Increased Detection Performance in Airport Security Screening Using the X-Ray ORT as Pre-Employment Assessment Tool

Diana Hardmeier, Franziska Hofer, Adrian Schwaninger

Abstract—**Detecting prohibited items in x-ray images of passenger bags is one of the most important tasks in aviation security. This screening process includes both, knowledge-based and image-based factors. That is, the knowledge about which items are prohibited and what they look like in x-ray images (knowledge-based factors) and the ability to cope with bag complexity, superposition and rotation of the threat item (image-based factors). The X-Ray ORT was developed to measure how well screeners and novices can cope with image-based factors. Schwaninger, Hardmeier, and Hofer (2004) could show that image-based factors are rather independent of knowledge and therefore can only be partly enhanced through training. As these image-based factors are very important in all x-ray screening tasks, using the X-Ray ORT as pre-employment assessment tool should result in a remarkable increase in detection performance of screeners in the future. To test whether the X-Ray ORT is a useful tool to select job applicants, detection performance of screeners selected with and without the X-Ray ORT was compared in the Prohibited Items Test (PIT), which mainly measures knowledge-based factors. This means that one group of job applicants (all novices) was hired using the X-Ray ORT, whereas the other group was hired without the X-Ray ORT. Both groups of screeners had undergone initial classroom training and a minimum of one year working experience in screening carry-on baggage when they took the PIT. Results evidence that in fact detection performance in the PIT is significantly higher for the group selected with the X-Ray ORT than detection performance of screeners selected without the X-ray ORT.**

Furthermore, results reveal reliable and valid measurement of detection performance in both tests, the ORT and the PIT.

Index Terms—**Aviation Security, Object Recognition, Pre-Employment Assessment, Reliability, Validity.**

I. Introduction

NOWADAYS, civil aviation has become more important and passenger flow still increases yearly. As a result, work load in aviation security increases enormously. To ensure effective and efficient work, it is very important to select and train people accurately. One of the most important tasks of aviation security screeners is detecting prohibited items such as guns, knives, improvised explosive devices (IEDs) and other prohibited items in passenger bags. During rush hours at checkpoints the decision whether a bag is OK (i.e. contains no prohibited item) or NOT OK (contains a prohibited item) has to be made within four seconds. This short time requires both, the profound knowledge about prohibited items and their appearance in x-ray images, as well as the ability to cope with image-based factors such as bag complexity, superposition and rotation of the threat item.

Referring to a general visual cognition model, recognition is defined as a successful matching of the stimulus representation with the visual memory representation. Based on this model [1] revealed two main factors in detecting threat items in x-ray screening, knowledge-based and image-based factors. First, screeners have to know which objects are prohibited and what they look like in x-ray images in order to recognize them (knowledge-based factors). As the appearance of prohibited items in x-ray images can differ remarkably from real life, training is very important in order to recognize them. In addition, it could be shown that an effective training system like X-Ray Tutor can significantly increase detection performance by reducing the false alarm rate. That is, through training screeners learn to distinguish reliably similar looking threat and non-threat items. Second, image-based factors influence detection performance in X-ray images enormously. [1] have shown three different types of image-based factors, namely bag complexity, superposition and rotation of the threat item. A threat item is more difficult to detect if it is shown in a close-packed bag as other objects can distract attention (effect of bag complexity). In addition, the more the threat item is superimposed by other objects in the bag, the harder it becomes to detect it (effect of superposition). Furthermore, a rotated threat item is more difficult to detect than a threat item shown in the frontal view (effect of viewpoint). These image-based factors are relatively independent of training and therefore rather referred to visual abilities. The ability to cope with image-based factors can be measured using the X-Ray ORT. This test consists of 256 x-ray images, half of them including either a gun or a knife.

Manuscript received February 25, 2006.

D. Hardmeier is with the University of Zurich, Department of Psychology, 8032 Zurich, Switzerland (phone: +41-44-3852; fax: +41-44-3856; e-mail: d.hardmeier@psychologie.unizh.ch).

F. Hofer is with the University of Zurich, Department of Psychology, 8032 Zurich, Switzerland (e-mail: f.hofer@psychologie.unizh.ch).

A. Schwaninger is with the University of Zurich, Department of Psychology, 8032 Zurich, Switzerland and Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany (e-mail: a.schwaninger@psychologie.unizh.ch).

These threat items are shown in the frontal and rotated view, more or less superimposed by other objects in the bag, in a close-packed or rather empty bag.

The above described image-based factors are supposed to play a key function in the x-ray screening process. Coping with bag complexity, superposition and viewpoint of a threat item can only be partly enhanced through training and is therefore rather dependent on the ability of each screener (see also [1]). Because these abilities play an important role in all x-ray image interpretation processes, screeners who have the relevant visual abilities should not only have a much better detection performance when untrained, but also after training and some working experience, compared to screeners who are less endued with image-based factors. To test this assumption, we compared detection performance of screeners, who were hired one year before using the X-Ray ORT with detection performance of screeners, who were selected using not fully standardized selection procedures. To compare detection performance of the two groups, we used the Prohibited Items Test (PIT). The PIT is a test including all kinds of prohibited items in x-ray images and therefore allows measuring knowledge-based factors in x-ray screening. This test provides a good possibility to measure the screener's detection performance of prohibited items independently of the selection process. Furthermore, reliability and validity of both tests, the ORT and PIT were evaluated.

II. METHOD

A. Participants

Two groups of aviation security screeners participated in this study. The experimental group consisted of 101 participants (71 male and 30 female) between 19 and 55 years (M = 35.25 years, SD = 9.79 years), who were all hired as security screeners based on the results of the X-ray ORT, which was used as part of the pre-employment assessment procedure. When taking the X-ray ORT, these job applicants had no x-ray image interpretation experience at all. Besides the X-Ray ORT, this group had to pass the color blindness test, an English test and a job interview as well in order to get employed. These screeners had about one year working experience in x-ray screening when this study was conducted (i.e. when taking the PIT).

The control group consisted of 453 screeners (141 male and 312 female) between 24 and 65 years (M = 48.94 years, SD = 9.09 years), who were hired without the X-ray ORT, but using an old selection procedure, which consisted of a color blindness test, an oral English test and a job interview. Working experience of these aviation security screeners varied from two years to 26 years (M = 9.71 years, SD = 5.50 years) when conducting this study (i.e. when detection performance in the PIT was compared to the experimental group).

B. Material

1) The X-Ray Object Recognition Test (X-Ray ORT)

The X-Ray ORT consists of 256 x-ray images and measures mainly image-based factors in x-ray screening. Therefore, only guns and knives are used as these threat objects are known by most people independent of visual experience or training and therefore are also well known by novices. Furthermore, all images are shown in black and white to eliminate color-diagnostic information for experts. To measure how good test candidates can cope with image-based factors, the image-based factors bag complexity, superposition and viewpoint are varied systematically with each other. That is, eight guns and eight knives were each combined with two bags with low complexity levels and two bags with high complexity levels, but once only little and once more superimposed by other objects in the bag. Furthermore, each bag is shown once with and once without a threat item. That is, half of the trials in the X-Ray ORT are completely harmless bags and contain neither a gun nor a knife. In the test, each image is shown for four seconds on the computer screen. Then, the test candidate has to decide whether the bag is OK (contains no gun and no knife) or NOT OK (contains a gun or a knife) by clicking the respective button on the screen. Additionally, test candidates are asked to indicate how sure they are in their decision clicking on a 50 point rating scale on the screen. For a closer description of the test design refer to [2].

2) Prohibited Items Test (PIT)

The Prohibited Items Test (PIT) was developed to measure how well aviation security screeners know what prohibited items look like in x-ray images. The PIT contains all kinds of prohibited items and thus measures mainly knowledge-based factors in x-ray screening. All prohibited items in the PIT can be classified into seven categories by ECAC, ICAO and EU prohibited items lists. A total of 19 guns, 27 sharp objects, 14 hunt and blunt instruments, 5 highly inflammable substances, 17 explosives, 3 chemicals and 13 other prohibited items (such as ivory, crocodile) are shown. In total the PIT includes 160 trials, half of them including prohibited items and half of them containing no prohibited items at all. 68 of the trials containing a prohibited item included exactly one prohibited item, whereas the other twelve trials included two or three prohibited items at once[1]. As this test was developed to measure mainly knowledge-based factors in x-ray images, all threat items were shown in an easy view, combined with bags of medium complexity level and medium superposition. Thus, all three image-based factors are kept relatively constant in the PIT. Furthermore, all images were shown in color to provide a realistic test environment.

Test taking procedure in the PIT was similar to the X-Ray ORT. First, a self-explanatory instruction was shown

[1] This was done to assure face validity. In reality more than one prohibited item can be in a passenger bag. Note that only bags including one prohibited item were used for analysis.

explaining the task followed by some exercise trials to familiarize the participants with the test taking procedure. After each of the six exercise trials a visual feedback was given whether the bag was OK (contains no prohibited item) or NOT OK (contains at least one prohibited item). In the test itself no more feedback was given to the test candidates. In the PIT, all images are displayed for a maximum of ten seconds on the screen. Test candidates have to decide whether the bag contains one or more prohibited items by clicking the OK or NOT OK button on the screen. If the bag is judged as NOT OK, screeners have to indicate to which of the seven categories the prohibited item(s) belongs to by clicking on the respective button(s)[2]. Besides giving the answer OK or NOT OK, test candidates have to indicate how sure they are in their decision by clicking on a 50 point rating bar on the screen. Pressing the space bar, the next image is shown. There are four blocks of trials, after which test candidates could take an individual short break if wanted. The order of blocks is counterbalanced across four groups of participants. Within each block the order of trials is random.

C. Procedure

To test whether the X-Ray ORT is a useful pre-employment assessment tool, detection performance of screeners selected without the X-Ray ORT[3] and screeners who were hired with the X-Ray ORT was compared using the test results in the PIT. All screeners who were hired with the X-Ray ORT had completed a classroom training and about one year of working experience when taking the PIT. Experience of screeners selected without the X-Ray ORT varied between two and 26 years when taking the PIT (for more details on detection performance and working experience see [3]).

III. Results

All test results were calculated using the "nonparametric" detection performance measure A' (see [4], [5]). A' takes into account the hit rate (i.e. bags containing a prohibited item judged as NOT OK) as well as the false alarm rate (i.e. harmless bags judged as NOT OK). This is especially important considering the task of an aviation security screener. A screener, who judges nearly all bags as NOT OK, would for sure have a high hit rate, but at the same time a very high false alarm rate and thus be very inefficient in his job. A good screener is expected to recognize most forbidden objects without being mistaken. For further information on detection performance measures, calculation and assumptions about A' see [6], [7] or [8].

A. Reliability and Validity of the X-Ray ORT

Reliability of the X-Ray ORT is very high for trained aviation security screeners and novices. Cronbach Alpha values range from .887 to .966 for screeners and from .907 to .970 for novices. As well split-half reliabilities (> .781 for screeners and > .778 for novices) support reliable measurement of detection performance using the X-Ray ORT. For more details about reliability of the X-Ray ORT see [2].

Different validity measures of the X-Ray ORT were evaluated by [2] in order to determine whether the test measures what it is supposed to measure and whether it can be used in making accurate decisions. Internal, convergent and discriminant validity measures evidence the former, whereas criterion-related validity refers to the correctness of decisions. Large effects of bag complexity, superposition and viewpoint could be shown for aviation security screeners and novices and support high internal validity. Furthermore, convergent and discriminant validity could be shown based on all 453 screeners selected with the old selection procedure correlating results in the X-Ray ORT with results in the PIT ($r = .61$, $p < .001$) and results in the computer-based questionnaire (CBQ) ($r = .27$, $p < .001$), respectively. The CBQ is a multiple choice questionnaire including airport specific questions about safety and security regulations at airports. Therefore, neither the ORT nor the PIT should show a high correlation with the CBQ. Criterion-related validity was examined by correlating detection performance in the X-Ray ORT with on-the-job performance measured by Threat Image Projection (TIP) data ($r = .41$, $p < .001$). TIP systems project fictional threat images into real passenger bags during work. Therefore, TIP allows measuring on-the-job detection performance. After each TIP image screeners receive a feedback message that a fictional threat item was present. TIP data were aggregated over a period of 17 months of 86 aviation security screeners. Detection performance was calculated using A' scores, i.e. hit and false alarm rates. The correlation between the X-Ray ORT and TIP data evidences that abilities measured with the X-Ray ORT are indeed important determinants of detection performance on-the-job. For more details about calculation of these validity measures see also [2].

B. Reliability and Validity of the PIT

As for the X-Ray ORT, Cronbach Alpha and split-half reliabilities were calculated with 453 aviation security screeners for the PIT. All reliability measures are based on percentage corrects (PC), i.e. hits and correct rejections, as well as on confidence ratings (CR), i.e. how sure screeners were in their decision. Based on signal detection theory, reliabilities were calculated for N trials (bags without a prohibited item) and SN trials (bags with prohibited items) separately. All reliability measures are listed in Table 1 for the two groups of screeners separately. All values are very similar for both groups and support reliable measurement of detecting threat items in x-ray images. Cronbach Alpha values are ranging from .870 to .943 and split-half reliabilities from .864 to .944.

[2] The answer to which of the seven categories the prohibited item(s) belonged to, was not used for the data analysis.

[3] Screeners selected without the X-Ray ORT had to take a color blindness test, as well as a common job interview.

TABLE I

RELIABILITY ANALYSES (PIT)

Reliability Coefficients		PC SN	PC N	CR SN	CR N
Screeners (Control Group N=453)	Cronbach Alpha	.874	.901	.910	.928
	Split-half (Guttman)	.871	.914	.900	.936
Screeners (Experimental Group N=101)	Cronbach Alpha	.908	.943	.870	.883
	Split-half (Guttman)	.878	.944	.877	.864

Cronbach Alpha values and split-half reliabilities (Guttman) of the PIT calculated for screeners selected without the X-Ray ORT (N=453) and screeners selected with the X-Ray ORT (N=101): PC = percentage correct, CR = confidence ratings, SN = signal plus noise trials, N = noise trials.

Validity of the PIT can be examined calculating convergent, discriminant and criterion-related validity. These measures were calculated based on all 453 aviation security screeners who were selected without using the X-Ray ORT as pre-employment assessment tool. Convergent validity was tested correlating test scores in the PIT with test scores in the X-Ray ORT. A' scores in the PIT correlated significantly with A' scores in the X-Ray ORT (r = .61, p < .001) indicating convergent validity. This rather high correlation makes sense because both tests investigate x-ray image interpretation and obviously also in the PIT image-based factors are relevant. Furthermore, correlation between A' scores in the PIT with percentage correct answers in the computer-based questionnaire (CBQ) indicates discriminant validity (r = .26, p < .001). As for the X-Ray ORT, criterion-related validity was estimated using threat image projection (TIP) data of the same TIP-library used for the validation of the X-Ray ORT (for more details about this library please see [2]). Correlation between test results in the PIT and on-the-job detection performance (TIP data) was r = .54 (p < .001). Thus, test results in the PIT can be used to predict on-the-job performance of screeners to a certain degree.

C. Evaluation of the X-Ray ORT as pre-employment assessment tool

In order to investigate whether the X-Ray ORT is a valuable tool for pre-employment assessment, the mean detection performance of both groups in the PIT was compared (see Figure 1). A significant difference in detection performance of prohibited items between screeners selected without the X-Ray ORT and the group hired with the X-Ray ORT can be shown. The job applicants who were selected with the X-Ray ORT are significantly better in detecting prohibited items in x-ray images, $t(552) = 14.51$, p < .001 one year after employment. To test whether the difference in detection performance is influenced by the age of screeners or working experience (see [3] for the influence of these factors on x-ray detection performance) an analysis of covariance (ANCOVA) with selection procedure as between-participants factor and age and working experience as covariates was conducted. Results show that even if these two covariates are considered, detection performance of the screeners selected with the X-Ray ORT is significantly higher compared to the other screeners, with an effect size of η^2 = .07, $F(1, 548)$ = 38.82, MSE = .004, p < .001.

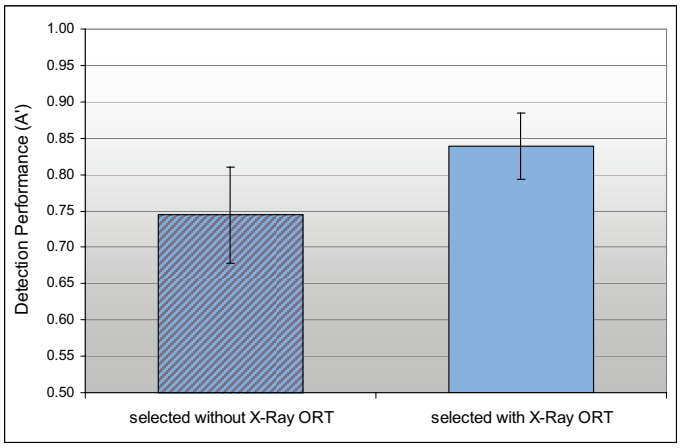

Fig. 1. Detection performance with standard deviations in the PIT for screeners selected without the X-Ray ORT (Control Group) left and screeners selected with the X-Ray ORT (Experimental Group) right.

IV. DISCUSSION

Overall, the results of the present study show that employing job applicants based on their results in the X-Ray ORT results in an increased detection performance in x-ray screening one year after employment, compared to detection performance of screeners selected without the X-Ray ORT. In this study, detection performance of two groups of screeners, each with a working experience of at least one year, was compared using the PIT, a computer-based x-ray screening test which measures rather knowledge-based factors in x-ray image interpretation. Compared to the screeners of the control group, who were not selected using the X-Ray ORT, job applicants who were hired based on the results in the X-ray ORT as pre-employment assessment tool performed significantly better in the PIT one year after employment. The effect size with η^2 = .07 ([9]) is still eminent, even when possible influences of the factors age and working experience are considered as covariates. Therefore, the ability how well someone can cope with the image-based factors, i.e. bag complexity, superposition and viewpoint, predicts detection performance in x-ray screening to a certain degree at a later date.

In addition, statistical analyses show for both x-ray performance measurement instruments high reliability and validity. The X-Ray ORT is not only a highly reliable and valid tool for measuring how well *novices* can cope with image-based factors in x-ray images, but also how well aviation security screeners with several years of working experience can handle these factors [1]. Furthermore, in this study we show that the PIT is a reliable instrument for measuring visual knowledge in the x-ray image interpretation task. Cronbach Alpha values were all > .87 and Guttman split half reliabilities > .86. Validity of the PIT was examined

calculating convergent, discriminant and criterion-related validity. The large correlation between the PIT and the X-Ray ORT (r =.61) supports convergent validity. A rather low correlation of r = .26 between the PIT and the CBQ (a test measuring general knowledge about security issues at airports) evidence discriminant validity. Furthermore, criterion-related validity of the PIT is also quite high (r = .54).

To further investigate whether the X-Ray ORT used as pre-employment assessment tool can also predict on-the-job detection performance, Threat Image Projection (TIP) data could be measured. Currently, this is examined in a recently started study, in which the two screener groups will be compared with regard to their TIP performance. Because TIP data are only reliable when a large TIP-library with realistic images is used and data are aggregated over several months [10], results are not yet available. As both tests, the X-Ray ORT and the PIT show high criterion-related validity, it can be assumed that the X-Ray ORT can effectively predict on-the-job detection performance.

Besides the importance of a valid and reliable pre-employment assessment procedure, intensive individual adaptive computer-based training (CBT) is also very important to improve detection performance of security screeners during work (see for example [11] for an evaluation study of CBT). In this context, it would be interesting if screeners with high values in the X-Ray ORT show a larger training effect than screeners with low performance in the X-Ray ORT. It could be assumed that screeners who are good in coping with image-based factors profit more from training than screeners who have problems with image-based factors. This is currently also under investigation.

ACKNOWLEDGMENT

This research was financially supported by Zurich Airport Unique, Switzerland. We are thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study.

REFERENCES

[1] A. Schwaninger, D. Hardmeier, and F. Hofer, "Measuring visual abilities and visual knowledge of aviation security screeners," *IEEE ICCST Proceedings*, vol. 38, pp. 258–264, 2004.

[2] D. Hardmeier, F. Hofer, and A. Schwaninger, "The X-Ray Object Recognition Test – A reliable and valid instrument for measuring visual abilities needed in x-ray screening," *IEEE ICCST Proceedings*, vol. 39, pp. 189–192, 2005.

[3] J. Riegelnig and A. Schwaninger, "The Influence of Age and Gender on Detection Performance and the Criterion in X-Ray Screening," *ICRAT Proceedings*, submitted for publication.

[4] J.B. Grier, "Nonparametric indexes for sensitivity and bias: Computing formulas," *Psychological Bulletin,* vol. 75, 424-429, 1971.

[5] R. E. Pastore, E. J. Crawley, M.S. Berens, and M. A. Skelly, ""Nonparametric" A' and other modern misconceptions about signal detection theory," *Psychonomic Bulletin & Review*, vol. 10, no. 3, pp. 556-569, 2003.

[6] D. M. Green and J. A. Sweets, *Signal Detection Theory and Psychophysics.* New York: Wiley, 1966.

[7] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behavior Research, Instruments, & Computers*, vol. 31, no. 1, pp. 137–149, 1999.

[8] N. A. MacMillan and C. D. Creelman, *Detection theory: A user's guide.* Cambridge: University Press, 1991.

[9] J. Cohen, *Statistical power analysis for the behavioural sciences.* New York: Hillsdale, 1988.

[10] F. Hofer and A. Schwaninger, "Using threat image projection data for assessing individual screener performance. In: C.A.Brebbia, T. Bucciarelli, F. Garzia, and M.Guarascio, *Transactions on the Built Environment (82), Safety and Security Engineering* (pp. 417-426). Wessex: WIT Press, 2005.

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