = .143$]). To further explore the effect of group, we tested whether work experience of professionals (years: $M = 6.83$, $SD = 5.82$) can explain some variance. However, there was no significant correlation between performance in the ORT 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 $R^2 = .72$).

We also tested whether zGv and zGsm are only predicting detection performance due to their association with general intelligence. While adding SPM as an additional predictor of performance, the model fit did not improve in the LT search task (adjusted R^2 = .12; Wald test: $F(1, 234) = 0.17$, $p = .682$), thereafter removing zGv, zGsm, and zGs reduced the model fit significantly (adjusted R^2 = .05, Wald test: $F(3, 234) = 7.67$, $p < .001$). Likewise, adding SPM as an additional predictor of X-ray inspection performance did not improve the model fit (adjusted R^2 = .73; Wald test: $F(1, 234) = 0.31$, $p = .718$). However, thereafter removing zGv, zGsm, and zGs reduced the model fit significantly (adjusted R^2 = .68, Wald test: $F(3, 234) = 12.22$, $p < .001$).

In a last step, we calculated the same analyses using response times (RT) as dependent variables (Table 8). For the L/T search task, a significant regression equation was found $F(2,235) = 10.95$, $p < .001$, with an adjusted R^2 of .14. zGs and zGv were significant predictors of response times (Table 8). The same analysis was calculated again with group as moderator variable however the moderation did not improve the model fit (adjusted R^2 = .14) and the comparison of the two models using the Wald test did not reach statistical significance [$F(3, 232) = 0.26$, $p = .85$]). For the ORT, the regression equation was also significant $F(4,235) = 12.74$, $p < .001$, with an adjusted R^2 of .16. Group and zGv were significant predictors of response times (Table 8). Using group as moderator variable did improve the model fit (adjusted R^2 = .18) and the comparison of the two models using the Wald test did not reach statistical significance [$F(3, 232) = 2.37$, $p = .07$]). Again, to further explore the effect of group we entered work experience as an additional variable what did not improve the model fit (adjusted R^2 = .18).

4. 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 that the results do extrapolate to real-world inspection where people search their visual fields for targets that are more complex, ambiguous and less salient (e.g., Radvansky & Ashcraft, 2016). To answer to what degree results from a traditional visual search task are comparable to an X-ray image inspection task and vice versa, both tasks and two populations were examined. We used a theoretical model with three known facets of visual-cognitive abilities (visual processing *Gv*, short-term memory *Gsm* and processing speed *Gs*) and tested students and professionals using the same experimental stimuli. Based on our results, the following research questions shall now be answered: (1) Do the same 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?

Separate multiple linear regression analyses were calculated for both visual search tasks, in order to predict performance based on Gv, Gsm, Gs and group. Taken together, it was shown that Gv is an overall predictor of detection performance and response times while Gs solely predicts response times. According to the CHC-theory, visual processing (Gv) describes a broader ability to perceive, analyze, synthesize, and think with visual patterns, including the ability to store and recall visual representations. Both, the LT search task and the ORT require visual processing abilities, i.e. 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). Participants with higher Gv scores therefore performed faster and better. As

mentioned in the introduction, there is a content-related overlap between visual processing (Gv) and short-term memory (Gsm). Also within our data, the coefficient for Gsm was the same for the LT-search task and the ORT, but Gsm only reached significance as predictor for the ORT (due to higher variance). 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 stimuli of the LT-search task and the ORT, one would assume that Gsm might be especially important for an inspection task like an ORT, which uses more complex and realistic stimuli and needs more top-down processing and the use of memory capacity while simple letters can be easily reminded. It can be further assumed that short-term memory is becoming even more important when predicting performance in tasks with increasing complexity and unknown features that need previous knowledge.

Processing speed, the ability to quickly and accurately perceive visual details, similarities and differences could predict response times in the LT letter search task and came close to significance in the ORT ($p = .051$). Participants with higher Gs scores therefore performed faster. This result was anticipated as processing speed was found to be relevant in terms of efficiency (Salthouse, 1996). Having a look at the correlational analyses revealed that there was a speed-accuracy tradeoff for the LT search task (higher RT lead to more accuracy), i.e. students and professionals who were looking longer for a target were also better at detecting it. Students also had a speed-accuracy tradeoff in the ORT while professionals did not. This result reveals that speed alone cannot be accountable for screener's performance in the ORT.

Further, correlational analyses revealed that the scales of the VCTB highly correlate with SPM, a known measure for fluid intelligence (Raven, Styles, & Raven, 1998). We were therefore interested in whether Gv and Gsm are only predicting detection performance due to their association with general intelligence. If this would be the case, then SPM could serve as the only predictor of visual search performance. However, replacing Gv, Gsm, and Gs with SPM resulted in a worse model fit, for both ORT and the letter search task. It can therefore be argued that the VCTB scales measure additional information which account for visual search performance.

Taken together, we can answer our first research question, on whether the same visual-cognitive abilities can predict performance in a traditional visual search task and an X-ray inspection task with "yes". Both tasks created for this study can be described as a form of a conjunction search tasks where targets differ from distractors while having some similar features. The literature further suggests that visual search with complex objects (e.g. X-ray image inspection) appears to rely on the same processes as conjunction search (e.g. LT-search) with less complex, contrived stimuli (Alexander & Zelinsky, 2011; 2012). In our study, the tasks were created in a similar way, to make them comparable. If a more simplified traditional visual search task or an X-ray inspection task with domain-specific knowledge is used as comparison, a direct translation of the underlying abilities has to be taken with caution. Implications based on current results could be that either a simple, short and therefore more ecologically version of the visual-cognitive test battery (Gv and Gs scales) could be used to measure abilities and predict performance in students and professionals. Or in an applied setting, the X-ray ORT can be used as criterion for abilities. As there are big individual differences in visual-cognitive abilities, it should be tested whether someone is suited to perform highly in a visual search and inspection task. Especially in regard to X-ray screening, airports could conduct pre-employment assessments that test for certain

visual abilities and aptitudes when recruiting new personnel. However, we believe that visual-cognitive abilities might get less important as predictor for performance for tasks where domain-specific knowledge is necessary. For example, when radiologist search for cancer in mammograms or screeners search for improvised explosive devices that include unknown features, training for these features should have a higher 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 with repeated exposure to visual search tasks.

To answer our second research question, whether results differ between students and professionals, we could show that the tested populations are comparable as the moderation effect of group did not improve the model fit. We found the same performance and response times for both populations in the LT-search task, while professionals performed better and faster on the X-ray inspection task. We could show that the visual-cognitive abilities tested in this study have a comparable influence on performance and response times independent of the tested population, why the group effect cannot be due to differences in these aptitudes. Also, while we found a speed-accuracy tradeoff for both tasks within the student population, professionals did not show this effect for the ORT. Alternative explanations for this group effect could either be due to differences in the task or differences in the baseline experience of the tested populations. Firstly, even though we used comparable tasks based on conjunction search, the objects and features did still differentiate. To have a fair comparison, we created a traditional visual search task with a high difficulty and an X-ray image inspection task with no need of domain-specific knowledge. Features of guns and knives, as well as letters like L or T are known from everyday life experience and can therefore be detected without specific experience and training. As group differences only occurred for the ORT, a difference in the populations conducting the tasks might be the better explanation. Even though there is no need for domain-specific knowledge when conducting the ORT, professionals still performed better and faster and with no speed-accuracy tradeoff. The professionals already had years of training and experience with X-ray image inspection and therefore a lot of domain-specific knowledge in this area. However, when including work experience as an additional variable, the model fit did not improve. Therefore, the amount of work experience per se did not seem to be the responsible variable of the group effect. However there is still a baseline difference between students and professionals regarding the familiarity with objects in the ORT. Based on the literature, more familiar objects possibly need fewer recognized features in order to be successfully identified (Koller et al., 2009) and features are known and recognized better and faster with repeated exposure (Halbherr et al., 2013; Koller et al. 2008; Koller et al., 2009; McCarley et al., 2004; Schwaninger & Hofer, 2004). Interestingly, despite the work experience of professionals, the visual-cognitive abilities were still relevant for a high performance and fast response times. It therefore seems that these abilities cannot be completely compensated through training and work experience. Another explanation could be that the professionals participating in this study all passed a pre-employment assessments test for these visual abilities and aptitudes (e.g. X-Ray Object Recognition Test; see Hardmeier et al., 2005; Hardmeier & Schwaninger, 2008). It could therefore be possible that they already have higher visual-cognitive abilities than the average population and show a smaller variance in the VCTB scales. Therefore results might be different if professionals were tested that had not been recruited based on a baseline assessment. Lastly, the students tested in our study revealed to be a very heterogeneous sample what’s not directly

comparable to a typical university-student's sample. The question therefore arises what would happen with the regression results if the tested sample is more homogenous on some variables.

Based on our results, we can therefore answer our second research question with "no" as the results of our study did not differ between students and professionals. This leads to the assumption that students as tested in our study can be used as a sample to make predictions of professionals. Obviously, conducting an X-ray ORT is not the same task as screeners conduct it at checkpoints, in particular regarding target prevalence, coloring of images and target categories. We therefore believe that results would look different if a task is used where domain-specific knowledge is needed, as students would lack the knowledge of specific features to perform well in this kind of visual inspection task. Prohibited items which are rather uncommon or have not been seen before (e.g. improvised explosive devices, IEDs) become very difficult to recognize without training to recognize certain features of these threats (Schwaninger, 2004a; Schwaninger, 2005a). 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, 1996). Depending on the similarity of objects and its features presented in an X-ray image to those stored in visual memory, the searcher will then decide whether the respective object is harmless or not. As students lack this visual memory, they would not perform well in such a task. The implications based on our second research questions have therefore to be taken with caution as the comparability of the two populations is dependent on the task.

Future directions could be to test another population with more heterogeneous visual-cognitive abilities to get more variance. Further challenges of the comparison of X-ray image inspection and traditional visual search include low target prevalence, variation in target visibility, an unknown target set and the possible presence of multiple targets (for recent reviews see Biggs & Mitroff, 2014; Mitroff, Biggs, & Cain, 2015). Tasks could therefore be compared with a variation in target prevalence or in task difficulty (e.g. searching for multiple targets).

5. Conclusion

With this study we tried to answer to what degree results from a traditional visual search task can be translated to an X-ray image inspection and vice versa and if students and professionals are comparable. Based on the tasks and populations tested in this study, results can be seen as transferrable to a certain point. The understanding that a traditional visual search task without domain-specific knowledge can be generalize from students to professionals can be a huge advantage, allowing for certain applied studies to be run with relatively easier accessible populations. Further, comparing visual-cognitive abilities and their influence on visual performance and response times showed that the same specific visual-cognitive abilities were able to predict performance and response times in both tasks. As the same cognitive processes are underlying the tested tasks, future searchers can be differentiated based on some very specific abilities.

6. Acknowledgments

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

Author Contributions Statement

All authors substantially contributed to the conceptualization of the manuscript as well as to the aquisition, 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 aquisition, 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.

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.

References

Alexander, R., & Zelinsky, G. (2011). Visual similarity effects in categorical search. *Journal of Vision. 11(8*): 9.

Alexander, R., & Zelinsky, G. (2012). Effects of part-based similarity on visual search: The Frankenbear experiment. *Vision Research. 54*: 20–30.

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.

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

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.

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. *HCI and Usability for Education and Work, Lecture Notes in Computer Science, 5298,* 419-438. doi:10.1007/978-3-540-89350-9_30

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,* Zurich Switzerland, October 5-8.

Bravo, M. J. & Farid, H. (2004). Recognizing and segmenting objects in clutter. *Vision research, 4,* 385-396.

Carroll. J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: 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: Tribute to Arthur R. Jensen* (pp. 5–22). San Diego: Pergamon.

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

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* (Vol. 59, pp. 147–181). New York: Springer.

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

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96(3*), 433-458.

Eckstein, M. (2011). Visual search: A retrospective. *Journal of Vision, 11(5),* 1–36.

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. Doi:10.3758/BF03198804

European Commission (2015). Commission Implementing Regulation (EU) 2015/1998. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R1998&from=DE

Flanagan, D. P., & Dixon, S. G. (2014). The Cattell-Horn-Carroll theory of cognitive abilities. *Encyclopedia of Special Education*. Published online 22 Jan 2014. doi:10.1002/9781118660584.ese0431

Flanagan, D. P., & Harrison, P. L. (2005). *Contemporary intellectual assessment: Theories, tests, and issues. (2nd Edition).* New York, NY: The Guilford Press

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. doi:10.1016/j.actpsy.2009.12.009

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

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM) (1 ed.)*. Thousand Oaks, CA: Sage.

Halbherr, T., Schwaninger, A., Budgell, G., & Wales, A. (2013). Airport security screener competency: a cross-sectional and longitudinal analysis. *International Journal of Aviation Psychology, 23(2)*, 113-129.

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. *IEEE ICCST Proceedings*, *39*, 189-192.

Hardmeier, D., & Schwaninger, A. (2008). Visual cognition abilities in x-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008*, Fairfax, Virginia, USA, June 1-4, 2008, 311-316.

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, ICRAT 2006*, Belgrade, Serbia and Montenegro, June 24-28, 2006, 393-397. doi:10.5167/uzh-97986

Helmholtz, H. (1896/1989). *Physiological Optics* (1896 - 2nd German Edition, translated by M. Mackeben, from Nakayama and Mackeben, Vision Research 29:11, 1631 - 1647, 1989)

Hoffman, J. E. (1975). Hierarchical stages in the processing of visual information. *Perception & Psychophysics, 18*, 348-354.

Hoffman, J. E. (1978). Search through a sequentially presented visual display. *Perception & Psychophysics, 23*, 1-11.

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).* Deutsche Bearbeitung und Normierung nach J. C. Raven, 2nd ed. Frankfurt: Pearson Assessment.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1),* 1-55.

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 758-799). New York, NY: Guilford Press.

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:10.1080/00140130802526935

Koller, S., 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

Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology, 7,* 341–370

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

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

Krupinski, E. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic Radiology, 3(2),* 137 – 144.

Kumada, T., & Humphreys, G. W. (2002). Cross-dimensional interference and cross-trial inhibition. *Perception and Psychophysics, 64*, 493-503.

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

Nodine, C. F, & Kundel, H. L. (1987). The cognitive side of visual search in radiology. In: O'Regan, J. K., Levy-Schoen, A., eds. *Eye movements: from physiology to cognition*. North-Holland: Elsevier Science; 573–582.

Lavie, N., & DeFockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin & Review, 12 (4),* 669–674.

Luck, S. J. & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Martin, N. A. (2006). *Test of visual perceptual skills (TVPS-3), 3rd ed*. Novato, CA: Academy Publishers.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport screening. *Psychological Science, 15*, 302–306.

McElree, B; Carrasco, M (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *Journal of experimental psychology. Human perception and performance. 25(6),* 1517–39.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp.136–182). New York: Guilford.

Mitroff, S. R., Biggs, A. T., & Cain, M. S. (2015). Multiple-target visual search errors: Overview and implications for airport security. *Policy Insights from the Behavioral and Brain Sciences, 2(1),* 121-128.

Nakayama, K., & Martini, P. (2010). Situating visual search. *Vision Research*. doi: 10.1016/j.visres.2010.09.003.

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton, Century, Crofts.

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

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14620349

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. *The Quarterly Journal of Experimental Psychology, 62(7),* 1430–1454.

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

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.

Reavis, E.A., Frank, S.M., Greenlee, M.W., & Tse, P.U. (2016). Neural correlates of context-dependent feature conjunction learning in visual search tasks. *Human brain mapping, 37 (6*): 2319–30.

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

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

Roper, Z. 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.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48(2),* 1-36.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103(3),* 403-428.

Schwaninger, A. (2005a). *Objekterkennung und Signaldetektion*. In B. Kersten (Ed.), Praxisfelder der Wahrnehmungspsychologie (pp. 108-132). Bern: Huber.

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

Schwaninger, A. (2006). Airport security human factors: From the weakest to the strongest link in airport security screening. *Proceedings of the 4th International Aviation Security Technology Symposium,* Washington, D.C., USA, November 27 – December 1, 2006, 265-270.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.

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

Shen, J., Reingold, E. M, & Pomplun, M. (2003). Guidance of eye movements during conjunctive visual search: the distractor-ratio effect. *Canadian journal of experimental psychology, 57(2),* 76–96.

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

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

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:10.1177/001872087802000502

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42(5),* 893-98.

Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (n.d.). Detection Measures for Visual Inspection of X-ray Images of Passenger Baggage. Under Review.

Treisman, A. (1988). Features and objects: the fourteenth Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology, 40A,* 201–236.

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. *Cogn Psychol. 12(1)*: 97–136. doi:10.1016/0010-0285(80)90005-5

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: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:10.3758/BRM.41.3.937

Watson, D. G., & Humphreys, G. W. (1997). Visual marking: Prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review, 104*, 90-122.

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

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*, 202–238.

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

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences, 7*, 70–76.

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.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology Human Perception Performance, 15(3),* 419-33.

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. *Journal of Experimental Psychology: General*, *136*(4), 623–638. doi:10.1037/0096-3445.136.4.623.

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

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:10.1016/j.cub.2009.11.066

Tables

Table 1

Description of Participants in the Experiment

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

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

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	Standartisized	Cronbach's α: .86-.94 Split-Half: .81-.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*)	Standartisized	Cronbach's α: .74 Test-Retest: .71	-
SPM	SPM: Speed-Test	Standartisized	Cronbach's α: .97-1.00 Split-Half: > .90 Test-Retest: .80-.90	Correlations with nonverbal IQ

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

Table 3

Definition of Hit, False Alarm, Miss, and Correct Rejection According to SDT (Green & Swets, 1966).

Stimulus	Target-present response	Target-absent response

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

Table 4

Means and Standard Deviations

		Students			Professionals		
	Max. Score	*n*	*M*	*SD*	*n*	*M*	*SD*
da ORT	3.5	128	1.6	0.3	112	2.6	0.4
RT ORT	4.0	128	3.2	1.1	112	2.6	0.7
d' L/T	3.5	128	1.0	0.5	112	1.0	0.5
RT L/T	15.0	128	8.1	1.3	112	8.2	1.4
Gs	116	128	80.9	13.7	112	64.2	16.6
Gv	48	128	37.4	5.1	112	36.3	6.2
Gsm	31	128	21.7	5.5	112	19.8	5.7

Note. *n* = number of participants; *M* = mean; *SD* = standard deviation

Table 5

Correlational Analyses

Students

	da ORT	RT ORT	d' L/T	RT L/T	SPM	Gs	Gsm	Gv
da ORT	-							
RT ORT	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

	da ORT	RT ORT	d' L/T	RT L/T	SPM	Gs	Gsm	Gv

da ORT	-							
RT ORT	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

Note. Spearman Correlations; * indicates $p < 0.05$; ** indicates $p < 0.01$ and *** indicates $p < 0.001$.

Table 6

Multiple Linear Regression Analyses for the LT letter search task

	β	*SE β*	*t-value*	p-value
zGs	-0.013	0.078	-0.164	0.870
zGsm	0.119	0.079	1.513	0.132
zGv	0.299	0.078	3.830	0.000***
Group	0.058	0.139	0.416	0.678

Note. * indicates $p < 0.05$; ** indicates $p < 0.01$ and *** indicates $p < 0.001$.

Table 7

Multiple Linear Regression Analyses for X-ray inspection task

	β	*SE β*	*t-value*	p-value
zGs	-0.039	0.044	-0.893	0.373
zGsm	0.104	0.044	2.348	0.019*
zGv	0.195	0.044	4.463	0.000***
Group	1.668	0.078	21.370	0.000***

Note. * indicates $p < 0.05$; ** indicates $p < 0.01$ and *** indicates $p < 0.001$.

Table 8

Multiple Linear Regression Analyses with Response Times (RT)

LT letter search task	*β*	*SE β*	*t-value*	p-value
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 inspection task	***β***	***SE β***	***t-value***	**p-value**
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***

Note. * indicates $p < 0.05$; ** indicates $p < 0.01$ and *** indicates $p < 0.001$.

Figures

Figure. 1. Three examples of X-ray images from the SBST.

Figure. 2. Example of an image from the L/T-letter search task. Image containing several pseudo-Ls as distractors, and one target T against a grey background.

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

Yanik Sterchi, Nicole Hättenschwiler, and Adrian Schwaninger

University of Applied Sciences and Arts Northwestern Switzerland

Author Note

Yanik Sterchi, Nicole Hättenschwiler, and Adrian Schwaninger, University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Olten, Switzerland.

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

Email: yanik.sterchi@fhnw.ch, Phone: +41 62 957 27 08

Abstract

In visual inspection tasks, such as airport security and medical screening, researchers often use the detection measures *d′* and *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

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

X-ray baggage screening at airports is an essential component for securing air transportation. To prevent passengers from bringing potential threats on to 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).

Table 1

Outcome of Decisions Depending on Stimulus Using the Terminology of Visual Search, Signal Detection Theory, and X-ray Baggage Inspection

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

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 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 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 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 Figure *1*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 Figure 1a and Figure *1*d). SDT assumes that the criterion can be shifted, with a *liberal* criterion resulting in a higher HR and FAR; and a *conservative* criterion, 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

$$d' = z(HR) - z(FAR), \quad (1)$$

where *z* is the inverse of the cumulative distribution function of the standard normal distribution (Green & Swets, 1966). The receiver operating characteristic (ROC) curve (Figure 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, *zROC*), they form lines with slope 1 and *d'* as their intercept (Figure 1b).

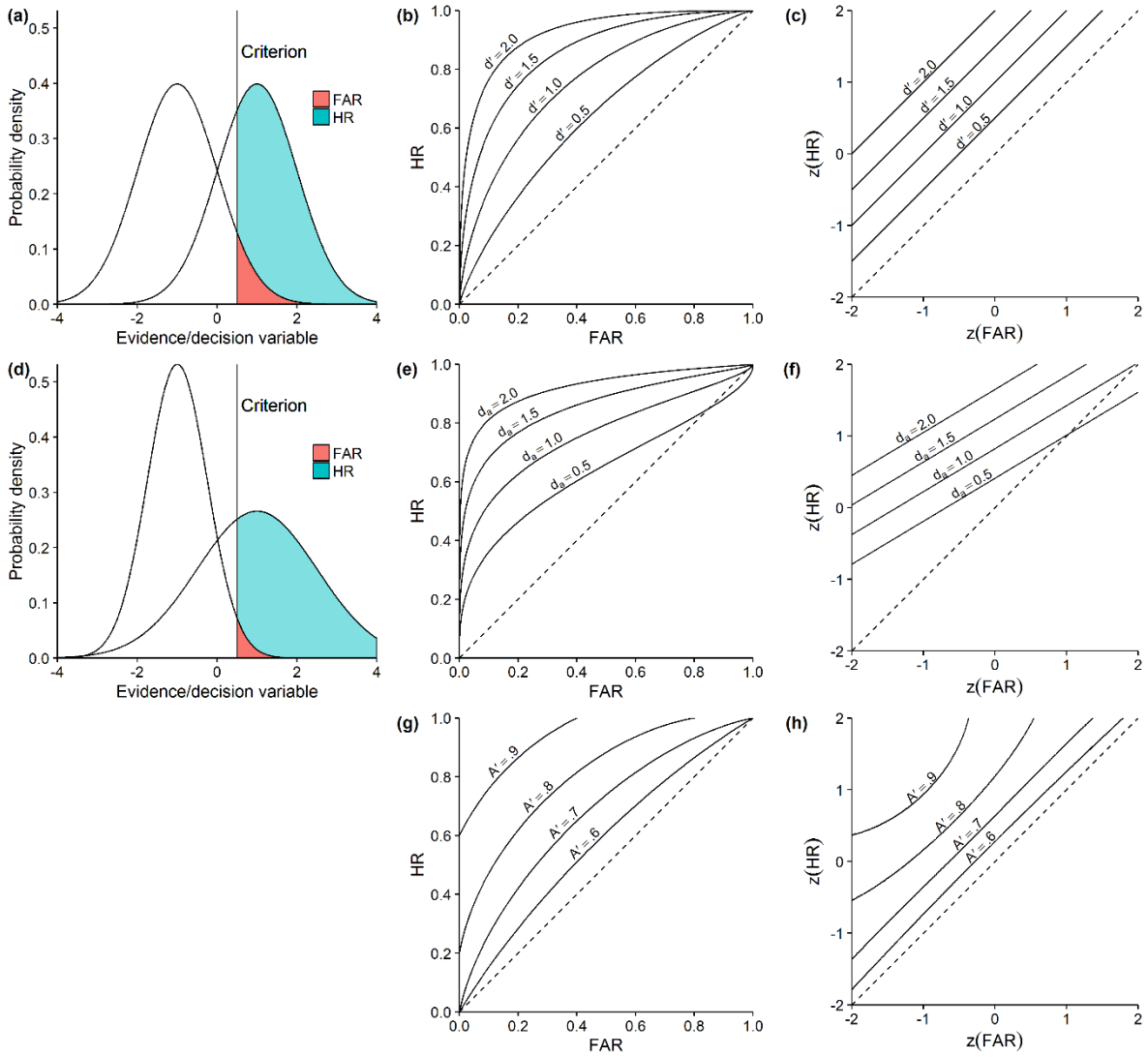

Figure 1. Illustration of noise and signal-plus-noise distribution (first column), receiver operating characteristic (ROC) curves (second column), and ROC curves in *z*-transformed space (zROC; third column) corresponding to *d'* (first row), d_a (second row), and *A'* (third row).

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:

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

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:

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

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 curve shown in Figure 1.

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 about 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 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 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 (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 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 reduced HR and FAR in low target prevalence conditions with averaged results seeming to lie on a *z*ROC line with a slope of 0.6. Two further publications (Godwin, Menneer, Cave, & Donnelly, 2010; Van Wert, Horowitz, & Wolfe, 2009) reported *z*ROC 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 *z*ROC slope of 0.6 indicates that the target-absent (noise) distribution has a smaller standard deviation than the target-present (signal-plus-noise) distribution. This is generally plausible because the prohibited items that have to be detected vary and therefore bring additional variation into the X-ray image. Consistent with this assumption, several studies have provided converging evidence that detection performance is influenced by several factors that possibly increase variability in X-ray images containing a prohibited item. Not only prohibited item categories (e.g., guns, knives, improvised explosive devices, and other prohibited items; see Halbherr et al., 2013; Koller et

al., 2009, 2008) but also different items within one category vary in difficulty (Bolfing, Halbherr, & Schwaninger, 2008; Schwaninger, Michel, & Bolfing, 2007).

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 *z*ROC space, as assumed by both *d'* and *A'*, and to estimate the *z*ROC 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, 2010; 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 ROC shape, 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 Figure 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.

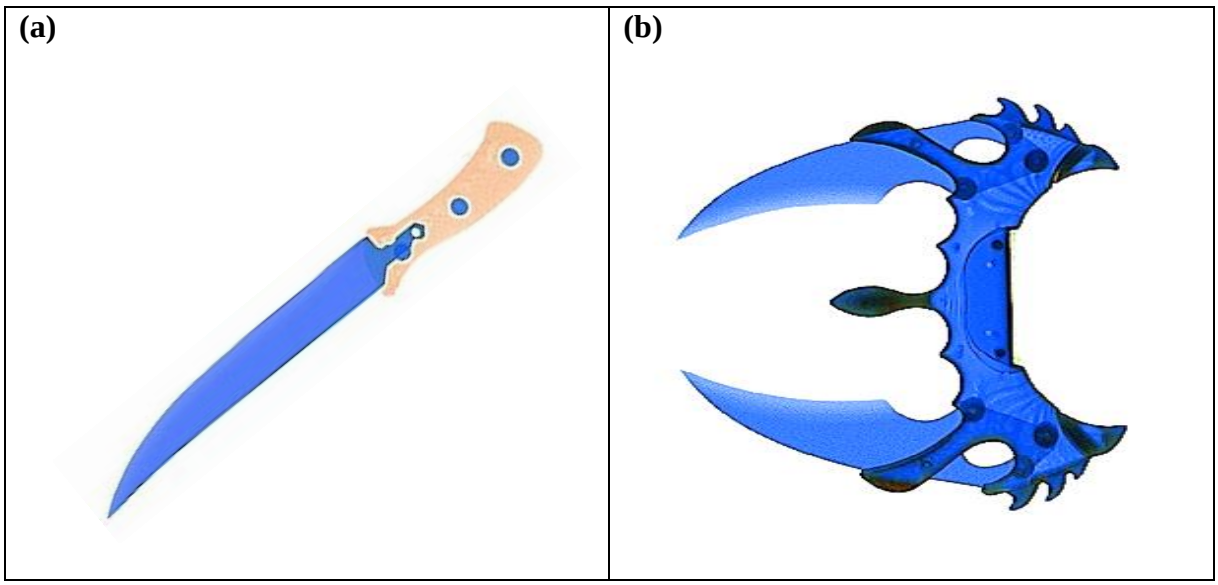

Figure 2. Two examples of the prohibited item category *knife*: (a) example of a known target item and (b) example of a novel target item (Asian combat knife).

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-inch TFT LCD screen with a resolution of 1,280 × 1,024 pixels placed 50 to 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, 10 practical 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.

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, 2010; Wolfe et al., 2007; Wolfe & Van Wert, 2010). For zROC 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*, this is, the rejection rate. A manipulation check revealed an effect of the instruction on the rejection rate with a Cohen's *d* of 0.58. However, 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. 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.

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 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.

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.

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)

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 size of d' and d_a were identical.

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]

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) 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).

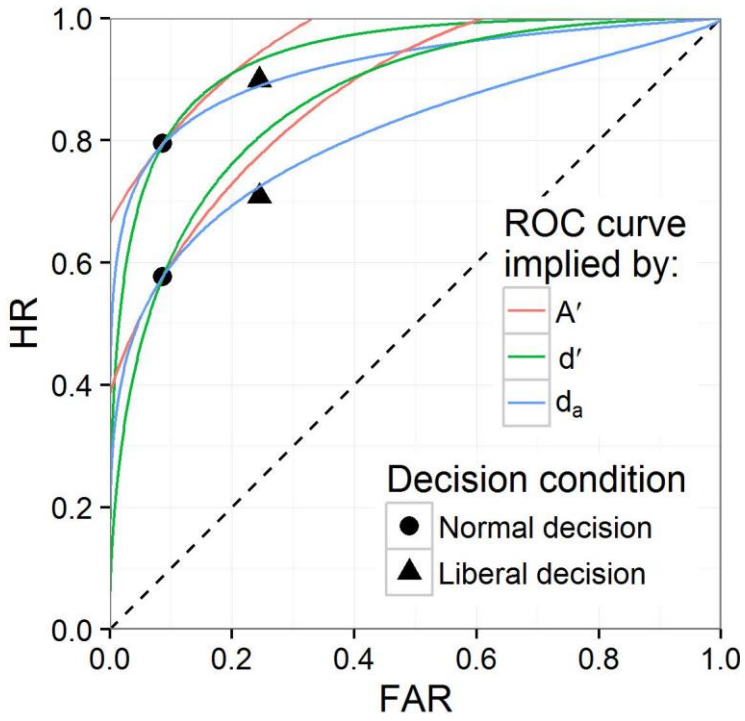

Figure 3. ROC curves implied by d', A', and d_a estimated by the pooled hit rate (HR) and false alarm rate (FAR) of the normal decision condition for known prohibited items (higher HR) and novel prohibited items (lower HR).

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 Figure 3, the resulting four points defined by the pooled HR and FAR fit well 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, 2010; 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, 2010; 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 hit and false alarm rate 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, 2010; 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'.

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, estimations of the slope parameter with the data from Experiment 1 resulted in large confidence intervals. 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 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 (gun, knife, IED, or bare explosive). Each item occurred twice, depicted from two different viewpoints. 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).

Procedure

The participants were tested in groups of maximally six screeners at a time. To become familiar with the test, the screeners performed eight practice trials. The test did not provide feedback on the correctness of responses. 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. 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*).

Data Analysis

For each participant, the HR and FAR were calculated for the different levels of confidence rating according to (Macmillan & 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, & Shen, 1998). The slope parameter as a ratio is inappropriate for directly calculating a mean because steep slopes result in large numbers (a horizontal *z*ROC has a slope of zero and a vertical

*z*ROC has a slope of infinity). 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 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).

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.

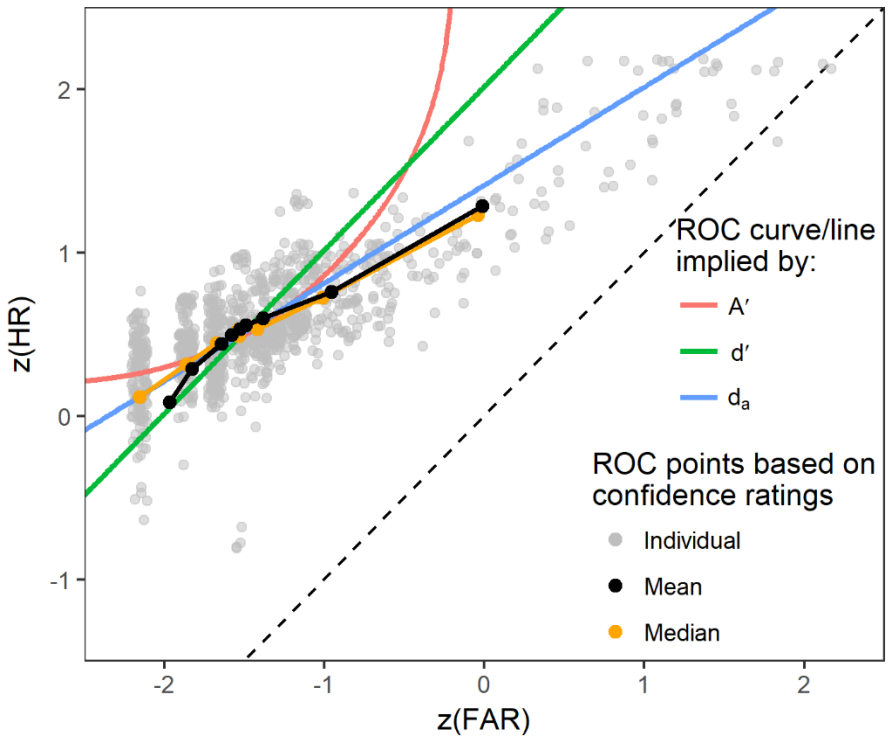

Figure 4. Individual (grey; jittered) and pooled (black) empirical *z*ROC curves, the lines corresponding to the mean *A'*, *d'*, and d_a with a slope of 0.6, and the chance line (dashed).

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 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 estimated using the LABROC3 algorithm (Metz et al., 1998) are illustrated in *Figure 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]).

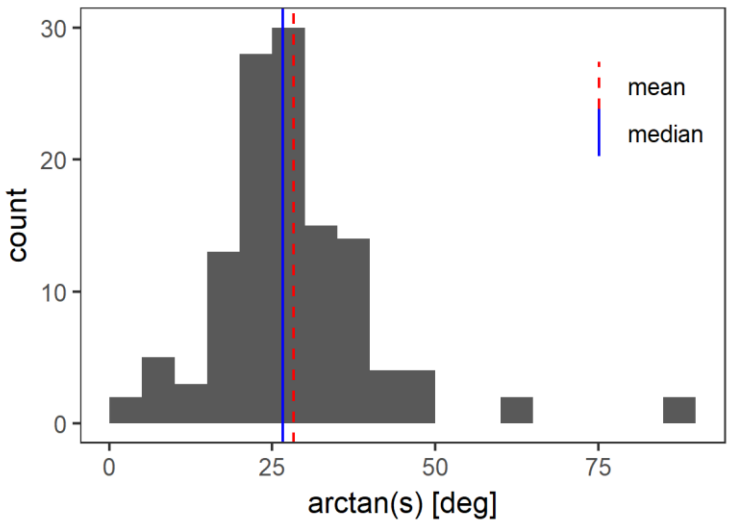

Figure 5. Distribution, mean (red dashed line), and median (solid blue line) of arctan-transformed individual slope parameters.

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, 2010; Wolfe et al., 2007) and another study that reported a slope of 0.56 (Wolfe & Van Wert, 2010).

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 concerns only values very close to the limits of the ROC space. However, for low sensitivity (e.g., for difficult items or unexperienced 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 Figure 1e). It would therefore be reasonable to assume that the *z*ROC slope converges to unit slope with decreasing sensitivity. Such a convergence has been found repeatedly in research on recognition memory (Brown & Heathcote, 2003; Glanzer, Kim, Hilford, & 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, 2010; 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., 2010; 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. We discuss in the next section, 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, 2010; 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 that 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 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, 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 et al., 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, which 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 *z*ROC slope of 0.5 to 0.6 often seems to provide the better measure. However, while it is not clear what factors determine the *z*ROC slope, we recommend testing d_a with a slope of 0.5 in addition to d_a with a slope of one (i.e., d') as the upper and lower bound, respectively. Another approach is to gather confidence ratings and use A_g as 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

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

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 *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'). Collecting confidence ratings allows to directly estimate the *z*ROC 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.

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: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: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. *HCI and Usability for Education and Work. 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008,* 419–438. doi: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:10.3758/BF03195493

Brunstein, A., & Gonzalez, C. (2011). Preparing for novelty with diverse training. *Applied Cognitive Psychology*, *25*(5), 682–691. doi: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: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: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:10.1518/155723408X299843

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. doi: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: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: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:10.1037/0278-7393.25.2.500

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. doi: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. doi: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: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: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: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: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:10.1109/CCST.2004.1405409

Huang, L., & Pashler, H. (2005). Attention capacity and task difficulty in visual search. *Cognition*, *94*(3), B101–B111. doi: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: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: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: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: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: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: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: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: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: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: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: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: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:10.1037/h0026862

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1), 125–126. doi: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: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: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: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: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: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:10.1024/1662-9647/a000020

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in X-ray screening using image measurements. *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80*(6), 481–488. doi: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: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: 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: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: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: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: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: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: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: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:10.1080/00140139.2014.993429

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 A1 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 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 A2 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 A3 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 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 dataset 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 A3 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).

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 *z*ROC 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.

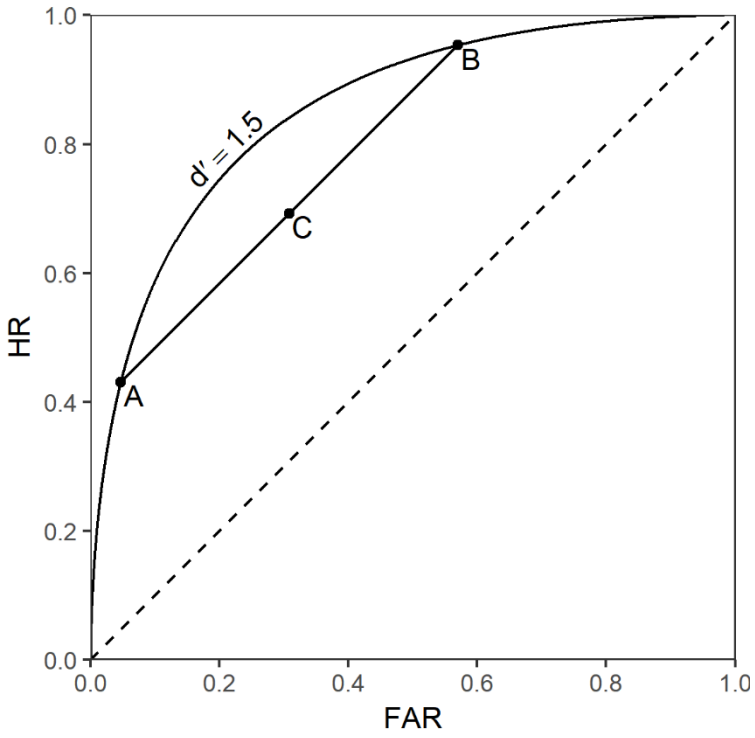

Figure A1. When the two points A and B from the same receiver operating characteristic (ROC) curve are averaged, the resulting ROC point C is below the original ROC curve.

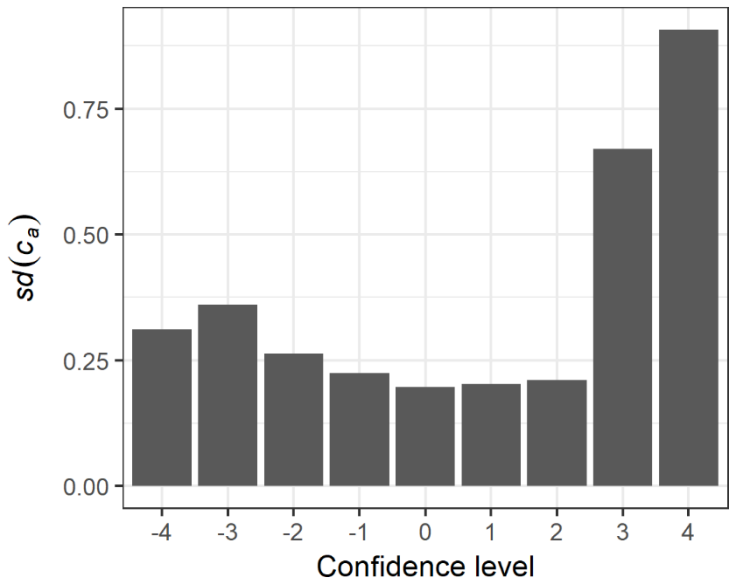

Figure A2. Between subject standard deviation of c_a (based on a slope of 0.6) for each confidence level.

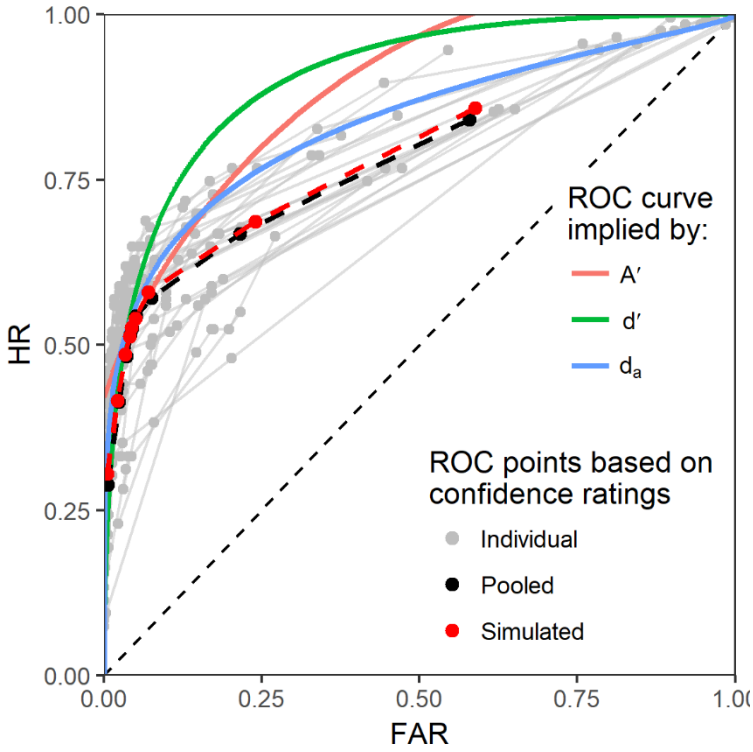

Figure A3. Receiver operating characteristic (ROC) points based on individual (grey) and pooled confidence rating data of dataset 2 (black, dashed), created from a simulation (red, dashed), as assumed by the average d' (green), d_a (blue), and A' (red).