

THE X-RAY OBJECT RECOGNITION TEST (X-RAY ORT) – A RELIABLE AND VALID INSTRUMENT FOR MEASURING VISUAL ABILITIES NEEDED IN X-RAY SCREENING

Diana Hardmeier, Franziska Hofer, Adrian Schwaninger

Department of Psychology, University of Zurich, Switzerland

ABSTRACT

Aviation security screening has become very important in recent years. It was shown in [1] that certain image-based factors influence detection when visually inspecting x-ray images of passenger bags. Threat items are more difficult to recognize when placed in close-packed bags (effect of bag complexity), when superimposed by other objects (effect of superposition), and when rotated (effect of viewpoint). The X-Ray Object Recognition Test (X-Ray ORT) was developed to measure the abilities needed to cope with these factors. In this study, we examined the reliability and validity of the X-Ray ORT based on a sample of 453 aviation security screeners and 453 novices. Cronbach Alpha and split-half analysis revealed high reliability. Validity was examined using internal, convergent, discriminant and criterion-related validity estimates. The results show that the X-Ray ORT is a reliable and valid instrument for measuring visual abilities needed in x-ray screening. This makes the X-Ray ORT an interesting tool for competency and pre-employment assessment purposes.

1. INTRODUCTION

One of the most important tasks in airport security screening is the visual inspection of passenger bags using x-ray imaging systems. During rush hours, screeners have only a few seconds to decide whether a bag is OK (i.e. it

contains no prohibited item) or NOT OK (i.e. it contains a prohibited item). Understanding the underlying visual cognition processes of this task is very important in order to train and select people appropriately for the x-ray screening job. A screener has to know which items are prohibited and what they look like in x-ray images of passenger bags. This is dependent on training and expertise [2, 3]. In addition to such knowledge-based factors, [1] and [4] have identified three image-based factors, which are illustrated in Figure 1. Threat items are more difficult to detect when they are in a close-packed bag (effect of bag complexity). Objects in x-ray images are often superimposed by other objects, which can also affect detection performance (effect of superposition). When threat objects are rotated they can become more difficult to recognize (effect of viewpoint).

The X-Ray Object Recognition Test (X-Ray ORT) is a tool to measure the visual abilities needed to cope with these image-based factors [1, 4]. In this study we examined the reliability and validity of the X-Ray ORT. Reliability measures, such as Cronbach's Alpha and split-half reliabilities were assessed with two groups (novices and experts) of 453 participants each. Validity estimates included internal, convergent, discriminant and criterion-related measures.

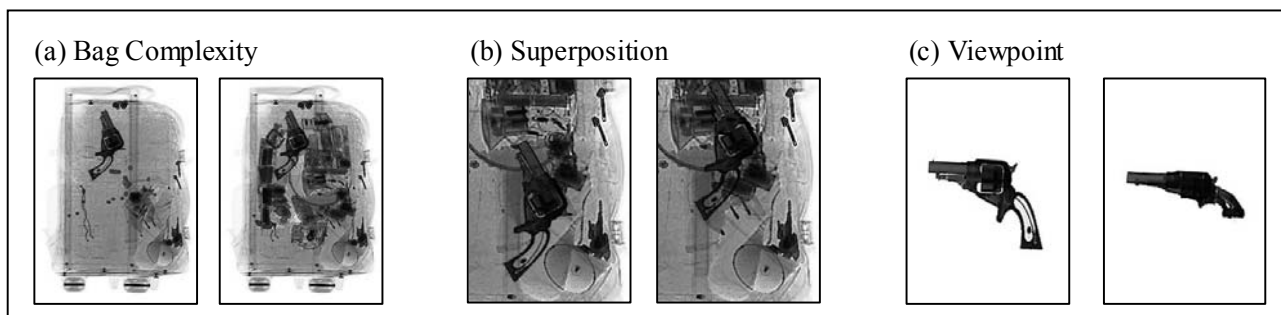


Figure 1: Image-based factors relevant in x-ray screening: (a) bag complexity, (b) superposition, (c) viewpoint.

2. METHOD

2.1. Participants

453 aviation security screeners (141 male and 312 female) between 24 and 65 years ($M = 48.94$ years, $SD = 9.09$ years) and 453 novices (333 male and 120 female) between 19 and 56 years ($M = 36.44$ years, $SD = 10.77$ years) participated in this study. All screeners had at least three years of experience in x-ray screening.

2.2. Materials and Procedure

In the X-Ray ORT, x-ray images of passenger bags are shown in black and white only because novices do not know how to interpret color information (which is in fact used to code different materials). To further reduce knowledge-based factors resulting from training or experience, only guns and knives with common shapes are used in the X-Ray ORT. Moreover, all threat items are presented before the test starts (8 guns for ten seconds followed by 8 knives for 10 seconds). Half of the threat items are shown in a frontal view, the other half in a rotated view.

All threat items are combined with bags of different bag complexities (low and high) using different levels of superposition (low and high). Each threat item is shown from two viewpoints (easy vs. difficult). The difficulty levels of bag complexity, superposition and viewpoint were determined visually by two raters. Each bag was used twice, once with a threat item (threat image) and once without (harmless image). Thus, the X-Ray ORT consists of a total of 256 test trials: 2 threat categories (guns, knives) * 8 (exemplars) * 2 (bag complexities) * 2 (superpositions) * 2 (views) * 2 (harmless images vs. threat images). Based on results from pilot studies, guns were more superimposed by other items in the bag and more rotated than knives in order to achieve a similar level of image difficulty.

The task in the X-Ray ORT is to visually inspect x-ray images of passenger bags for the presence of a gun or a knife. Each image is presented for 4 seconds on the screen in order to match visual inspection times at airports during periods of high passenger flow. For each trial, test candidates have to decide whether the bag is OK (no threat item included) or NOT OK (gun or a knife included) and indicate on a 90 point rating scale how sure they are in their decision (confidence ratings). All responses are made by clicking buttons on the screen. By pressing the space bar, the next trial is initiated.

Before the actual test starts, candidates are exposed to several screens with instructions as well as eight practice trials (half of them with a threat item and half of them without). None of the threat items and bags of the practice trials are used in the actual test. Whereas practice trials contain feedback on whether the correct response was given (OK vs. NOT OK), the actual test does not contain any feedback. The test is subdivided into four blocks and

participants are allowed to take a short break after finishing a block. Trials are randomized within each block and block order is counterbalanced across four groups of participants using a Latin square design. The X-Ray ORT takes about 45 minutes to complete.

3. RESULTS

3.1. Reliability of the X-Ray ORT

Cronbach Alpha and Guttman split-half reliabilities were calculated for novices and experts. Reliability measures were based on hits and correct rejections (PC = percentage correct) as well as on confidence ratings (CR). Reliability was calculated separately for x-ray images of bags including a threat item (SN trials) and for x-ray images of harmless bags (N trials). The high reliability coefficients in Table 1 show that the X-Ray ORT is a reliable instrument for measuring visual abilities that are needed when visually inspecting x-ray images of passenger bags.

TABLE I
RELIABILITY ANALYSES

Reliability Coefficients		PC SN	PC N	CR SN	CR N
Screeners	Alpha	.887	.944	.926	.966
	Split-half	.781	.840	.840	.904
ORT	Alpha	.907	.946	.932	.970
	Split-half	.778	.871	.807	.939

Note. PC = Percent Correct, CR = Confidence Ratings, SN = Bags containing a threat ("Signal plus Noise Trials"), N = Bags containing no threat ("Noise-Trials")

3.2. Validity of the X-Ray ORT

Individual A' scores were calculated based on the percentage of hits and false alarms over all trials of the X-Ray ORT for each participant. The advantage of A' over d' is that it requires no a priori assumption about the underlying noise and signal plus noise distributions. For further information on these and other detection measures see [5, 6, 7].

3.2.1. Internal validity

Individual A' scores were subjected to a three-way analysis of variance (ANOVA) with bag complexity, superposition and view difficulty as within-participant factors. This analysis was done for both groups of participants (experts and novices) separately. The main effects are displayed in Figure 2. ANOVA results of aviation security screeners showed highly significant main effects of bag complexity (low vs. high) with an effect size of $\eta^2 = .80$, $F(1, 452) = 1851.83$, $p < .001$, superposition (low vs. high) $\eta^2 = .55$, $F(1, 452) = 548.10$, $p < .001$, and

view (easy vs. difficult) $\eta^2 = .70$, $F(1, 452) = 1044.01$, $p < .001$. Some interactions reached statistical significance but

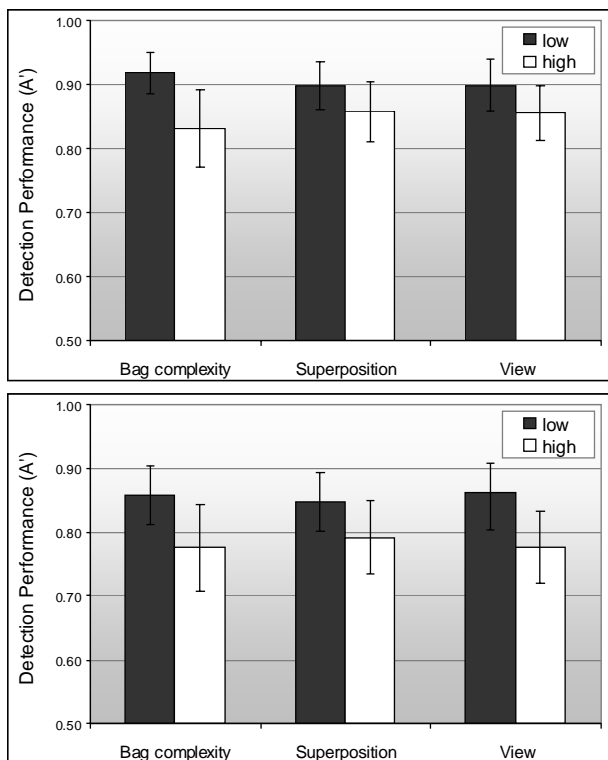


Figure 2: Effects of image-based factors in the X-Ray ORT, error bars represent standard deviations: TOP: aviation security screeners, BOTTOM: novices.

their effect sizes η^2 were small relative to the effect sizes of the main effects: bag complexity * superposition $\eta^2 = .06$, $F(1, 452) = 27.69$, $p < .001$, superposition * view $\eta^2 = .08$, $F(1, 452) = 37.90$, $p < .001$ and bag complexity * superposition * view $\eta^2 = .01$, $F(1, 452) = 6.55$, $p < .05$.

Similar results were observed for novices. There were again highly significant main effects with large effect sizes: bag complexity (low vs. high) $\eta^2 = .69$, $F(1, 452) = 1012.20$, $p < .001$, superposition (low vs. high) $\eta^2 = .64$, $F(1, 452) = 817.19$, $p < .001$, and view (easy vs. difficult) $\eta^2 = .72$, $F(1, 452) = 1137.67$, $p < .001$. Again, some interactions were significant, but their effect sizes were rather small when compared to the effect sizes of the main effects. bag complexity * superposition $\eta^2 = .10$, $F(1, 452) = 48.01$, $p < .001$, bag complexity * view $\eta^2 = .10$, $F(1, 452) = 51.25$, $p < .001$, superposition * view $\eta^2 = .11$, $F(1, 452) = 55.35$, $p < .001$ and bag complexity * superposition * view $\eta^2 = .02$, $F(1, 452) = 8.64$, $p < .01$.

In summary, large main effects of image-based factors (bag complexity, superposition and view difficulty) were found both for novices and experts. The large variances between individuals (see standard deviations in Figure 2) show that people differ remarkably with regard to how well they can cope with image difficulty resulting from these image-based effects. Interestingly, only small mean

differences in A' between novices and experts were found. This is consistent with the assumption that the X-Ray ORT measures relatively stable visual abilities that are needed to cope with effects of bag complexity, superposition and view difficulty. However, we are currently conducting further studies in order to investigate whether these abilities can be trained when using an individually adaptive computer-based training system (X-Ray Tutor) that takes the image-based effects into account.

3.2.2. Convergent and discriminant validity

Convergent validity was examined using the X-Ray ORT data from the aviation security screener group since all of them have taken also the Prohibited Items Test (PIT). The PIT is an image interpretation competency test using color x-ray images that contain different kinds of forbidden objects according to international prohibited items lists (for details see [1]). A' scores in the ORT correlated significantly with A' scores in the PIT, $r = .61$, $p < .001$, indicating high convergent validity. Discriminant validity was tested by correlating the X-Ray ORT with results obtained with the CBQ. The CBQ is a computer based multiple choice questionnaire about safety and security regulations on airports. As expected, the correlation with the X-Ray ORT was rather low, $r = .27$, indicating sufficient discriminant validity.

3.2.3. Criterion-related validity

Criterion-related validity was examined by correlating X-Ray ORT scores with on the job performance measured with threat image projection (TIP). With this technology of current x-ray screening equipment it is possible to display fictional threat images during regular x-ray screening operations. Screeners receive feedback after each TIP image so that no negative impact on the screening operation occurs. The TIP library used in this study consisted of 1028 combined threat images (CTIs). These CTIs were created as follows: 64 improvised explosive devices (IEDs) were combined with 8 bags of different image difficulties rated by 8 x-ray screening experts. Each bag was also displayed without the IED. A TIP to bag ratio of 1 to 30 was used. This means that one TIP image was shown within about 30 x-ray images of real passenger bags. Half of the TIPs contained a threat item, the other half did not. This allowed obtaining valid hit and false alarm rates (see [6] for further information). TIP data was available from 86 screeners. On the job performance was estimated using TIP data aggregated over 17 months. A' and d' scores were calculated from hit and false alarm rates in TIP and in the X-Ray ORT. Large correlations between X-Ray ORT and TIP performance were found: $r = .41$, $p < .001$ for A' scores and $r = .51$, $p < .001$ for d' scores. These rather high correlations suggest that the abilities measured by the X-Ray ORT are indeed very important determinants of on the job performance in x-ray screening.

4. DISCUSSION

According to [1] and [4] detection of threat items in x-ray images of passenger bags depends on image-based factors such as bag complexity, superposition by other objects, and view difficulty of the threat item resulting from its position within the bag. The X-Ray ORT has been developed to measure how well people can cope with these image-based factors. In this study, the reliability and validity of the X-Ray ORT was examined. Cronbach Alpha and split half analyses revealed that this test is a very reliable instrument. Cronbach Alpha coefficients were found to be high in both samples of participants ($\alpha > .89$ for experts and $\alpha > .91$ for novices). Further evidence of reliability was revealed by split-half reliabilities (Guttman) which were quite high as well ($r > .78$ for both groups). Internal validity was examined using ANOVA. Highly significant main effects with large effect sizes were found for bag complexity, superposition and view difficulty. Whereas some interactions reached statistical significance, their effect sizes were rather small when compared to the main effects. This indicates high internal validity regarding the assumption of three image-based factors that are conceptually independent. It should also be noted that large differences between individuals were found with regard to how well they could cope with effects of bag complexity, superposition and view difficulty. Interestingly, this accounted both for novices and screeners. Convergent validity was assessed by correlating X-Ray ORT scores with the results in the PIT, which is a computer-based image interpretation competency test. The large correlation of $r = .61$ supported convergent validity. Discriminant validity was estimated by correlating with the CBQ, a computer-based multiple choice exam on theoretical knowledge needed in airport security operations. In order to support discriminant validity a low correlation should be found. This was the indeed case since the X-Ray ORT correlated with CBQ scores only with $r = .24$. Criterion-related validity was calculated by correlating detection scores in the X-Ray ORT with on the job performance measured using threat image projection (TIP). Correlations of $r = .41$ using A' scores and $r = .51$ using d' scores indicated good criterion-related validity.

In summary, the results of different reliability and validity analyses showed that this test provides a very useful, reliable and valid instrument to assess visual abilities needed in x-ray screening of passenger bags. This makes the X-Ray ORT an interesting tool for competency and pre-employment assessment purposes in airport security and other areas in which x-ray screening is applied.

5. REFERENCES

- [1] Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings*, 38, 258-264.
- [2] Schwaninger, 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* (pp. 147-156). Wessex: WIT Press.
- [3] Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International*, FEB/2004, 31-36.
- [4] Schwaninger, A. (2003b). Evaluation and selection of airport security screeners, *AIRPORT*, 02/2003, 14-15.
- [5] Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*, New York: Wiley.
- [6] Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings*, 38, 303-308.
- [7] MacMillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. Cambridge: University Press.

6. 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.