Increasing X-ray image interpretation competency of cargo security screeners

Stefan Michel, Marcia Mendes, Jaap C. de Ruiter, Ger C.M. Koomen, Adrian Schwaninger

1. Introduction

In the last decade, airport security has been substantially enhanced in the areas of passenger, cabin baggage and hold baggage screening. However, air cargo, whether transported on passenger planes or commercial cargo aircrafts, has not yet received a similar level of attention. Nevertheless, the explosion of a passenger plane or commercial cargo aircraft, has not yet received a similar level of attention. Nevertheless, the explosion of a passenger plane or commercial cargo aircraft could highly impact the world economy, commerce or the global supply chain. At least since 2010 it is clear that cargo is an attractive target for terrorists (TIME, 2010). Consequently, X-ray cargo screening for the inspection of unit load devices (ULDs) and containers is emerging quickly and becoming a common feature at ports and airports. A big advantage of X-ray screening is that images of the contents of a ULD or container can be created quickly, without causing any physical changes within the containers. Concerning X-ray screening technology, remarkable improvements have been achieved over the last decades. For the field of cargo X-ray screening, modern X-ray inspection systems especially tailored to the examination of containers, trucks and rail cars exist, which produce images similar to those obtained through traditional X-ray baggage screening machines at airports (e.g., Reed, 2008, 2009).

Despite the advances in technology, the actual decision whether an X-ray image of a container or ULD contains a prohibited item or not still needs to be taken by a human operator, i.e. a cargo security screener. Cargo X-ray screening, in particular, must be considered an extremely challenging task. The large sizes of containers make it very difficult for the cargo security screeners to detect the proportionally small threat items (e.g. an improvised explosive device (IED)) and contraband goods. Moreover, it could be assumed that perpetrators would intentionally disguise or hide their materials or decompose contraband into component parts distributing it within the container. Still, screening officers are expected to assess the...
contents of large containers or ULDs within a few minutes only. Therefore, it is essential to consider human factors in security screening and design systems to support the decision-making of human operators (e.g., Harris, 2002; Kraemer et al., 2009). Selection and training of screening personnel is a highly relevant factor for improving man-machine system performance. The most sophisticated machines become worthless if the people who operate them and visually inspect the X-ray images are not qualified to do so (Al-Fandi et al., 2009; Hardmeier et al., 2005; Schwaninger et al., 2005).

Research on X-ray security screening of passenger baggage at airports has shown that performance of human operators (i.e. airport security screeners) is determined by several factors (e.g., Michel and Schwaninger, 2009; Schwaninger, 2006). First of all, not every person has the potential to become a good screening officer, as certain aptitudes and abilities are required for this task (Hardmeier et al., 2005; Hardmeier and Schwaninger, 2008; Schwaninger et al., 2005). For that reason it is advisable to apply scientifically proven selection tests, such as the X-ray Object Recognition Test (X-Ray ORT), as part of pre-employment assessment procedures, so that right from the beginning only people well suited for the screening job are chosen for this task (Hardmeier et al., 2005; Schwaninger et al., 2005). Furthermore, in order to detect prohibited items reliably, a screening officer must know which items are prohibited and what they look like in X-ray images. Many prohibited items are rarely seen in everyday life and might look quite different from reality when displayed as an X-ray image (e.g. IEDs). According to object recognition theories and visual cognition, object shapes not similar to the ones stored in visual memory are difficult to recognize (Graf et al., 2002; Schwaninger, 2004a, 2004b). Schwaninger et al. (2005) found that the detection of prohibited items is dependent on knowledge-based and image-based factors. Knowledge-based factors refer to knowledge required for the detection of prohibited items, i.e. knowledge on which items are allowed vs. prohibited, what objects look like in reality, and what they look like in X-ray images of passenger bags or containers. This underlines the importance of class-room, computer-based (CBT) and on-the-job training. Image-based factors relate to the difficulty of inspecting an X-ray image. According to Schwaninger (2003) and Schwaninger et al. (2005) three image-based factors are relevant in X-ray screening: viewpoint of an object, superposition of an object by other objects and the complexity of a bag/container. Depending on perceptual experience and the ability to mentally rotate objects, rotated images are more difficult to recognize (effect of viewpoint). Moreover, when prohibited objects are superimposed by other objects (effect of superposition), or when bags are closely-packed and contain many items which could attract attention (effect of complexity), the identification of prohibited items becomes more challenging. Results of different studies could show that coping with image difficulty resulting from these factors can be improved through training (e.g., Hardmeier et al., 2006b; Koller et al., 2008). Yet, a person’s learning process is greatly dependent on visual abilities and cannot merely be attributed to training (Hardmeier et al., 2006b). Regarding knowledge-based and image-based factors, individually adaptive computer-based training (CBT), e.g. with X-Ray Tutor (XRT), is considered to be a very powerful tool for achieving and maintaining a good X-ray image interpretation competency in cabin baggage and hold baggage screening (Halbherr et al., 2013; Hardmeier et al., 2006b; Koller et al., 2008; Schwaninger and Hofer, 2004). For one thing, it allows exposing screening officers to objects they usually do not encounter in everyday life (e.g. IEDs) and for another thing, screeners get trained to identify all sorts of objects in different views (rotations), when superimposed by other objects and in bags (containers) of different complexities.

Considering the objective of increasing container security at ports and airports by improving X-ray screening, the aim of this project was to find out whether the screening officers’ performances in detecting prohibited items in cargo X-ray images can be enhanced through implementing recurrent CBT. Moreover, another essential aspect concerning screening performance is the speed at which an operator performs bag/container searches while maintaining an optimal performance level. Results of previous studies with XRT revealed that training not only increased detection performance of screening officers but also reduced reaction time (Koller et al., 2009; Michel et al., 2007b; Schwaninger and Hofer, 2004; Schwaninger and Wales, 2009; Wales et al., 2009). This is relevant for (cargo) X-ray screening applications where throughput matters.

State-of-the art X-ray screening machines, especially for cabin and hold baggage screening, are able to produce high quality and material coded colored images. Additionally, they offer a variety of so-called “image-enhancement functions” (IEFs) such as color inversion, edge-enhancement, organic only, metal only etc. (for illustrations, see e.g. Michel et al., 2007a). IEFs are applied to bring out details which is obscured or to highlight certain features (e.g., organic content). However, previous studies (Klock, 2005; Michel, Koller et al., 2008) investigating the usefulness of IEFs for the cargo baggage (CBS) and hold baggage screening (HBS) have casted doubt on whether IEFs are useful. Results varied for the different threat types (guns, knives, IEDs and other threat items). The findings of these studies highlight the importance of systematically investigating the usefulness of IEFs in order to optimize human–machine interaction. In order to obtain colored X-ray images, dual energy technology (high and low energy beams) is required (for more details see Fainberg, 1992; Ogorodnikov and Petruin, 2002). For cargo X-ray screening machines, the application of dual energy technology is still difficult and of high costs. In this study, images were recorded with single energy technology, providing greyscale images. As humans can only discern a few dozen grey level values while they can distinguish thousands of colors (Gonzales and Woods, 2002), it could be assumed that the application of color would help to better distinguish between objects. Moreover, color adds vivacity to images und can thus decrease boredom or fatigue and increase attention (Abidi et al., 2005). Color coding could therefore significantly improve detection performance in X-ray screening. In our study, we investigated if showing the original greyscale image in pseudo colored images could improve the detection of prohibited items in cargo X-ray images, and, if so, for which kind of items or images this application was of particular help.

Previous studies focusing on cabin baggage screening have proven the effectiveness of CBT in increasing X-ray image interpretation of screening officers (Halbherr et al., 2013; Koller et al., 2008; Michel et al., 2007b; Schwaninger, 2004b; Schwaninger et al., 2008, 2007; Schwaninger and Wales, 2009). Recurrent CBT is considered to be an indispensable prerequisite for achieving and maintaining good operational performance at security checkpoints. For all airports within the EU, recurrent training consisting of image recognition training and testing either in the form of classroom or computer-based training is obligatory for all persons operating X-ray equipment (European Commission, 2010). For this study, a new CBT system for the area of cargo X-ray screening was created (Cargo-XRT) and tested with cargo security screening officers at an international European airport. The main aim of this study was to investigate if X-ray image interpretation competency of cargo security screeners could be increased by introducing CBT with the C-XRT. Moreover, it was examined whether detection performance is better when pseudo colored images are displayed compared to when the original greyscale images are shown. Further analyses were conducted in order to find out which type of threat items
(security vs. customs related items) were more difficult for screening officers to detect in a ULD or container, as well as how much time screening officers needed to judge an image, and if reaction times could be decreased through training.

2. Method and procedure

2.1. Participants and procedure

Altogether, 40 employees (9 working in administration (4 males) and 31 experienced cargo security screeners (25 males)) participated in this study. The experimental design is shown in Fig. 1. Three groups were formed (two experimental groups and one control group). Administrative personnel working in customs served as a control group. In order to make sure to have equally balanced groups with regard to visual abilities needed for X-ray image interpretation, all participants conducted the X-Ray Object Recognition Test (X-Ray ORT) in advance (see 2.2.1). Based on the X-Ray ORT results, the cargo security screeners were distributed into the two experimental groups. A univariate ANOVA using the detection performance scores $A'$ (Pollack and Norman, 1964) of the X-Ray ORT was conducted to test the equality of all three groups in terms of their visual abilities. No significant difference between the groups (grey scale, pseudo color, control group) was revealed $F(2, 36) = .032, p = .969, \eta^2 = .002$. Both experimental groups underwent weekly recurrent training of at least 20 min per week with C-XRT for approximately three months (resulting in a total of $M = 8.95$ h of training, $SD = 5.68$), while the control group did not conduct any training. Detection performance was measured using the Cargo X-Ray CAT (C-CAT) (see 2.2.3). Measurements were taken before training started and again after three months of training. The two experimental groups conducted different versions of the C-CAT (grey scale X-ray images ($n = 15$) vs. pseudo color X-ray images ($n = 16$)). The control group ($n = 9$) only conducted the greyscale test version (see Fig. 1).

2.2. Materials

2.2.1. X-Ray Object Recognition Test (X-Ray ORT)

As aforementioned, not every person has the potential to become a good X-ray screening officer. Certain aptitudes and abilities are required to cope with the above mentioned three image-based factors (viewpoint, superposition and image complexity) relevant in X-ray screening, which can be measured using the X-Ray ORT (Hardmeier et al., 2005; Schwaninger et al., 2005). In the X-Ray ORT, X-ray images of passenger bags are displayed. As novices would not know how to interpret color information (the coding of materials), images are only shown in greyscale. Furthermore, only objects with common shapes (guns and knives), which would be identifiable without experience or special knowledge, are used. Threat items are combined with bags of different complexities (low and high) and with different levels of superposition (low and high). Every threat item is shown from two different viewpoints (easy vs. difficult). Each bag image is used twice, once without (non-threat image) and once containing a prohibited item (threat image). Altogether, the test consists of 256 trials. The task of the candidate taking the X-Ray ORT is to visually inspect the X-ray images and decide whether a gun or knife is present. Each image is presented for four seconds. For more detailed information on the X-Ray ORT see Schwaninger et al. (2005), Schwaninger et al. (2004) and Hardmeier et al. (2005, 2006a).

2.2.2. Cargo X-Ray Tutor (C-XRT) training system

X-Ray Tutor (XRT) is a computer-based training system developed to enhance X-ray image interpretation competency of X-ray screening officers (Schwaninger, 2004a). During training with XRT, X-ray images of passenger bags are presented to the screening officers which they have to visually inspect in order to decide whether a bag can be regarded as harmless (OK) or not (NOT OK). In case a bag is judged as NOT OK the screener has to identify the threat item in the bag by clicking on it. After each response, feedback is provided, informing the screener whether the response was correct or not. If the bag contains a prohibited item, detailed information about this item (what kind of threat item and what it looks like when X-rayed) as well as its location in the bag is displayed (Schwaninger, 2004a). One core advantage of XRT is that it is individually adaptive, i.e. training sessions are created based on each trainee’s individual performance and learning progress. The XRT training starts by showing threat items depicted in easy (canonical) views, with low superposition and low bag complexities. The difficulty of each of these factors is then increased depending on each individual’s learning progress, using sophisticated image processing algorithms (see Boling et al., 2008).

For this study, a new version of XRT was created and adapted to the field of cargo X-ray screening. The C-XRT was based on the same scientific background as earlier XRT versions, also accounting for

![Fig. 1. Experimental design of the study.](image-url)
the three relevant image-based factors and also being individually adaptive and level based. The XRT training systems for CBS and HBS use a scientifically validated image merging algorithm (Mendes et al., 2011), where an X-ray image of a fictional threat item (FTI) is automatically inserted into an X-ray image of a passenger bag during training. As ULDs or containers are much more complex with regard to size and filling, automatic merging could not yet be applied for the C-XRT. Instead, combined threat images (CTIs) were shown in the C-XRT. For this, X-ray images of ULDs/containers and FTIs were recorded separately. The FTIs were then merged into the ULD/container images by screening experts using an X-ray image merging-software based on the machine specific X-ray image format. The screening experts made sure that each CTI was realistic with regard to the placement of the FTI and that the FTIs were technically visible, i.e. they were potentially detectable for the human operator. All CTIs were double checked by an additional expert group to ensure that they were realistic.

The C-XRT version used in this study only included the most important features of XRT. Besides a zoom function, this version also offered the application of different IEFs. Images in the C-XRT could be viewed using the following options (similar to the ones the screening officers could apply at the X-ray machine during work): Normal (Greyscale), black & white square root (BW SQRT), black & white logarithmic (BW LOG), black & white histogram (BW HIST) and pseudo color (PC).

2.2.3. Cargo X-Ray Competency Assessment Test (C-CAT)

The X-Ray Competency Assessment Test (X-Ray CAT) is a reliable and valid instrument developed to measure X-ray image interpretation competency of security screeners, i.e. how well they can detect different types of prohibited items in X-ray images of passenger bags (Koller and Schwaninger, 2006). Similarly as in XRT, X-ray images of passenger bags are displayed on a computer screen. The security screeners have to visually inspect these for prohibited items and decide whether a bag can be considered as harmless (OK) or whether it contains a prohibited object (NOT OK).

For the field of cargo X-ray screening, a new version of the X-Ray CAT (Cargo X-Ray CAT, C-CAT) had to be created. Two versions were designed, one with the original greyscale images, and one where these images were color coded (pseudo color images). For the color coding, an X-ray machine specific algorithm was used which derived the colors by the different levels of grey on the greyscale images (not through the differentiation of materials). Apart from the coloring, both tests were identical, containing the same images of containers/ULDs (see Fig. 2 for an example).

The C-CAT consisted of 240 trials based on 120 different X-ray images of cargo containers and ULDs. All images were once displayed without prohibited items and once containing prohibited goods or threat items. The C-CAT contained 36 customs related and 24 security related objects, which were visually similar to objects included in the training system, but not identical. The number and types of prohibited items were defined and selected in collaboration with customs experts. The customs related objects were 24 drug items, eight weapons and four “miscellaneous” (e.g., Chinese medicine, watches etc.) items. The security related objects were inert IEDs or explosives, as defined and prepared by TNO. The comparisons between the 24 security and 24 drug items were of special interest. The items from these two categories could be further divided into three different weight categories (light/medium/heavy), resulting in eight items per weight class for each of these two groups. All objects were once shown with low superposition and once with high superposition. This added up to 120 threat images altogether. The applied merging-software displayed the superposition value. This was computed using a function that calculated the difference between the pixel intensity values of the ULD/container image with the FTIs and subtracting the ULD image without the FTIs, using the following formula (see Koller et al., 2008):

$$ SP = \frac{\sqrt{\sum (I_{SN}(x,y) - I_{H}(x,y))^2}}{\text{Object size}} $$

SP: superposition; $I_{SN}$: greyscale intensity of the SN (signal plus noise) image (contains a prohibited item); $I_H$: greyscale intensity of the N (noise) image (contains no prohibited item); $x, y$: image co-ordinates; object size: number of pixels of the FTI where $R$, $G$ and $B$ are $<253$.

Using this equation, the superposition value is independent of the size of the prohibited item. Additionally to detection performance, the C-CAT allowed measuring reaction times, i.e. the time a screener needed to search a ULD/container before taking a decision. The screening officers were instructed to analyze the images as fast and accurately as possible. A maximum of five minutes was set before the image disappeared on the screen.

3. Results and discussion

According to signal detection theory (SDT), four possible types of responses exist (Green and Swets, 1966): hit, false-alarm, correct rejection and miss. In this study, A’ (Pollack and Norman, 1964) was applied as a measure for detection performance. It considers the hit rate as well as the false-alarm rate and can be calculated by the following formulae (Grier, 1971):

When $H > F$: \[ 5 + \frac{[H - F](1 + H - F)}{[4H(1 - F)]} \] \[ (1) \]

When $H < F$: \[ 5 + \frac{[F - H](1 + F - H)}{[4F(1 - H)]} \] \[ (2) \]

$H$ is the hit rate and $F$ the false alarm rate. If performance is below chance, i.e. when $H < F$, Equation (2) must be used (Aaronson and Watts, 1987).

Due to the security confidential nature of performance values, these are not displayed in this paper. In order to provide meaningful results, relative differences between the two measurements and effect sizes are reported. The reported effect sizes are interpreted based on Cohen (1988). For t-tests, $d$ between .20 and .49 represents a small effect size; $d$ between .50 and .79 represents a medium effect size; $d$ $> 0.80$ represents a large effect size. For analysis of variance (ANOVA) statistics, $\eta^2$ between .01 and .05 represents a small effect size; $\eta^2$ between .06 and .13 represents a medium effect size; $\eta^2$ $\geq .14$ represents a large effect size.

Fig. 3 shows the detection performance of the two experimental groups and the control group for both C-CAT measurement times, i.e. at T1 (before training started) and at T2 (after approximately three months of training). As can be seen in the figure, there were large improvements as a result of training for both experimental groups (greyscale and pseudo color). Separate pairwise $t$-tests revealed significant differences between the two measurement times with large effect sizes for the greyscale group $t(14) = -5.38$, $p < .001$, $d = 1.39$, and the pseudo color group, $t(15) = -3.96$, $p < .01$, $d = 1.16$. No significant difference was found for the control group $t(8) = -1.2$, $p = .099$, $d = .04$. This suggests that improvements found in the experimental groups are solely due to training and cannot be explained by repeated exposure to the C-CAT.

---

1. Netherlands organization for Applied Scientific Research (TNO).
Additionally, a repeated measures ANOVA using A’ scores was conducted with test date (first, second) as a within-participants factor and group (greyscale vs. pseudo color) as a between-participants factor. There was a large effect for test date $F(1, 29) = 40.457, p < .001, \eta^2 = .582$, and no effect for group $F(1, 29) = .039, p = .845, \eta^2 = .001$. The interaction between test date and group was not significant, $F(1, 29) = .021, p = .887, \eta^2 = .001$. Hence, improvements were similar for both groups and it cannot be concluded that either one of the image versions at test (greyscale vs. pseudo color) led to a better detection performance.

In summary, a highly significant increase in detection performance measured with the C-CAT could be observed after three months of training. Altogether, the experimental groups (greyscale and pseudo color groups combined) increased their detection performance significantly by 14.9% ($t(30) = -6.47, p < .001, d = 1.28$). Thus, the results imply that training with the C-XRT did help to improve X-ray image interpretation competency of the cargo security screeners. Compared to earlier studies on cabin baggage screening where the authors are in the possession of the data (e.g., Halbherr et al., 2013; Koller et al., 2008; Michel, de Ruiter et al., 2007; Schwaninger and Wales, 2009), the absolute detection performance scores achieved in the C-CAT at both measurement times of this study are lower. However, it must be taken into account that the cargo security screeners have only been training for a very short period of time ($M = 8.95$ h, $SD = 5.68$). Further increases could be expected with a continuation of training. Moreover, the lower detection performance scores of the cargo security screeners at the baseline measurement underline the relevance of training.

Another repeated measures ANOVA was conducted including the factor superposition. A’ scores for test date (first, second) and superposition (low vs. high) as within-participants factors and group (greyscale vs. pseudo color) as between-participants factors were used. Large main effects of test date and superposition were revealed, while again no significant effect for group was found (see Table 1, a, for all results). Furthermore, the interaction of superposition and group was significant. As Fig. 4 indicates, the effect of superposition was as expected at $T_2$ for both groups (greyscale and pseudo color), i.e. higher detection performance when superposition is low. This was not so clear at $T_1$ which might be
The standard error of the mean represents the standard deviation of the sample-mean’s estimate of a population mean (Everitt, 2003).

explained by the relatively low detection performance of both groups before training started.

As described earlier, the threat items contained in the C-CAT were made up of 36 customs and 24 security related items. The security related items contained inert explosives and IEDs, the customs related items were 24 drug items, eight weapons (guns and knives) and four “miscellaneous” (e.g., Chinese medicine, watches etc.) items. Fig. 5 suggests that increases in detection performance after training resulted for every category. Both experimental groups achieved highest scores for the detection of IEDs and lowest for the category miscellaneous, guns, knives.

Table 1

Results of the ANOVA’s conducted with detection performance (A’) as the dependent variable.

<table>
<thead>
<tr>
<th>Factor</th>
<th>df</th>
<th>F</th>
<th>η²</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1, 29</td>
<td>38.19</td>
<td>.568</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Superposition (S) (low vs. high)</td>
<td>1, 29</td>
<td>9.70</td>
<td>.251</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Group (G) (greyscale vs. pseudo)</td>
<td>1, 29</td>
<td>.05</td>
<td>.002</td>
<td>&lt;.831</td>
</tr>
<tr>
<td>T × G</td>
<td>1, 29</td>
<td>.06</td>
<td>.002</td>
<td>&lt;.804</td>
</tr>
<tr>
<td>S × G</td>
<td>1, 29</td>
<td>8.40</td>
<td>.225</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>T × S</td>
<td>1, 29</td>
<td>15.41</td>
<td>.347</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>T × S × G</td>
<td>1, 29</td>
<td>1.01</td>
<td>.034</td>
<td>&lt;.322</td>
</tr>
<tr>
<td>b</td>
<td>1, 29</td>
<td>37.90</td>
<td>.566</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Category (C) (drugs: miscellaneous, guns, knives; explosives; IEDs)</td>
<td>2, 31, 67</td>
<td>33.45</td>
<td>.536</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Group (G) (greyscale vs. pseudo)</td>
<td>1, 29</td>
<td>.24</td>
<td>.008</td>
<td>&lt;.626</td>
</tr>
<tr>
<td>T × G</td>
<td>1, 29</td>
<td>.01</td>
<td>.000</td>
<td>&lt;.915</td>
</tr>
<tr>
<td>T × C</td>
<td>2, 18, 63</td>
<td>4.40</td>
<td>.132</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>C × G</td>
<td>3, 67</td>
<td>.09</td>
<td>.003</td>
<td>&lt;.965</td>
</tr>
<tr>
<td>T × C × G</td>
<td>3, 63</td>
<td>.06</td>
<td>.002</td>
<td>&lt;.982</td>
</tr>
<tr>
<td>c</td>
<td>1, 29</td>
<td>36.97</td>
<td>.560</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Category (C) (customs 24 vs. security 24)</td>
<td>1, 29</td>
<td>41.94</td>
<td>.591</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Group (G) (greyscale vs. pseudo)</td>
<td>1, 29</td>
<td>.12</td>
<td>.004</td>
<td>&lt;.733</td>
</tr>
<tr>
<td>T × G</td>
<td>1, 29</td>
<td>.01</td>
<td>.000</td>
<td>&lt;.921</td>
</tr>
<tr>
<td>C × G</td>
<td>1, 29</td>
<td>.71</td>
<td>.024</td>
<td>&lt;.407</td>
</tr>
<tr>
<td>T × C</td>
<td>1, 29</td>
<td>9.78</td>
<td>.252</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>T × C × G</td>
<td>1, 29</td>
<td>.21</td>
<td>.007</td>
<td>&lt;.653</td>
</tr>
</tbody>
</table>

Separate pairwise t-tests were conducted to compare detection performance at the first and second measurement for both groups and each threat category (Table 2). For both groups, increases in detection performance were revealed for drugs, explosives and IEDs. The comparisons of the effect sizes d between the t-tests of the four categories confirm that the training effect was particularly large for IEDs. Both experimental groups hardly differed with respect to the scores achieved for each category at each measurement time. A repeated measures ANOVA with test date (first, second) and category (drugs; miscellaneous, guns, knives; explosives; IEDs) as within-participants factors and group (greyscale vs. pseudo color) as between-participants factor confirmed this assumption (results can be seen in Table 1, b). There was no main effect of group (greyscale vs. pseudo color) and no interaction involving group reached statistical significance. Thus, with either kind of image (greyscale vs. pseudo color), prohibited items were detected equally well.

As Fig. 6 suggests, when comparing the 24 security related items with the 24 drug items (customs related items) a considerably larger increase for the security related items could be found. Separate pairwise t-tests for both experimental groups and both measurement times were conducted for these two categories. Effect sizes d were calculated to examine this difference (see Table 3) and indeed much larger effect sizes were found for security related items.

Moreover, Fig. 6 illustrates the similarity in detection performance of the two experimental groups (greyscale and pseudo color). The results of a repeated measures ANOVA with the within-participants factors test date (first, second) and category (drug items vs. security items) and the between-participants factor group (greyscale vs. pseudo color) are reported in Table 1, c. Large main effects were revealed for test date and category, whereas again no main effect was found for group and no interaction involving group reached statistical significance. The significant interaction between test date and category could indicate that the identification of the security related items was easier to learn than the identification of the drug related items, and therefore increases in detection

Fig. 3. Mean detection performance A’ with standard errors of the mean for the greyscale group, pseudo color group and control group at the first (T1) and second (T2) measurement. For security reasons, actual A’ scores are not displayed in the figures. The standard error of the mean represents the standard deviation of the sample-mean’s estimate of a population mean (Everitt, 2003).

Fig. 4. Mean detection performance A’ with standard errors of the mean for the greyscale group and pseudo color group at the first (T1) and second (T2) measurement, broken up by superposition.
performance were larger for this category. Figs. 5 and 6 also imply that the customs related items (including the 24 drug items, eight weapons and four “miscellaneous”), in general, seemed to be more difficult to recognize than the security related items.

In order to optimize throughput where relevant, it is of special interest to investigate how much time cargo security screeners need to search a container/ULD reliably and whether reaction time can be reduced through training. Reaction time (RT) was measured (in milliseconds) by the C-CAT software as the time from X-ray image onset until a response button was pressed (i.e. the OK or NOT OK button). For all three groups, a decrease in reaction time could be observed between the first and second measurement (see Fig. 7).

A repeated measures ANOVA with test date (first, second) as within-participants factor and group (greyscale, pseudo color, control) as between-participants factor revealed a large main effect for test date, and no significant effects for group and the interaction between test date and group (see Table 4). At first sight, the results of the ANOVA could be interpreted as a general effect resulting from repeated exposure to the C-CAT. However, it must be pointed out that such a general effect applies only to RTs. While they were lower at T2 for all groups, the improvement in detection performance was only observed for the two groups that received training, but not for the control group (see Fig. 3 and statistical results reported above).

Finally, it must be taken into consideration that the RTs observed in this study were measured in controlled test situations. In reality, cargo security screeners often take more time to inspect a container/ULD.

Further analyses of training data investigating the relationship between test performance and training behavior (number of images seen during training and the number of hours trained) revealed significant correlations. Both, linear and logarithmic correlations were conducted, to test which trend fits better. A linear correlation of \( r = .379, p < .05 \) and a logarithmic correlation of \( r = .414, p < .05 \) between the number of images seen during training by the cargo security screeners and performance in the C-CAT at the second measurement (Fig. 8a) could be found. Moreover, when correlating the difference in detection performance of all cargo security screeners between the two measurements with the number of hours trained in the meantime, a linear correlation of \( r = .464, p < .05 \) and a logarithmic correlation of \( r = .464, p < .01 \) was found (Fig. 8b). All correlations indicate that training with C-CAT can improve detection performance.

---

Table 2
Results of the t-tests comparing detection performances between the first (T1) and second (T2) measurement for each category.

<table>
<thead>
<tr>
<th></th>
<th>T1 (14)</th>
<th>T2 (44)</th>
<th>p value</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Greyscale group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drugs T1–T2</td>
<td>−3.54</td>
<td></td>
<td>&lt;.01</td>
<td>.95</td>
</tr>
<tr>
<td>Misc., guns, knives T1–T2</td>
<td>−1.03</td>
<td>−.307</td>
<td>.21</td>
<td>.71</td>
</tr>
<tr>
<td>Explosives T1–T2</td>
<td>−3.67</td>
<td></td>
<td>&lt;.01</td>
<td>1.06</td>
</tr>
<tr>
<td>IEDs T1–T2</td>
<td>−4.64</td>
<td></td>
<td>&lt;.001</td>
<td>1.29</td>
</tr>
<tr>
<td><strong>Pseudo color group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drugs T1–T2</td>
<td>−2.41</td>
<td></td>
<td>&lt;.05</td>
<td>.71</td>
</tr>
<tr>
<td>Misc., guns, knives T1–T2</td>
<td>−1.47</td>
<td>−1.47</td>
<td>.147</td>
<td>.29</td>
</tr>
<tr>
<td>Explosives T1–T2</td>
<td>−3.29</td>
<td></td>
<td>&lt;.01</td>
<td>1.16</td>
</tr>
<tr>
<td>IEDs T1–T2</td>
<td>−4.18</td>
<td></td>
<td>&lt;.001</td>
<td>1.36</td>
</tr>
</tbody>
</table>

---

Fig. 5. Mean detection performance \( A' \) with standard errors of the mean for the greyscale group and pseudo color group at the first (T1) and second (T2) measurement, broken up by category (drugs; miscellaneous, guns, knives; explosives, IEDs).

Fig. 6. Mean detection performance \( A' \) with standard errors of the mean for the greyscale group and pseudo color group at the first (T1) and second (T2) measurement, broken up by category (customs vs. security).
XRT did have an effect on the performance in the C-CAT and, thus, led to an increase in detection performance. For the existing data, a slightly better fit was found with the logarithmic functions. Usually, benefits from practice follow nonlinear functions, as improvement is rapid at first, but eventually evens out after the practitioner becomes more skilled (e.g., Ebbinghaus, 1971; Heathcote and Brown, 2000; Ritter and Schooler, 2001; Thorndike, 1913). Hence, it is likely that logarithmic functions explain these distributions better, which is consistent with our results.

4. Summary and conclusion

The main aim of this study was to investigate whether X-ray image interpretation competency of cargo security screeners can be increased by computer-based training (CBT). Consistent with results of previous studies in cabin baggage screening (Koller et al., 2008; Michel, de Ruiter et al., 2007; Schwaninger, 2004b; Schwaninger et al., 2008, 2007; Schwaninger and Wales, 2009), a highly significant increase in detection performance was found for cargo security screeners after approximately three months of training. Yet, it should be noted that even after training detection performance was still in need of improvement. In this study, the cargo security screeners had only trained for a comparably short period of time. In earlier studies, approximately 20 min of training per week were conducted for at least 6–12 months between measurements (Koller et al., 2008; Michel, de Ruiter et al., 2007). Hence, even larger improvements in X-ray image interpretation competency of cargo screeners could be expected with more training.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Results of the t-tests comparing detection performances between the first (T1) and second (T2) measurement for the 24 customs related (drugs) and 24 security related items.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t(14)</td>
</tr>
<tr>
<td>Greyscale group</td>
<td></td>
</tr>
<tr>
<td>Customs 24 (drugs) T1–T2</td>
<td>-3.54</td>
</tr>
<tr>
<td>Security 24 T1–T2</td>
<td>-5.74</td>
</tr>
<tr>
<td>Pseudo color group</td>
<td></td>
</tr>
<tr>
<td>Customs 24 (drugs) T1–T2</td>
<td>-2.41</td>
</tr>
<tr>
<td>Security 24 T1–T2</td>
<td>-4.29</td>
</tr>
</tbody>
</table>

![Fig. 7. Reaction times (in seconds) with standard errors of the mean for the greyscale, pseudo color and control groups at the first (T1) and second (T2) measurement.](image)

![Fig. 8. a Correlation between detection performance A' at the second measurement (T2) and the number of images seen during training. b. Correlation between the differences in detection performance A' between the two measurements and the number of hours trained.](image)
Results further revealed that viewing X-ray images in pseudo color at test did not lead to a better detection performance compared to when X-ray images were shown in greyscale. Thus, our findings imply that the application of pseudo color on its own as an image enhancement function (IEF) does not necessarily improve detection performance. Although consistent with earlier research on IEFs (Klock, 2005; Michel, Koller, et al., 2007) this does not mean that IEFs such as pseudo color are useless. It cannot be ruled out that a serial application of different IEFs may result in better detection performance compared to a condition where only greyscale images are available. Further research is needed to investigate this possibility.

Last but not least it should be noted that the prohibited items used in this study were all technically visible. In reality, prohibited items could be hidden and distributed in the ULDs/containers in ways that could make them even more difficult to spot. Thus, further improvements of man-machine system performance in cargo security screening must not only rely on training, but also on technological enhancements such as automatic threat detection and screener assist technologies (e.g., Elbert, 2009; Singh and Singh, 2003).

Acknowledgments

We are thankful to the Dutch Customs and TNO (Netherlands Organization for Applied Scientific Research) for their good collaboration and support in conducting the study.

References


Eilbert, R.F., 2009. Chapter 6


