IEEE ICCST Proceedings, 38, 303-308

Reliable and Valid Measures of Threat Detection Performance in X-ray Screening

Franziska Hofer, Adrian Schwaninger
Department of Psychology, University of Zurich, Switzerland

Abstract—Over the last decades, airport security technology has evolved remarkably. This is especially evident when state-of-the-art detection systems are concerned. However, such systems are only as effective as the personnel who operate them. Reliable and valid measures of screener detection performance are important for risk analysis, screener certification and competency assessment, as well as for measuring quality performance and effectiveness of training systems. In many of these applications the hit rate is used in order to measure detection performance. However, measures based on signal detection theory have gained popularity in recent years, for example in the analysis of data from threat image projection (TIP) or computer based training (CBT) systems.

In this study, computer-based tests were used to measure detection performance for improvised explosive devices (IEDs). These tests were conducted before and after training with an individually adaptive CBT system. The following measures were calculated: pHit, d', Δm, Az, A', $p(c)_{max}$. All measures correlated well, but ROC curve analysis suggests that "nonparametric" measures are more valid to measure detection performance for IEDs. More specifically, we found systematic deviations in the ROC curves that are consistent with two-state low threshold theory of [9]. These results have to be further studied and the question rises if similar results could be obtained for other X-ray screening data. In any case, it is recommended to use A' in addition to d' in practical applications such as certification, threat image projection and CBT rather than the hit rate alone.

***Index Terms*—human factors in aviation security, hit rate, signal detection theory, threat detection in X-ray screening, computer based training system, threat image projection.**

This research was financially supported by Zurich Airport Unique, Switzerland.

Franziska Hofer, University of Zurich, Department of Psychology (e-mail: fhofer@allgpsy.unizh.ch).

Adrian Schwaninger, University of Zurich, Department of Psychology (e-mail: aschwan@allgpsy.unizh.ch).

I. Introduction

Technological progress enabled state-of-the-art X-ray screening to become quite sophisticated. Current systems provide high image resolution and several image "enhancement" features (e.g. zoom, filter functions such as negative image, edge detection etc.). But technology is only as effective as the humans that operate it. This has been realized more and more in recent years and the relevance of human factors research has increased substantially. Note that during rush hour, aviation security screeners often have only a few seconds to inspect X-ray images of passenger bags and to judge whether a bag contains a forbidden object (NOT OK) or whether it is OK. Threat object recognition does largely depend on perceptual experience and training [1]. Most object recognition models agree with the view that the recognition process involves matching an internal representation of the stimulus to a stored representation in visual memory (for an overview see [2] and [3]). If a certain type of forbidden item has never been seen before, there exists no representation in visual memory and the object becomes very difficult to recognize if it is not similar to stored views of another object. Besides the aspect of memory representation, image-based factors can also affect recognition substantially (for a detailed discussion see [4]). When objects are rotated, they become more difficult to recognize (effect of view). In addition, objects can be superimposed by other objects, which can impair detection performance (effect of superposition). Moreover, the number and type of other objects in the bag challenge visual processing capacity, which can affect detection performance as well (effect of bag complexity).

While CBT can increase detection performance substantially [1], recent results suggest that training effects are smaller for increasing the ability to cope with image-based factors such as effects of view, superposition and bag complexity [4]. However, this conclusion relies on the availability of reliable and valid measures of detection performance. For example the hit rate is not a valid measure for estimating the detection performance in a computer-based test in which screeners are exposed to X-ray images of passenger bags and have to take OK / NOT OK decisions. The reason is simple: A candidate could achieve a high hit rate by simply judging most bags as NOT OK. In order to distinguish between a liberal response bias and true detection ability, the false alarm rate needs to be taken into account as well. This is certainly part of the reason why signal detection theory (SDT)

has been used for analyzing X-ray screening data (see for example [5] and [1]). In general, reliable and valid measures of detection performance are certainly very important for risk analysis, screener certification and competency assessment, as well as for measuring quality performance and effectiveness of training systems.

The objective of this study was to compare different measures of screener detection performance. As clearly shown by [1], detection performance depends substantially on perceptual experience and training – at least for certain types of threat items. In order to evaluate different performance measures we used computer-based tests before and after CBT. Using a baseline test and a test after the training period makes it possible to compare different detection measures with regard to their reliability and validity while taking effects of training into account. The following detection measures were compared in this study: pHit, d', Δm, Az, A', $p(c)_{max}$. These measures and the corresponding detection models are summarized in the following section.

II. Detection Models and Performance Measures

Signals are always detected against a background of activity, also called noise. Thus, detecting threat objects in passenger bags could be described as typical detection task, where the signal is the threat object and the bag containing different harmless objects constitutes the noise. A correctly identified threat object corresponds to a hit, whereas a bag, which contains no threat item judged as being harmless, represents a correct rejection. Judging a harmless bag as being dangerous is a false alarm, whereas missing a forbidden object in a bag represents a miss. In every detection situation, the observer must first make an observation and then make a decision about this observation.

Signal detection theory [6], [7], and threshold theories [8], [9] are two fundamentally different approaches of conceptualizing human perception. The main difference between the two approaches is that threshold theories suppose a theoretical threshold, whereas in SDT the concept of a threshold is rejected in favor of an adjustable decision criterion. Threshold theories can be coarsely divided into high (e.g. single high-threshold theory or double high-threshold theory) and low threshold theories. These approaches assert that the decision space is characterized by a few discrete states, rather than the continuous dimensions of SDT. Single high threshold theory predicts ROC curves which are often not consistent with experimental data [10], [19]. The other types of threshold theories are low threshold theories originally described by [11] and [12]. The two-state low threshold theory is a slightly newer version by [9]. This theory is often as consistent with the data as are SDT models. But because no single sensitivity measure exists for this theory, it is not widely applied [7].

According to SDT, the subject's decision is guided by information derived from the stimulus and the relative placement of a response or decision criterion. An anxious person would set the criterion very low, so that a very small observation would lead to a signal response. In contrast, another person might set the criterion very high, so that the sensory observation needs to be very strong that this person would give a "signal present answer". It is important to note that different persons can have different criterion locations, and it is also possible that the same person changes the location of the criterion over time. For example the day after 9/11, many airport security screeners have moved their criterion in the direction that at the smallest uncertainty, they judged passenger bags as being dangerous. Although the hit rate increases, detection performance stays more or less stable, because also the false alarm rate increases.

In contrast to signal detection theory, threshold theories suppose that not the locus of the criterion causes the answer but a theoretical threshold. In the two-state low threshold theory by [9], the threshold is assumed to exist somewhere between the middle and the upper end of the noise distribution. During a sensory observation, an observer is in the detect state if the observation exceeds threshold and in the nondetect state if the observation is below threshold. The response one makes in either state may be biased by nonsensory factors. A person can say yes when in the nondetect state or say no when in the detect state. Manipulating variables such as payoff and signal probability changes the observers' response bias when they are in either one of the two possible detection states. The main disadvantage of this low threshold theory is its lack of a single sensitivity measure that can be calculated from hits and false alarms.

Different signal detection measures and sensitivity measures coming from threshold theories exist. One of the most popular and very often used parametric SDT measures is d'. It is calculated by subtracting the standardized false alarm rate from the standardized hit rate. A detection performance of d' = 0 means that the screener had exactly the same hit and false alarm rate – in other words that this screener just guessed. This measure may only be calculated under the assumption that the theoretical signal-plus-noise distribution and the noise distribution are 1) normally distributed (binormal ROC curves) and 2) that their variances are equal. These assumptions can be tested with receiver-operating characteristic (ROC) curves, where the proportion of hits is plotted as a function of the proportion of false alarms at different locations of the criterion. Maximum likelihood (ML) estimation algorithms for fitting binormal ROC curves are available (see [13]-[15]). The second assumption can be tested with the slope of the standardized ROC curve (if the variances are equal the slope of the standardized ROC curve is 1). If the variances are unequal, another signal detection measure, Δm, is often used. One disadvantage of this measure is that it can only be computed when ROC curves are available. D' and Δm express sensitivity in terms of the difference between the means of the noise and signal-plus-noise distribution expressed in units of the noise distribution. If the ROC curves are not binormal, it is still possible to express sensitivity as the

area under the ROC curve.

Another well known measure, which is "nonparametric" (or sometimes also called "distribution-free") is A' and was first proposed by [16]. The term "nonparametric" refers to the fact that the computation of A' requires no a priori assumption about underlying distributions. A' can be calculated when ROC curves are not available and the validity of the normal distribution and equal variance assumptions of the signal-noise and noise distribution can not be verified.

A' can be calculated by the following formula [17]:

$$A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)],$$

whereas H is the hit rate and F the false alarm rate. If the false alarm rate is greater than the hit rate the equation must be modified [18], [19]:

$$A' = 0.5 - [(F - H)(1 + F - H)]/[4F(1 - H)]$$

As [20] have pointed out, this does not mean that these measures are an accurate reflection of their theoretical origin (i.e. that A' reflects the area under a reasonable ROC curve) or that A' is a distribution-free measure or fully independent of a response bias (see also [21]). Thus, the term "nonparametric" is somewhat misleading. A further disadvantage of A' is that it underestimates detection ability by an amount that is a function of the magnitude of bias and decision ability [20]. But because A' can be easily computed and no further assumptions on the underlying noise and signal-plus-noise distribution have to be made, researchers often use this measure when the assumptions of SDT are not fulfilled or cannot be tested. Another measure, which is sometimes used, is the unbiased proportion correct $p(c)_{max}$. This measure can be calculated from d' and used instead of A' (see [7], [20]). As d', it is independent of any response biases. Whenever using $p(c)_{max}$, double high-threshold theory is implied (for detailed information on double-high threshold theory see [22]; summarized in [6]).

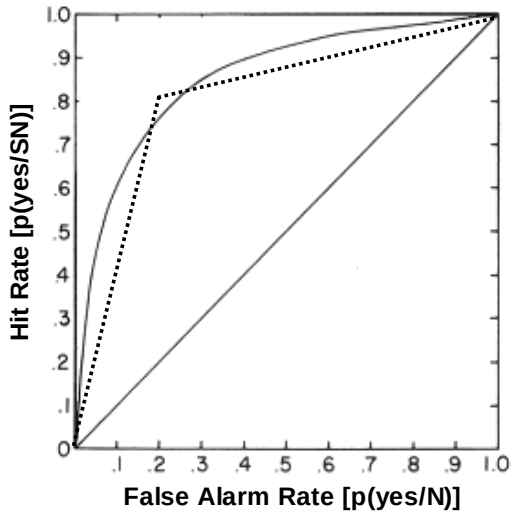

Fig. 1 ROC implied by signal detection theory (solid curve) and by two-state low threshold theory from [9] (dashed lines).

To investigate which performance measures for threat detection in X-ray images are valid and reliable for novices as well as for experts it is important to examine the form of ROC-curves prior and after training. Note that signal detection and threshold theories predict different forms of ROC curves. In linear coordinates the ROC curve predicted from the two-state low threshold theory of [9] are two straight lines, whereas the ROC curve predicted from signal detection theory is a symmetrical curve (see Figure 1).

III. Method

We used a computer-based training (CBT) system for improvised explosive devices (IEDs), which was developed based on object recognition theories and visual cognition. For detailed information on this CBT system (X-Ray Tutor) see [1], [23] and [24].

A. Participants

The original sample size of participants was seventy-two (fifty females) at the age of 23.9 – 63.3 years (M = 48.3 years, SD = 9.0 years). Data of ten participants were not included in the analyses of this study because at least for one test date the slope of the standardized ROC curve was between -0.1 and 0.1. Thus, for the analyses in this study data of sixty-two participants were used. None of the participants had received CBT before.

B. Training design

The detailed design of the data collection, design and material can be found in an evaluation study of the CBT published recently [1]. In summary, four groups of participants had access to the CBT from December 2002 to May 2003. There were four training blocks counterbalanced across four groups of trainees using a Latin Square design. Prior to each training block, performance tests were taken containing the IEDs of the following training block. This method allowed to measure training effectiveness for IEDs never seen before in a standardized way.

Training and testing blocks consisted of sixteen IEDs. For the training, sixteen difficulty levels were constructed by combining each IED with bags of different complexities. At test, only the two most difficult combinations of IEDs and bags were used. To test training effects for different display durations, each bag was presented for 4 and 8 seconds.

All four tests consisted of 128 trials: 16 (IEDs) * 2 (two most difficult levels) * 2 (4 & 8 sec display durations) * 2 (harmless vs. dangerous bags). The order of trial presentation was randomized. For each trial, participants had to judge whether the X-ray image of the bag contained an IED (NOT OK response) or not (OK response). In addition, confidence ratings from 0 (very easy) to 100 (very difficult) were assessed after each decision.

For the purposes of this study we analyzed data of the first detection performance test conducted in Dec/Jan 2003 (Test 1, prior training) and data of the third detection performance test conducted in March/April 2003 (Test 3, after 20 training sessions on average)[1].

IV. Results

In order to plot ROC curves, confidence ratings were divided into 10 categories ranging from 1 (bag contains no IED for sure) to 10 (bag contains an IED for sure). Figure 2 shows the pooled unstandardized ROC curves prior training and after 20 training sessions for display durations of four (a, c) and eight seconds (b, d). As can be seen in Figure 2, the ROC curves seem to be better fitted by two straight lines than by a bimodal ROC curve as would be predicted from SDT.

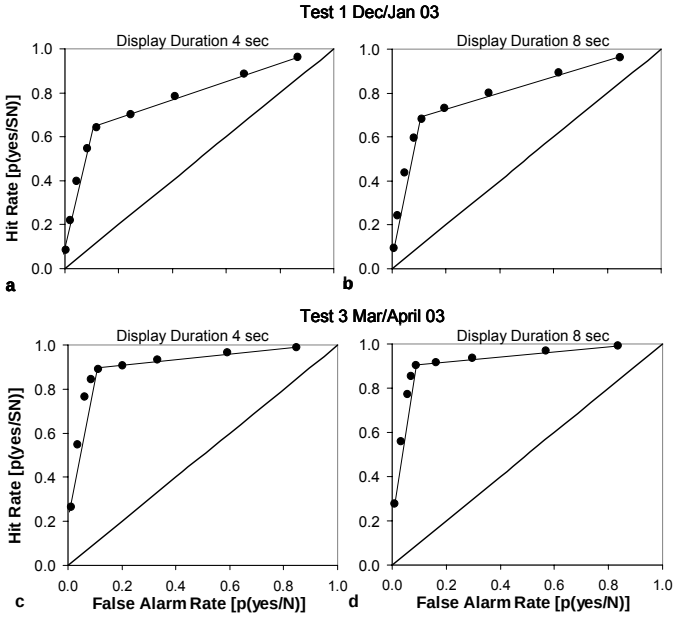

Fig. 2. Unstandardized ROC curves for the two test dates (prior and after training), based on pooled data from 62 participants. a, b) ROC curves prior training with display durations of 4 sec (a) and 8 sec (b). c, d) ROC curves after training with display durations of 4 sec (c) and 8 sec (d).

As mentioned in section II such ROC curves are predicted from two-state low threshold theory ([9]) and are not consistent with the Gaussian distribution assumptions of SDT. Standardized ROC curves are shown in Figure 3 and one can clearly see that none of them is linear as would be predicted by SDT. This was confirmed in individual ROC analyses that revealed significant deviations from linearity for several participants ($\chi 2$-tests). Interestingly, nonlinearity seems to be more pronounced after training (see bottom halves of Figure 2 and 3).

These results suggest the existence of a low threshold and challenge the validity of SDT for explaining the data obtained in this study. As a consequence, the use of parametric SDT measures as reliable and valid estimates of threat detection performance might be questioned – at least as far as detection of IEDs in X-ray images of passenger bags is concerned.

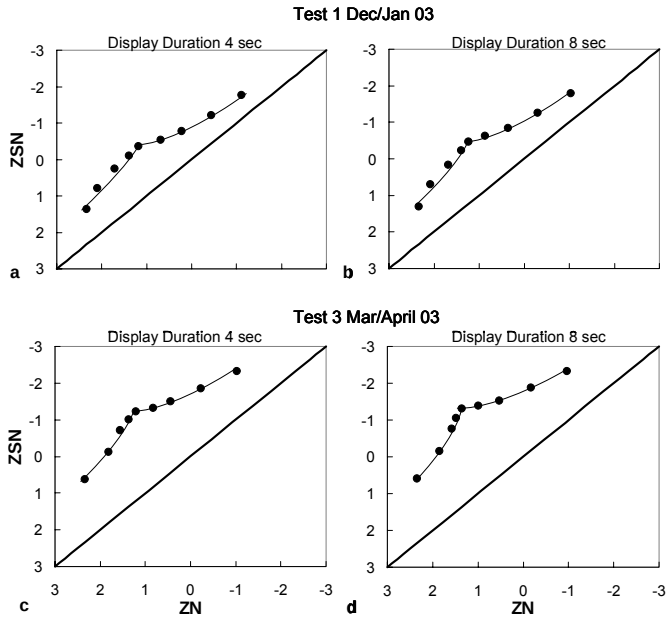

Fig. 3. Standardized detection ROC curves for the two test dates (prior and after training), based on pooled data for the 62 participants. a, b) ROC curves prior training for 4 sec (a) and 8 sec (b). c, d) ROC curves after training for 4 sec (c) and 8 sec (d).

However, it remains to be investigated whether our results can be replicated using other stimuli and whether similar results can be obtained when other threat categories are used (e.g. guns, knives, dangerous goods, etc.).

In any case it is an interesting question to what extent different detection measures correlate. Table 1 shows correlation coefficients (Pearson) between the detection measures d', Δm, Az, A' and proportion correct $p(c)_{max}$ calculated according to [7] for all four conditions (2 test dates and 2 display durations).

TABLE 1
Correlations between detection measures

	Display Duration 4 sec (*r*)					Display Duration 8 sec (*r*)				
Test1 Dec/Jan 03	*d'*	pHit	Δm	Az	A'	*d'*	pHit	Δm	Az	A'
pHit	0.76					0.78				
Δm	0.89	0.74				0.89	0.77			
Az	0.84	0.74	0.95			0.81	0.78	0.93		
A'	0.75	0.73	0.75	0.77		0.80	0.83	0.81	0.88	
$p(c)_{max}$	0.97	0.74	0.88	0.87	0.75	0.94	0.82	0.57	0.82	0.81
Test3 Mar/April 03	*d'*	pHit	Δm	Az	A'	*d'*	pHit	Δm	Az	A'
pHit	0.74					0.74				
Δm	0.64	0.49				0.69	0.53			
Az	0.73	0.79	0.68			0.41	0.42	0.51		
A'	0.77	0.87	0.46	0.70		0.77	0.85	0.51	0.38	
$p(c)_{max}$	0.97	0.80	0.88	0.85	0.85	0.95	0.78	0.61	0.44	0.83

Note. All *p*-values < .01.

[1] Standard deviation was 8 training sessions.

As it can be seen in Table 1, the correlations between the different measures are quite high. In general, there is a tendency of slightly smaller correlations after training, particularly for display durations of 8 sec.

Figure 4 visualizes the training effect using the different

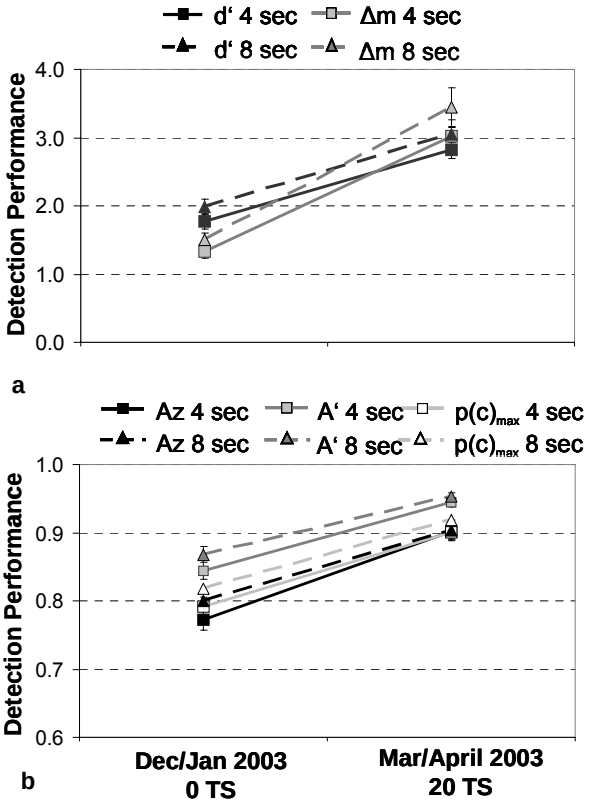

Fig. 4. Illustration of training effect by comparing performance prior training (Test 1, Dec/Jan 2003) and after 20 training sessions on average (Test 3, Mar/April 2003) for display durations of 4 and 8 sec. a) d', Δm, b) Az, A' and $p(c)_{max}$ TS = training sessions.

detection performance measures.

A large training effect is clearly apparent for each detection measure. While substantial differences in slope can be observed between d' and Δm (Figure 4a), the comparison between A', Az and $p(c)_{max}$ reveals relatively parallel lines (Figure 4b). Detection performance is slightly better for display durations of 8 vs. 4 seconds, which is apparent for all measures.

Statistical analyses are reported only for A' since ROC analysis could not support parametric SDT measures. A two-way analysis of variance (ANOVA) with the two within-participants factors test date (prior and after training) and display duration (4 and 8 sec) showed a significant effect of test date, $F(1, 61) = 14.37$, $MSE = 0.001$, $p < .001$, and display duration, $F(1, 61) = 86.95$, $MSE = 0.53$, $p < .001$. Effect sizes were high [25], with $\eta^2 = .19$ for test date, and $\eta^2 = .59$ for display duration. There was no significant interaction between test date and display duration, $F(1, 61) = 3.24$, $MSE = 0.001$, $p = .08$.

Internal reliability was assessed by calculating Cronbachs Alpha using hits (NOT OK response for threat images) and correct rejections (OK responses for non-threat images). Table 2 contains the reliability coefficients for the two test dates (prior and after training) and each group of trainees while pooling display durations of 4 and 8 seconds.

TABLE 2
RELIABILITY ANALYSES (CRONBACH'S ALPHA)

	Dec/Jan 2003 0 TS (Test 1)	Mar/April 2003 20 TS (Test 3)
Group A (N =12)	.92	.92
Group B (N = 15)	.90	.82
Group C (N = 17)	.90	.93
Group D (N = 18)	.91	.90

Note. Internal reliability coefficients broken up by participant group and test date.

V. DISCUSSION

The objective of this study was to compare different measures of X-ray detection performance while taking effects of training into account. To this end, computer based tests were used that were conducted before and after CBT.

From a regulators perspective the hit rate is sometimes the preferred measure when data from threat image projection (TIP) is used to judge the performance of screeners. However, the hit rate alone is not a valid measure because it is not possible to distinguish between good detection ability and a liberal response bias. For example an anxious screener might achieve a high hit rate only because most bags are judged as being NOT OK. In this case security is achieved at the expense of efficiency, which would be reflected in long waiting lines at the checkpoint. It would be much more beneficial to achieve a high degree of security without sacrificing efficiency. This implies a high hit rate and a low false alarm rate. SDT provides several measures that take the hit and false alarm rate into account in order to achieve more valid measures of detection performance. Although the use of parametric SDT measures is still very common (e.g. [26], [5], [1]), several studies have used "nonparametric" A' because its computation does not require a priori assumptions about the underlying distributions (e.g. [27], [28], [4]). ROC analysis can be used to test whether the assumptions of SDT are fulfilled. We found that standardized ROCs deviated from linearity both before and after training of IED detection. Interestingly, unstandardized ROC curves could be fitted very well by two straight lines, just as would be predicted from two-state low threshold theory of [9]. These results challenge the validity of SDT measures as estimates of threat detection performance, at least when detection of IEDs in X-ray images of passenger bags is concerned. It certainly remains to be investigated whether our results can be replicated with different stimulus material and other threat items than IEDs. In any case however, our findings suggest that other detection performance measures than those from SDT should be considered, too. As mentioned above, the calculation of A'

requires no a priori assumption about the underlying distributions, which has often been regarded as an advantage over SDT measures such as d' and Δm. In many applications such as risk analysis, quality performance measurement, and competency assessment based on TIP or CBT data, only hit and false alarm rates are available and multipoint ROCs can not be obtained to test the assumptions of SDT measures. At least in these cases it should be considered to use A' in addition to d', while certainly both measures are more valid estimates of detection performance than the hit rate alone.

Finally, it should be noted that the five psychophysical measures compared in this study were usually strongly correlated. More specifically, the measures that are most often reported in the detection literature, A' and d', correlated in all four test conditions with $r >= .75$. And in a recent study using computer-based tests with different types of threat items even higher correlations between A' and d' were found ($r > .90$, [4]).

ACKNOWLEDGMENT

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

[2] M. Graf, A. Schwaninger, C. Wallraven, and H.H. Bülthoff, "Psychophysical results from experiments on recognition & categorization," Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis; IST-2000-29375, 2002.

[3] A. Schwaninger, "Object recognition and signal detection," in Praxisfelder der Wahrnehmungspsychologie, B. Kersten and M.T. Groner, Eds., Bern: Huber, in press.

[4] A. Schwaninger, H. Hardmeier, and F. Hofer, "Measuring visual abilities and visual knowledge of aviation security screeners," *IEEE ICCST Proceedings,* this volume.

[5] J.S. McCarley, A., Kramer, C.D. Wickens, E.D. Vidoni, and W.R. Boot, "Visual Skills in Airport-Security Screening," *Psychological Science,* vol. 15, pp. 302-306, 2004.

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

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

[8] D.H. Krantz, "Threshold theories of signal detection," *Psychological Review*, vol. 76, pp. 308-324, 1969.

[9] R.D. Luce, "A threshold theory for simple detection experiments," *Psychological review*, vol. 70, pp. 61-79, 1963.

[10] G.A. Gescheider, *Psychophysics: The Fundamentals* (3rd Ed), Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 1998.

[11] J.A. Swets, "Is there a sensory threshold?," *Science*, vol. 134, pp. 168-177, 1961.

[12] J.A. Swets, W.P.Jr. Tanner, and T.G. Birdsall, "The evidence for a decision-making theory of visual detection," Electronic Defense Group, University of Michigan, Tech. Rep. No. 40, 1955.

[13] J.A. Swets and R.M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* New York: Academic Press, 1982.

[14] C.E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology*, vol. 24, pp. 234-245, 1989.

[15] D.D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating method data," *Journal of Mathematical Psychology*, vol. 6, pp. 487-496, 1969.

[16] I. Pollack and D.A. Norman, "A non-parametric analysis of recognition experiments," *Psychonomic Science*, vol. 1, pp. 125-126, 1964.

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

[18] D. Aaronson and B. Watt, "Extensions of Grier's computational formulas for A' and B'' to below-chance performance," *Psychological Bulletin,* vol. 102, pp. 439-442, 1987.

[19] J.G. Snodgrass and J. Corwin, "Pragmatics of measuring recognition memory: Applications to dementia and amnesia," *Journal of Experimental Psychology: General,* vol. 117, pp. 34-50, 1988.

[20] 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, pp. 556-569, 2003.

[21] N.A MacMillan, and C.D. Creelman, "Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias," *Psychonomic Bulletin & Review*, vol. 3, pp. 164-170, 1996.

[22] J.P. Egan, "Recognition memory and the operating characteristic," Hearing and Communication Laboratory, Technical Note AFCRC-TN-58-51, Indiana University, 1958.

[23] A. Schwaninger, "Training of airport security screeners," *AIRPORT*, *05*, pp. 11-13, 2003.

[24] A. Schwaninger, "Computer based training: a powerful tool to the enhancement of human factors," *Aviation security international*, February, pp. 31-36, 2004 .

[25] J. Cohen, *Statistical power analysis for the behavioral sciences.* New York: Erlbaum, Hillsdale, 1988.

[26] J.A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics – Collected Papers,* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 1996.

[27] A.D. Fisk and W. Schneider, "Control and Automatic Processing during Tasks Requiring Sustained Attention: A New Approach to Vigilance," *Human Factors*, vol.23, pp. 737-750, 1981.

[28] G.C. Prkachin, "The effects of orientation on detection and identification of facial expressions of emotion", *British Journal of Psychology*, vol. 94, pp. 45-62, 2003.