



# Investigating Training and Transfer Effects Resulting from Recurrent CBT of X-Ray Image Interpretation

**Saskia M. Koller (s.koller@psychologie.unizh.ch)**

University of Zurich, Department of Psychology, Binzmühlestrasse 14, 8050 Zürich, Switzerland

**Diana Hardmeier (d.hardmeier@psychologie.unizh.ch)**

University of Zurich, Department of Psychology, Binzmühlestrasse 14, 8050 Zürich, Switzerland

**Stefan Michel (s.michel@psychologie.unizh.ch)**

University of Zurich, Department of Psychology, Binzmühlestrasse 14, 8050 Zürich, Switzerland, and  
Max Planck Institute for Biological Cybernetics, Dept. Bühlhoff, Spemannstraße 38, 72076 Tübingen, Germany

**Adrian Schwaninger (a.schwaninger@psychologie.unizh.ch)**

University of Zurich, Department of Psychology, Binzmühlestrasse 14, 8050 Zürich, Switzerland, and  
Max Planck Institute for Biological Cybernetics, Dept. Bühlhoff, Spemannstraße 38, 72076 Tübingen, Germany

## Abstract

The importance of airport security has increased dramatically in the last years. Large investments into x-ray screening technology have been made in order to cope with the changed terrorist threat situation. However, the most expensive equipment is of limited value if the humans who operate it are not trained well enough to detect threat objects in x-ray images of passenger bags quickly and reliably. In this study we investigated whether adaptive computer based training (CBT) can be used to increase x-ray image interpretation competency of airport security screeners. To this end, we tested screeners before and after six months of weekly recurrent CBT using X-Ray Tutor (XRT). A control group of screeners was tested as well but this group did not receive training with XRT. Large increases in detection performance were found for the training group, which did also generalize to new threat objects that were not shown during training. The results of this study indicate that recurrent CBT can be a powerful tool to increase the x-ray image interpretation competency of screeners.

**Keywords:** Object recognition; perceptual learning; x-ray screening; human-computer interaction; airport security human factors.

## Introduction

In recent years, x-ray screening of passenger bags has become an essential component of airport security. Large investments were made into state-of-the art x-ray screening equipment. However, well trained human screeners are needed to operate the equipment appropriately in order to detect threat objects in passenger luggage within few seconds of inspection time. Object shapes that are not similar to ones stored in visual memory are difficult to recognize (e.g., Graf, Schwaninger, Wallraven, & Bühlhoff, 2002; Schwaninger, 2004, 2005). Thus, a prerequisite for good threat detection performance is knowledge about which objects are prohibited and what they look like in x-ray images. Schwaninger, Hardmeier, and Hofer (2005) have shown that x-ray screener performance depends on knowledge-based and image-based factors. Image-based

factors refer to image difficulty resulting from viewpoint variation of threat objects, superposition of threat objects by other objects in a bag, and bag complexity depending on the number and type of objects in the bag. The ability to cope with image-based factors is related to individual visual-cognitive abilities rather than a mere result of training. In contrast, knowledge-based factors refer to knowing which items are prohibited and what they look like in x-ray images of passenger bags. Because objects look quite different in x-ray images than in reality and because many threat objects are not known from everyday experience, computer-based and on the job training are important determinants of x-ray detection performance. Schwaninger et al. (2005) compared detection performance of novices with the one of trained aviation security screeners. A rather poor recognition of unfamiliar object shapes (e.g. self-defense gas spray, electric shock device etc.) in x-ray images was found for novices. For trained aviation security personnel, a much higher recognition performance was shown. Schwaninger and Hofer (2004) showed that adaptive computer-based training (CBT) can be very effective to increase the detection of improvised explosive devices (IEDs) in x-ray images of passenger bags. McCarley, Kramer, Wickens, Vidoni, and Boot (2004) reported a better performance after training for the detection of knives in x-ray images.

The purpose of this study was to investigate to what extent the previous findings can be expanded to other threat categories (e.g., guns and other prohibited items) and to examine transfer effects. The training group conducted weekly recurrent CBT using X-Ray Tutor (Schwaninger, 2004). The control group did not receive this type of training and conducted recurrent classroom training including another CBT system. Both groups of screeners were tested before and after 6 months using the X-Ray Competency Assessment Test (X-Ray CAT, Koller & Schwaninger, 2006). This test shows different kinds of prohibited items in x-ray images of passenger bags. Half of

the threat objects in the X-Ray CAT were not presented during the training sessions. This enabled measuring whether a transfer of the gained knowledge about trained objects to untrained but similar looking objects occurs.

## Method

### Participants

A total of 209 airport security screeners of a European airport participated in this study and conducted the X-Ray CAT 1.0.0 two times with an interval of six months. The training group consisted of 97 screeners who conducted weekly recurrent CBT of about 20 minutes using X-Ray Tutor (XRT) CBS 2.0 Standard Edition during the 6 months interval between the two test measurements. The control group consisted of 112 screeners and they did not conduct weekly recurrent CBT with XRT.

### Materials

The X-Ray CAT consists of 128 x-ray images of passenger bags. Each of the bags is used twice, once containing a prohibited item (threat image) and once without any threat object (Figure 1 displays an example of the stimuli). The threat items belong to four categories of prohibited items as defined in Doc 30 of the European Civil Aviation Conference (ECAC): guns, improvised explosive devices (IEDs), knives and other prohibited items (e.g., gas, chemicals, grenades etc.). The threat objects have been selected and prepared by experts of Zurich State Police, Airport division to be representative and realistic.



Figure 1: Example of an x-ray image of a passenger bag. The image on the right contains the prohibited item depicted separately on the bottom right.

For each threat category 16 exemplars are used (8 pairs). Each pair consists of two prohibited items that are similar in shape (see Figure 2). These were distributed randomly into two sets, set A and set B.



Figure 2: Example of two x-ray images of similar looking threat objects used in the test, one belonging to set A and B, respectively.

Every item is depicted from two different viewpoints. The easy viewpoint shows the object from a canonical perspective (Palmer, Rosch, & Chase, 1981) as judged by two security experts who captured the stimuli. The difficult viewpoint shows the threat item with an 85 degree horizontal rotation or an 85 degree vertical rotation relative to the canonical view. In each threat category half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally. Set A and B are equalized concerning the rotations of the prohibited objects. The effects of viewpoint are not analyzed in this study and will be reported elsewhere.

Every threat item is combined with a bag in a manner that the degree of superposition by other objects is similar for both viewpoints. This was achieved using a function that calculates the difference between the pixel intensity values of the bag image with the threat object minus the bag image without the threat object using the following formula:

$$SP = \frac{\sqrt{I_{SN}(x, y) - I_N(x, y)}}{\text{ObjectSize}}$$

SP = Superposition;  $I_{SN}$  = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item);  $I_N$  = Grayscale intensity of the N (Noise) image (contains no prohibited item); Object Size: Number of pixels of the prohibited item where R, G and B are < 253

Using this equation (division by object size), the superposition value is independent of the size of the prohibited item. This value can be kept relatively constant for the two views of a threat object, independent of the degree of clutter in a bag, when combining the bag image and the prohibited item. The bag images were visually inspected by aviation security experts to ensure they do not contain any other prohibited items. Harmless bags were assigned to the different categories and viewpoints of the threat objects in a way that their difficulty was balanced across all categories<sup>1</sup>. The false alarm rate (the rate at which screeners wrongly judged a harmless bag as containing a threat item) for each bag image served as measure of difficulty based on a pilot study with 192 screeners.

The X-Ray CAT is integrated in the XRT training system and takes about 20-30 minutes to complete. Each image is shown for a maximum of 10 seconds on the screen. Screeners have to judge whether the bag is OK (contains no prohibited item) or NOT OK (contains a prohibited item). Additionally, screeners have to indicate the perceived

<sup>1</sup> The eight categories of test images (four threat categories in two viewpoints each) are similar in terms of the difficulty of the harmless bags. This means, a difference of detection performance between categories or viewpoints can not be due to differences in the difficulty of the bag images.

difficulty of each image on a 100 point scale (difficulty rating). The difficulty ratings were not analyzed in study and will be reported elsewhere. The visible appearance of the test is the same as in training except there is no feedback and screeners do not have to click on the image to identify the threat object (see Figure 3). Feedback is provided only during training and informs the screener whether the image has been judged correctly or not. If the bag contains a threat item, it is highlighted by flickering after the screener responded with OK or NOT OK and the screener has the possibility to display information about the threat item (see Figure 3). As mentioned previously, during training, screeners have to click on the image and mark the object they perceive to be a threat item. This is not required during test mode.

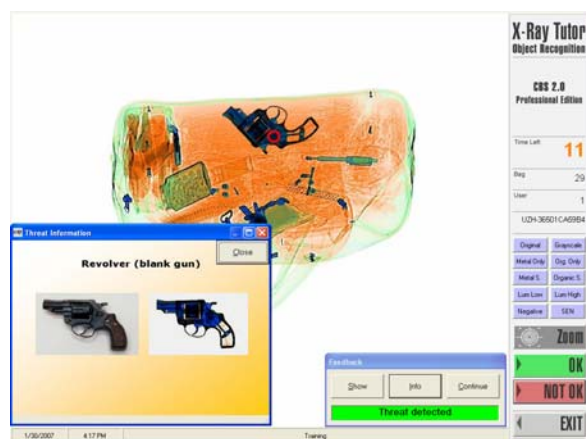


Figure 3: Screenshot of the XRT training system during training. At the bottom right a feedback is provided. If a bag contains a threat item, an information window can be displayed (see bottom left of the screen).

## Procedure

Screeners were randomly distributed into two groups. Both groups conducted the X-Ray CAT 1.0.0 without having trained with XRT before (baseline measurement). After test completion, only one group received recurrent adaptive CBT using XRT (training group). On average, each screener of the training group conducted 20.26 min recurrent training per week (SD = 3.65 min). After six months, both groups conducted the X-ray CAT again. This approach allows the comparison of the two test measurements and the performance of the two groups prior to and after training with XRT.

In order to measure a transfer effect, only the images of the prohibited items of test set A were included in training. They are part of the XRT CBS 2.0 SE training library, which contains 100 threat items belonging to the four threat categories (guns, IEDs, knives, other). Most of them are depicted from six different viewpoints. No bag image of the test appeared during training with XRT. During training, images containing a threat object are created at the point of

use, that is, test threat items (set A) and other threat items are digitally inserted into randomly selected bag images at random positions. For details on XRT see Schwaninger (2004).

## Results

Detection performance was calculated using the signal detection measure  $d'$  (Green & Swets, 1966), which takes into account the hit rate (correctly judged threat images as being NOT OK) and the false alarm rate (wrongly judged harmless bags as being NOT OK).

Figure 4 shows the detection performance of the first and second measurement for both screener groups. Performance values are not reported due to security reasons. However, effect sizes are reported for all relevant analyses and interpreted based on Cohen (1988), see Table 1.

Table 1: Classification of effect sizes according to Cohen (1988)

Effect size	$d$	$\eta^2$
small	0.20	0.01
medium	0.50	0.06
large	0.80	0.14

An analysis of variance (ANOVA) for repeated measures using  $d'$  scores with the within-participant factor measurement (first vs. second) and the between-participant factor group (trained vs. control) revealed a large main effect of measurement (first vs. second),  $\eta^2 = .40$ ,  $F(1, 207) = 138.66$ ,  $p < .001$ , a medium main effect of group (trained vs. control),  $\eta^2 = .13$ ,  $F(1, 207) = 31.22$ ,  $p < .001$ , and a large interaction of measurement and group  $\eta^2 = .34$ ,  $F(1, 207) = 105.55$ ,  $p < .001$ .

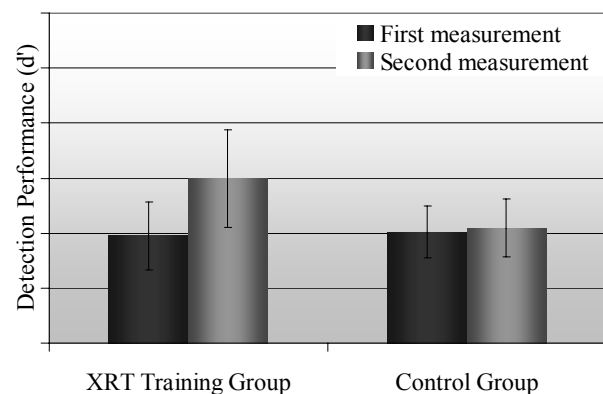


Figure 4: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement.

Separate pairwise  $t$ -tests of detection performance  $d'$  revealed no significant difference at the baseline

measurement between the two groups ( $p = .353$ ) and no significant difference for the control group in both measurements ( $p = .108$ ). However, there was a significant difference for the XRT training group between the first and the second measurement ( $p < .001$ ) with a large effect size of  $d = 1.39$ . There was also a significant difference between the two groups at the second measurement,  $p < .001$ , with a large effect size of  $d = 1.27$ .

Figure 5 shows the detection performance for each threat category separately for both groups at the first and the second measurement. A repeated-measures ANOVA with the within-participant factors measurement (first vs. second) and threat category (guns, IEDs, knives and other), and the between-participant factor group (XRT training vs. control) revealed the main effects and interactions given in Table 2a.

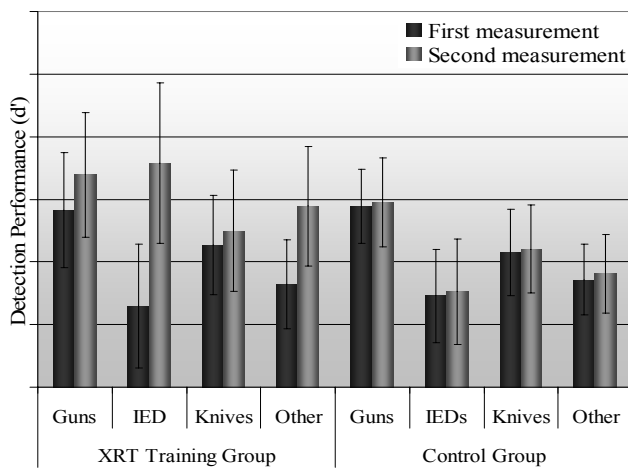


Figure 5: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for each threat category separately.

Separate pairwise  $t$ -tests were conducted to compare detection performance at the first and the second measurement for both groups and each threat category separately. The XRT training group showed a significant increase of the detection performance at the second measurement for each threat category (guns, IEDs and other threat objects, all  $p < .001$ , all  $d > 0.60$ , knives,  $p < .05$ ,  $d = 0.26$ ). Detection performance of the control group did not differ significantly between the two measurements (guns:  $p = .358$ , IEDs:  $p = .296$ , knives:  $p = .467$ , and other threat objects:  $p = .168$ ).

The results of the analysis considering the two sets of the test, set A and set B, are shown in Figures 6 and 7.

The results of the repeated measures ANOVA with the within-participant factors measurement (first vs. second) and test set (A vs. B) and the between-participant factor group (XRT training group vs. control group) can be seen in Table 2b. Pairwise  $t$ -tests showed a significant increase in detection performance at the second measurement for both

sets for the XRT training group (set A and B:  $p < .001$ ,  $d > 1.25$ ) but not for the control group.

Table 2: Results of the ANOVAs

	Factor	df	F	$\eta^2$	p
a)	Measurement (M)	1, 207	140.23	0.40	<.001
	Threat Category (T)	3, 621	222.7	0.52	<.001
	Group (G)	1, 207	37.57	0.15	<.001
	M x G	1, 207	108.16	0.34	<.001
	T x G	3, 621	29.36	0.12	<.001
	M x T	3, 621	76.5	0.27	<.001
	M x T x G	3, 621	74.78	0.27	<.001
b)	Measurement (M)	1, 207	138.39	0.40	<.001
	Group (G)	1, 207	32.64	0.14	<.001
	Test Set (S)	--	--	--	n.s.
	M x G	1, 207	104.08	0.34	<.001
	M x S	1, 207	8.72	0.04	<.01
	S x G	1, 207	17.31	0.08	<.001
	M x S x G	1, 207	7.92	0.04	<.01
c)	Measurement (M)	1, 207	146.15	0.41	<.001
	Threat Category (T)	3, 621	219.54	0.52	<.001
	Group (G)	1, 207	42.53	0.17	<.001
	M x G	1, 207	108.68	0.34	<.001
	T x G	3, 621	30.29	0.13	<.001
	M x T	3, 621	78.18	0.27	<.001
	M x S	1, 207	10.17	0.05	<.01
	T x S	3, 621	58.12	0.22	<.001
	M x T x G	3, 621	75.51	0.27	<.001
	M x S x G	1, 207	6.67	0.02	<.05

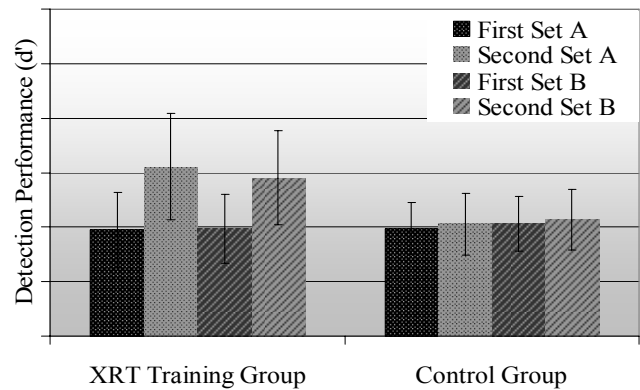


Figure 6: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for set A and set B separately.

An ANOVA for repeated measures with the within-participant factor set showed a very small significant main effect of set  $\eta^2 = .02$ ,  $F(1, 208) = 3.94$ ,  $p < .05$  at the first measurement. Pairwise  $t$ -tests comparing both sets within one group at the first measurement revealed a significant difference of the two sets only for the control group with only a small effect size ( $p < .01$ ,  $d = 0.17$ ) but not for the XRT training group ( $p = .676$ ).

An extended ANOVA with the additional within-participant factor threat category revealed the main effects and interactions as specified in Table 2c.

Pairwise *t*-tests confirmed a significant ( $p < .001$ , all  $d > 0.46$ ) increase in detection performance for the XRT training group for all threat categories per set except for knives (set A:  $p < .05$ ,  $d = 0.27$ , set B;  $p = .127$ ,  $d = 0.19$ ). The control group showed no significant change in detection performance at the second measurement for neither threat category per set (set A: guns  $p = .147$ , IEDs  $p = .202$ , knives  $p = .801$ , other threat objects  $p = .245$ ; set B: guns  $p = .974$ , IEDs  $p = .597$ , knives  $p = .235$ , other threat objects  $p = .123$ ).

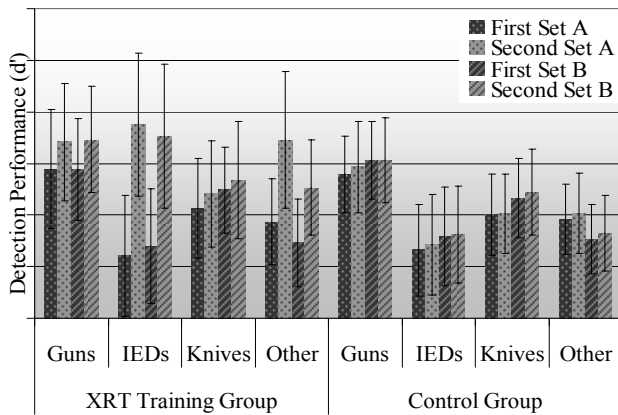


Figure 7: Detection performance with standard deviations for the XRT training group vs. the control group comparing first and second measurement for set A and set B and each threat category separately.

## Discussion

The results of this study confirmed earlier findings on x-ray detection performance of airport security screeners showing that adaptive CBT with X-Ray Tutor (XRT) results in substantial increases of detection performance (e.g., Hardmeier, Hofer, & Schwaninger, 2006; Schwaninger & Hofer, 2004; Schwaninger et al., 2005). In this study, the training group showed remarkable increases in detection performance for all types of threat objects (guns, knives, IEDs, and other prohibited items). For the control group, which did not conduct weekly recurrent CBT with XRT, no significant change in detection performance was observed. It should be noted that according to the security organization and their appropriate authority, the control group did recurrent training as mandated by national regulation during the whole duration of the study. This training was comparable in terms of the required training hours and included x-ray image interpretation training using another commercially available CBT. Thus, the improved performance in the training group reflects specific effects of training with XRT and they can not be explained by a "Hawthorne Effect". The largest training effect was found

for IEDs. It should be noted that as all other stimuli, the IEDs were developed by police experts of Zurich State Police, Airport Division. Especially the IEDs were quite sophisticated threat objects using components that are often not known to screeners without enhanced training in IED detection. Thus it is not surprising that before training,  $d'$  scores for IEDs were substantially smaller than for guns. However, after training, IED detection of the training group was very good and even slightly better than gun detection. This shows that the detection of IEDs is not difficult per se, but rather depending on the training of screeners.

Besides measuring training effects, the main aim of this study was to examine whether gained knowledge about trained threat objects can be transferred to similar looking objects. Since the X-Ray CAT is composed of two comparable sets (set A and set B) this can easily be tested by including the threat objects of one set (in this case set A) into the XRT system. A large transfer effect would mean a similarly higher detection performance after training for both sets. This was confirmed, as Figures 6 and 7 illustrate. The significant increase of the detection performance for the XRT training group was found for the trained test set A as well as for the untrained test set B. This implies a large transfer of the acquired knowledge about the visual appearance of trained objects (set A) to untrained but similar looking objects (set B). The comparison of the two sets A and B at the baseline measurement over all screeners showed a slightly significant difference ( $p < .05$ ) indicating that the two sets are not exactly equal in terms of image difficulty. But this possible objection to the transfer effect can be disapproved with two arguments: first, the effect size is only small according to the conventions by Cohen (1988, see also Table 1), and second, only the control group showed a significant difference ( $p < .01$ ) but not the XRT training group ( $p = .676$ ). Therefore, the transfer effect in the results of the XRT training group can be attributed to the training of set A only.

Transfer effects were revealed for all threat categories, i.e. for guns, IEDs and other threat objects. For knives, a significant training effect appeared only in the trained set A ( $p < .05$ ) but not in the untrained set B ( $p = .127$ ). Thus, there was no transfer effect for knives from set A to set B. Either the knives of the two sets were not similar enough in shape to allow a transfer effect, or the small training effect for knives is due to their shape. On one hand, knives show less diagnostic features which play an important role in object recognition compared to objects from other categories. On the other hand, the visual similarity of knives to harmless everyday objects (e.g., pen) is substantial. These factors could impede detectability and trainability and ultimately might have resulted in small transfer effects.

Contrary to our results, Smith, Redford, Gent, and Washburn (2005) found a large decrease in screeners' detection performance when specific trained objects were replaced with new images belonging to the same categories

(see also Smith, Redford, Washburn, and Tagliatalata, 2005). According to these authors, improvement in screening performance is attributable only to specific-token familiarity that developed for the original images and not to a category generalization. They state constraints on categorization and the use of category-general information when humans face visual complexity and have to identify targets within it. Our results can be interpreted in support of generalization of visual learning in x-ray image interpretation. However, it might be possible that the objects of the untrained set in our study are so similar to the trained objects that a specific-token familiarity led to the detection performance increase and not a true generalization effect. The lacking transfer effect in knives would along these lines mean that the objects in set A and set B are not similar enough in shape to generate a specific-token familiarity. Therefore only the learnt objects could generate a training effect but not the unlearnt ones. For Schwaninger and Hofer's (2004) findings of a large increase in detection performance of IEDs after recurrent CBT with other members of the category than those included in the test, it would mean, that those objects were very similar in order to create a specific-token familiarity and therefore a training effect.

For our future studies, it could also be interesting to increase the interval between the end of training and the testing of training transfer, as corresponding literature usually tests transfer of training after a considerable period of time in order to measure the stability of the transfer (e.g., Saks & Belcourt, 2006). However, most research is about organizational training and therefore training transfer is related to learning working skills and the generalization to the job context (Baldwin & Ford, 1988). In contrast our transfer refers to the transfer of visual knowledge about objects to other objects.

In any case, our findings show that the knowledge about the visual appearance of forbidden objects, which airport security screeners acquire during recurrent CBT, can be transferred to similar looking, but not previously seen objects. Thus, adaptive CBT can be a powerful tool to increase screeners' x-ray image interpretation competency.

### Acknowledgments

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). We thank four reviewers for valuable comments. Many thanks to Zurich State Police, Airport Division, for their help in creating the stimuli and the good collaboration for conducting parts of the study.

### References

Baldwin, T., & Ford, J.K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New York: Erlbaum, Hillsdale.
- Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis (IST-2000-29375)*.
- Green, D.M., & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
- Hardmeier, D., Hofer, F., & Schwaninger, A. (2006). The role of recurrent CBT for increasing aviation security screeners' visual knowledge and abilities needed in x-ray screening. *The 4<sup>th</sup> International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 – December 1, 2006*.
- Koller, S., & Schwaninger, A. (2006). Assessing X-ray image interpretation competency of airport security screeners. *Proceedings of the 2<sup>nd</sup> International Conference on Research in Air Transportation, ICRAT 2006 Belgrade, Serbia and Montenegro, June 24-28, 2006*, 399-402.
- McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., & Boot, W.R. (2004). Visual skills in airport screening. *Psychological Science, 15* (5), 302-306.
- Palmer, S.E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In I. Long and A. Baddeley (Eds.), *Attention and Performance IX*. Hillsdale, N.J.: Erlbaum.
- Saks, A.M, & Belcourt, M. (2006). An Investigation of Training Activities and Transfer of Training in Organizations. *Human Resource Management, 45* (4), 629-648.
- Schwanger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International, FEB/2004*, 31-36.
- Schwanger, A. (2005). Object recognition and signal detection. In B. Kersten (Ed.), *Praxisfelder der Wahrnehmungspsychologie*. Bern: Huber.
- Schwanger, A., Hardmeier D., & Hofer F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems, 20*(6), 29-35.
- Schwanger, A., & Hofer, F. (2004). 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*. Wessex: WIT Press.
- Smith, J.D., Redford, J.S., Gent, L.C., & Washburn, D.A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General, 134* (4), 443-460.
- Smith, J.D., Redford, J.S., Washburn, D.A., & Tagliatalata, L.A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31* (6), 1171-1185.