Screener Evaluation and Selection

a b c

The importance of aviation security has increased dramatically in recent years. As a consequence of the new threat situation it has become clear that certain aspects of airport security are in need of improvement. State-of-the-art x-ray technologies provide high resolution images, many image processing features and even automatic explosive detection systems. However, the last decision is always made by human operators. It is being realized more and more that the best equipment is of limited use, if the personnel who operates it is not selected well and trained enough. This is especially important because according to several aviation security experts the human operator is currently the weakest link in airport security.

Fig. 1: Relevant factors of threat detection. a Rotated objects are more difficult to recognize (viewpoint-dependence), b when objects are superimposed by other objects, detection performance decreases (effect of superposition), c identifying a threat item in a close-packed bag is more difficult (effect of bag complexity).

Human factors have gained much attention recently, and it has become clear that effective selection, evaluation and training of airport security personnel are crucial factors for increasing airport security and efficiency. Since June 2000 scientists from the University of Zurich have investigated human factors in x-ray screening. The research projects were conducted in close collaboration with Zurich State Police, Airport Division and were funded by Zurich Airport. Important insights were revealed for the following topics: (1) reliable measurements of threat detection, (2) screener evaluation and selection, (3) training of screeners, and (4) pre-employment assessment. In AIRPORT 3/2002 an overview of these studies was presented (page 20-21). In AIRPORT 1/2003 the first topic was presented in more detail (page 22–23). In this article topic (2) is discussed, i.e. how scientific methods can be applied to build reliable tests for screener evaluation and selection.

X-ray object recognition test X-ray ORT 1.0

In a recent study at Zurich Airport a new test was validated, which measures different visual abilities that are relevant for detecting threats in x-ray images. The development of this test was based on the results of scientific studies on visual cognition, i.e. the scientific investigation of how we process visual information. Many studies on object recognition have shown that recognition performance is often dependent on viewpoint. Therefore, a gun might be more difficult to detect when it is rotated (Fig. 1a). Since forbidden objects have to be detected within a baggage, the degree by which they are superimposed by other objects could influence detection, too (Fig. 1b). Because luggage differs remarkably in terms of complexity, a threat item could be easier to detect in a baggage, which contains relatively few other objects as opposed to a close-packed baggage (Fig. 1c). In addition to these three aspects, general detection performance across all conditions was calculated. A total of 80 airport security screeners took the test. The data was analyzed using signal detection theory, which provides scientific methods for measuring threat detection performance (see AIRPORT 1/2003, page 22–23). Figure 2 shows the main findings concerning viewpoint-dependence, superposition and baggage complexity. When guns or knifes were rotated, they were more difficult to detect, as indicated by the decrease in detection performance (d') in Figure 2, left. This result clearly shows that threat detection is viewpoint-dependent. Similarly, when forbidden objects were superimposed by other objects, it was more difficult to detect them, as illustrated in the middle of Figure 2 (effect of superposition). Detection was also impaired when a threat was in a close-packed bag, confirming the importance of baggage complexity (Fig. 2, right). All effects were highly significant, for weapons as well as for guns ($p<.001$). While this is the first scientific study that has measured the effects of viewpoint, superposition and baggage complexity it should be mentioned that airport security screeners know such effects from their own experience on the job since a long time. For example screeners usually put the bag on the conveyor belt themselves in order to

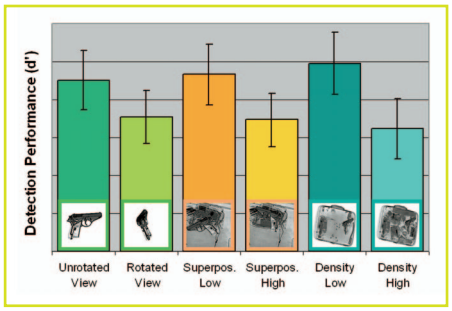

Fig. 2: Results of a study based on 80 screeners who took the x-ray object recognition test X-ray ORT 1.0. Rotation, superposition and density significantly affected threat detection performance (p<.001). Thin black bars represent the standard deviation between screeners.

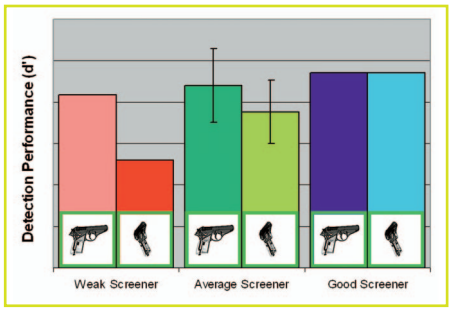

Fig. 3: Illustration for the remarkable differences between screeners (detection of guns only). The red bars on the left illustrate the performance of a relatively weak screener, who's detection performance is strongly impaired when a gun is rotated. The green bars in the middle represent the mean detection performance of all screeners who took the test. The blue bars on the right are the results of a good screener, who's detection performance is not dependent on viewpoint.

avoid that a passenger puts the bag on the conveyor belt in a way that could make a forbidden object difficult to recognize. Moreover, bags usually are hand-searched if a potential threat item is superimposed by other objects or if a baggage is close-packed.

A very interesting result was the fact that screeners differed remarkably in their detection performance. This is illustrated in Figure 3 for the aspect of view-dependent detection. The two red bars on the left of the figure are the detection performance values (d') of a relatively weak screener. Although this person was able to detect full views of threat objects in bags, detection was heavily impaired when the forbidden objects were rotated. The two green bars in the middle of Figure 3 represent the mean detection performance of all screeners who took the test (see also Fig. 2, left). The blue bars on the right of Figure 3 depict the results of a good screener, who can detect threats very well even if they are rotated. Such large differences between screeners were also found for the effects of superposition and density. The standard deviation (a statistical measure of variance) was relatively high in all conditions, confirming large differences between individuals (thin bars in Fig. 2 and Fig. 3). Therefore, it could be possible to increase security and efficiency substantially if good screeners are selected for demanding jobs and weak screeners receive special training. Moreover, reliable tests of x-ray detection performance are essential elements in a screener certification system. But what makes a good test? What are the basic elements to look for when judging the quality of a test?

Reliability, validity and standardisation

In order to judge the quality of a test three aspects need to be considered: reliability, validity, and standardisation. Reliability refers to the "consistency" or "repeatability" of measures and has to do with the quality of measurement. If for example an IQ test yields a score of 90 for an individual today and 125 a week later, it is not reliable. Reliability also can be a measure of a test's internal consistency, i.e. all questions or images should be measuring the same ability. Good tests have reliability coefficients which range from a low of .70 to above .90 (the theoretical maximum is 1.00). While there are different ways to compute the reliability of a test it is generally assumed that KR20 or Cronbach's alpha are the most accurate estimates of reliability available within classical test theory. The current version of the above mentioned x-ray object recognition test X-ray ORT 1.0 has a Cronbach alpha of .90, which makes it very useful for selection, evaluation and certification purposes.

Validity indicates whether a test is able to measure what should be measured. The face validity of X-ray ORT 1.0 is obviously very high, because the test is quite similar to what screeners do at work. The term concurrent validity refers to whether a test can distinguish between groups that it should be able to distinguish between. We compared the detection performance of students from the University of Zurich with the detection performance of airport security screeners from Zurich State Police, Airport Division. The differences were significant, which is an indicator of concurrent validity. Since in X-ray ORT 1.0 only knifes and guns are used and anybody knows how these objects look like, this test could also be used in a pre-employment assessment. Indeed, some students achieved quite high detection scores, suggesting that the test does measure relevant cognitive abilities that are not dependent on expertise. In order to establish convergent validity it has to be shown that measures that should be related are indeed related. For instance, if X-ray ORT 1.0 does measure general abilities that are important for threat detection, one would expect that if someone is good in detecting guns this person would also achieve a good detection performance for knifes. The convergent validity coefficient of the current version of X-ray ORT was r = .57, indicating that this test does indeed measure rather general threat detection abilities. Another validity measure is called predictive validity. For example it is planned to investigate how well X-ray ORT 1.0 can predict detection performance when used in a pre-employment assessment. Note however, that many tests have validity coefficients (correlations) of around .30 with "real world" behaviour and values higher than .50 are not often achieved. These aren't high correlations, and they emphasize the need to use an x-ray detection test in conjunction with other measures of relevant cognitive factors. This is the reason why the scientists at the University of Zurich use multiple regression and structural equation modelling in order to achieve a higher predictive validity than traditional approaches.

The third important aspect for judging the quality of a test is standardisation. Essentially, this involves trying out the test on a representative group of people in order to establish norms. When an individual takes the test, it can then be determined how far above or below the average her or his score is, relative to the normative group. Note that it is quite important to know how the normative group was selected. For instance, a good standardisation for a test that is used in order to evaluate the detection performance of screeners would imply that a large and representative sample of screeners take the test in order to serve as a good normative group.

In sum, this article has shown that scientific methods and results from visual cognition, signal detection and psychometrics can be combined in order to build reliable and valid tests. This can provide a solid basis for the selection, evaluation and certification of screeners, which are topics of increasing importance in airport security.

Adrian Schwaninger · University of Zurich
aschwan@allgpsy.unizh.ch