Assessment and Certification of X-Ray Image Interpretation Competency of Aviation Security Screeners

Report of VIA Project

Work Package 6

Version 2008-10-15

Adrian Schwaninger[1,2], Saskia M. Koller[1], and Anton Bolfing[1,2]

1)
University of Zurich
Department of Psychology
Visual Cognition Research Group (VICOREG)
Binzmühlestrasse 14/22
8050 Zurich
Switzerland

2)
Max Planck Institute for Biological Cybernetics
Department Human Perception,
Cognition and Action
Spemannstrasse 38
72076 Tübingen
Germany

Assessment and Certification of X-Ray Image Interpretation Competency of Aviation Security Screeners

Adrian Schwaninger, Saskia M. Koller, and Anton Bolfing

Competency Assessment in Aviation Security Screening

In response to the increased risk of terrorist attacks, large investments have been made in recent years in aviation security technology. However, the best equipment is of limited value if the people who operate it are not selected and trained appropriately to perform their tasks effectively and accurately. Latterly, the relevance of human factors has increasingly been recognized. One important aspect of human factors is the competency of the personnel who conducts security screening at airports (aviation security screeners) and assessment of their competency.

Competency assessment maintains the workforce certification process. The main aim of certification procedures is to ensure that adequate standards in aviation security are consistently and reliably achieved. Certification of aviation security screeners can be considered as providing quality control over the screening process. Using certification tests, important information on strengths and weaknesses in aviation security procedures in general as well as on each individual screener can be obtained. As a consequence, certification can also be a valuable basis for qualifying personnel, measuring training effectiveness, improving training procedures, and increasing motivation. In short, certification and competency assessment can be important instruments in improving aviation security.

The implementation of competency assessment procedures presents several challenges. First, what should be assessed has to be identified. Then, there should be consideration of how procedures for the certification of different competencies can be implemented. Another important challenge is international standardization, since several countries, organizations, and even companies are independently developing their own certification or quality control systems.

The following international documents refer to the certification and competency assessment of aviation security staff:

- EU Regulation 2320/2002
- ICAO Annex 17, 3.4.3[1]
- ICAO-Manual on Human Factors in Civil Aviation Security Operations (Doc. 9808)[2]
- ICAO Human Factors Training Manual (Doc. 9683), Part 1, Chapter 4, and Appendix 6, Appendix 32[3]
- ICAO Security Manual for Safeguarding Civil Aviation against Acts of Unlawful Interference, Doc. 8973, Chapter 4, I-4–45[4]
- ECAC Doc. 30, Chapter 12, and Annex IV-12A[5]

ECAC Doc. 30 of the European Civil Aviation Conference specifies three elements for *initial* certification of aviation security screeners:

- an X-ray image interpretation exam
- a theoretical exam
- a practical exam

The *periodical* certification should contain a theoretical exam and an X-ray image interpretation exam. Practical exams can be conducted if considered necessary.

This chapter covers the first element, that is, how to examine competency in X-ray image interpretation. First, human factors best practice guidance for assessing the X-ray image interpretation competency of aviation security screeners is provided. Three different possibilities are discussed, which can serve to measure competency in X-ray image interpretation: covert testing, threat image projection (TIP), and computer-based image tests. Second, on-the-job assessment of screener competency using TIP is discussed. Third, an example of a reliable, valid, and standardized computer-based test is presented which is now used at more than 100 airports worldwide to measure competency in X-ray image interpretation and also for certification purposes, the X-Ray Competency Assessment Test (X-Ray CAT). Fourth, the application of this test in an EU-funded project (the VIA Project) at several European airports is presented.

Requirements for Assessing Competency

One of the most important tasks of an aviation security screener is the interpretation of X-ray images of passenger bags and the identification of prohibited items within these bags. Hit rates, false alarm rates, and the time used to visually inspect an X-ray image of a passenger bag are important measures that can be used to assess the effectiveness of screeners at this task. A hit refers to detecting prohibited items in an X-ray image of a passenger bag. The hit rate refers to the percentage of all X-ray images of bags containing a prohibited item that are correctly judged as being NOT OK. If a prohibited item is reported in an X-ray image of a bag that does not contain such an item, this counts as a false alarm. The false alarm rate refers to the percentage of all harmless bags (i.e., bags not containing any prohibited items) that are judged by a screener as containing a prohibited item. The time taken to process each bag is also important, as it helps in determining throughput rates and can indicate response confidence.

The results of an X-ray image interpretation test provide very important information for civil aviation authorities, aviation security institutions, and companies. Moreover, failing a test can have serious consequences, depending on the regulations of the appropriate authority. Therefore, it is essential that a test should be fair, reliable, valid, and standardized. In the last 50 years, scientific criteria have been developed that are widely used in psychological testing and psychometrics. These criteria are essential for the development of tests for measuring human performance. A summary of the three most important concepts, namely reliability, validity, and standardization, is now presented.

Reliability

We mean reliability to refer to the “consistency” or “repeatability” of measurements. It is the extent to which the measurements of a test remain consistent over repeated tests of the same participant under identical conditions. If a test yields consistent results for the same measure, it is reliable. If repeated measurements produce different results, the test is not reliable. If, for example, an IQ test yields a score of 90 for an individual today and 125 a week later, it is not reliable. The concept of reliability is illustrated in Figure 1. Each point represents an individual person. The x-axis represents the test results in the first measurement and the y-axis represents the scores in the same test of the second measurement. Figures 5.1a–c represent tests of different reliability. The test in Figure 1a is not reliable. The score a participant achieved in the first measurement does not correspond at all with the test score in the second measurement.

The reliability coefficient can be calculated by the correlation between the two measurements. In Figure 1a, the correlation is near zero, that is, r = 0.05 (the theoretical maximum is 1). The test in Figure 1b is somewhat more reliable. The correlation between the two measurements is 0.50. Figure 1c shows a highly reliable test with a correlation of 0.95.

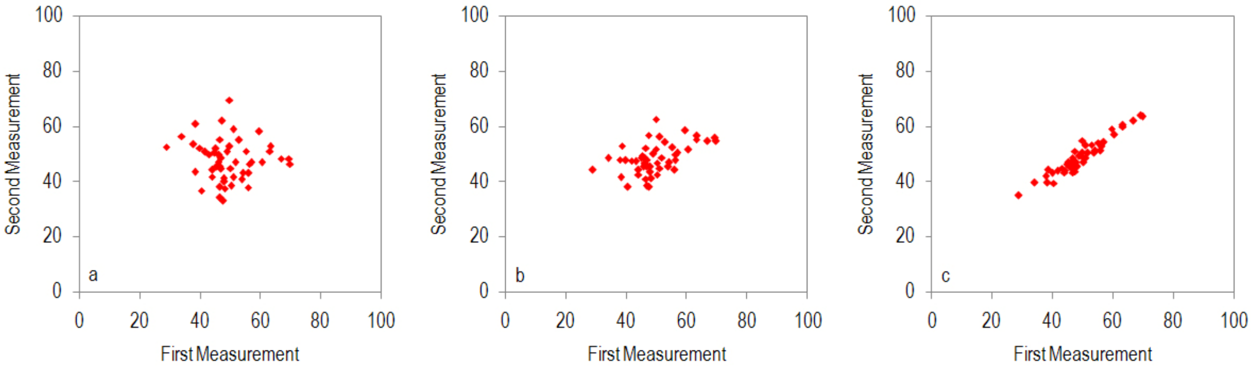


Figure 1 Illustration of different correlation coefficients.
Left: r = 0.05, middle: r = 0.50, right: r = 0.95.

The reliability of a test may be estimated by a variety of methods. When the same test is repeated (usually after a time interval during which job performance is assumed not to have changed), the correlation between the scores achieved on the two measurement dates can be calculated. This measure is called *test-retest reliability*. A more common method is to calculate the *split-half reliability*. In this method, the test is divided into two halves. The whole test is administered to a sample of participants and the total score for each half of the test is calculated. The split-half reliability is the correlation between the test scores obtained in each half. In the alternate forms method, two tests are created that are equivalent in terms of content, response processes, and statistical characteristics. Using this method, participants take both tests and the correlation between the two scores is calculated (*alternate forms reliability*). Reliability can also be a measure of a test's internal consistency. Using this method, the reliability of the test is judged by estimating how well the items that reflect the same construct or ability yield similar results. The most common index for estimating the internal reliability is Cronbach's alpha. Cronbach's alpha is often interpreted as the mean of all possible split-half estimates. Another internal consistency measure is KR 20 (for details see standard text books on psychometrics such as for example Fishman & Galguera, 2003; Kline, 2000; Murphy and Davidshofer, 2001[6]).

Acceptable tests usually have reliability coefficients between 0.7 and 1.0. Correlations exceeding 0.9 are not often achieved. For individual performance to be measured reliably, correlation coefficients of at least 0.75 and Cronbach's alpha of at least 0.85 are recommended. These numbers represent the minimum values. In the scientific literature, the suggested values are often higher.

Validity

Validity indicates whether a test is able to measure what it is intended to measure. For example, the hit rate alone is not a valid measure of detection performance in terms of discriminability (or sensitivity), because a high hit rate can also be achieved by judging most bags as containing prohibited items. In order to measure detection performance in terms of discriminability (or sensitivity), the false alarm rate must be considered, too.[7]

As with reliability, there are also different types of validity. The term *face validity* refers to whether a test appears to measure what it claims to measure. A test should reflect the relevant operational conditions. For example, if a test for measuring competency in X-ray image interpretation contains a representative sample of X-ray images of bags and screeners have to decide whether the depicted bags contain a prohibited item, it is *face valid. Concurrent validity* refers to whether a test can distinguish between groups that it should be able to distinguish between (e.g., between trained and untrained screeners). In order to establish *convergent validity,* it has to be shown that measures that should be related are indeed related. If, for example, threat image projection (TIP, i.e., the insertion of fictional threat items into X-ray images of passenger bags) measures the same competencies as a computer-based offline test, one would expect a high correlation between TIP performance data and the computer-based test scores. Another validity measure is called predictive validity. In *predictive validity,* the test's ability to predict something it should be able to predict is assessed. For example, a good test for pre-employment assessment would be able to predict on-the-job X-ray screening detection performance. *Content validity* refers to whether the content of a test is representative of the content of the relevant task. For example, a test for assessing whether screeners have acquired the competency to detect different threat items in X-ray images of passenger bags should contain X-ray images of bags with different categories of prohibited items, according to an internationally accepted prohibited items list.

Standardization and development of population norms

The third important aspect of judging the quality of a test is standardization. This involves administering the test to a representative group of people in order to establish norms (a normative group). 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. It is important to know how the normative group was selected, though. For instance, for the standardization of a test used to evaluate the detection performance of screeners, a meaningful normative group of a large and representative sample of screeners (at least 200 males and 200 females) should be tested.

In summary, competency assessment of X-ray image interpretation needs to be based on tests that are reliable, valid, and standardized. However, it is also important to consider test difficulty, particularly if results from different tests are compared with each other. Although two tests can have similar properties in terms of reliability, an easy test may not adequately assess the *level* of competency needed for the X-ray screening job.

Assessment of X-ray Image Interpretation Competency

Currently, there are several methods used to assess X-ray image interpretation competency: covert testing (infiltration testing), threat image projection (TIP), and computer-based image tests.

Covert testing

Covert testing, as the exclusive basis for individual assessment of X-ray image interpretation competency, is only acceptable if the requirements of reliability, validity, and standardization are fulfilled. For covert testing to achieve these requirements, a significant number of tests of the same screener is necessary in order to assess competency reliably. More research is needed to address this issue and it should be noted that this chapter does not apply to principles and requirements for covert testing used to verify compliance with regulatory requirements.

Threat image projection (TIP)

Screener competency can also be assessed using TIP data if certain requirements are met. In cabin baggage screening, TIP is the projection of fictional threat items into X-ray images of passenger bags during the routine baggage screening operation. This way, the detection performance of a screener can be measured under operational conditions. Using *raw* TIP data alone does not provide a reliable measure of individual screener detection performance. Data needs to be *aggregated* over time in order to have a large enough sample upon which to perform meaningful analysis. In order to achieve reliable, valid, and standardized measurements, several other aspects need to be taken into account as well when analyzing TIP data. One requirement is to use an appropriate TIP library. This should contain a large number of threat items, which represent the prohibited items that need to be detected and which feature a reasonable difficulty level. See the section on reliable measurement of performance using TIP for more information on how to use TIP data for measuring X-ray detection performance of screeners.

Computer-based X-ray image interpretation tests

Computer-based X-ray image interpretation tests constitute a valuable tool for standardized measurements of X-ray image interpretation competency. These tests should consist of X-ray images of passenger bags containing different prohibited objects. The categories of items should reflect the prohibited items list and requirements of the appropriate authority, and it should be ensured that the test content remains up to date. The test should also contain clean bag images, that is, X-ray images of bags that do not contain a prohibited object. For each image, the screeners should indicate whether or not a prohibited object is present. Additionally, the screeners can be requested to identify the prohibited item(s). The image display duration should be comparable to operational conditions.

Test conditions should be standardized and comparable for all participants. For example, the brightness and contrast on the monitor should be calibrated and similar for all participants. This applies equally to other monitor settings that could influence detection performance (e.g., the refresh rate). In order to achieve a valid measure of detection performance, not only hit rates but also false alarm rates should be taken into account. An additional or alternative measure would be to count the number of correctly identified prohibited items (in this case, candidates have to indicate where exactly in the bag the threat is located).

The test should be reliable, valid, and standardized. Reliability should be documented by scientifically accepted reliability estimates (see above, section on reliability). If possible, validity measures should also be provided (see above, section on validity). Individual scores should be compared to a norm that is based on a large and representative sample of screeners (see above, section on standardization).

The probability of detecting a prohibited item depends on the knowledge of a screener as well as on the general difficulty of the item. Image-based factors such as the orientation in which a threat item is depicted in the bag (view difficulty), the degree by which other objects are superimposed on an object (superposition), and the number and type of other objects within the bag (bag complexity) influence detection performance substantially[8]. Tests should take these effects into account.

Certification of X-Ray Image Interpretation Competency

As indicated above and as specified in ICAO Annex 17, 3.4.3, individuals carrying out screening operations should be certified initially and periodically thereafter. Certification can not only be considered as providing quality control over the screening process; but also as a valuable basis for awarding personnel a qualification, measuring training effectiveness, improving training procedures, and increasing motivation. Certification data provides important information on strengths and weaknesses in aviation security procedures in general as well as on individual screeners. Furthermore, standardized certification can help in achieving international standardization in aviation security. However, this is very challenging, since many countries, organizations, and companies develop their own certification and quality control systems. The present section gives a brief overview of how a certification system can be implemented.

As mentioned above, certification of screeners should contain a theoretical exam and an X-ray image interpretation exam. For periodic certification, practical exams can be conducted if considered necessary, unlike the initial certification, where practical exams are required. The exams should meet the requirements of high reliability and validity and standardization (see above).

The X-ray image interpretation exam should be adapted to the domain in which a screener is employed, that is, cabin baggage screening, hold baggage screening, or cargo screening. Since not every threat object always constitutes a threat during the flight, depending on where aboard the aircraft it is transported, screeners should be certified according to their domain. The certification of cabin baggage screeners should be based on cabin baggage X-ray images that contain all kinds of objects that are prohibited from being carried on in cabin baggage (guns, knives, improvised explosive devices, and other prohibited items). Objects that are prohibited from being transported in the cabin of an aircraft do not necessarily pose a threat when transported in the hold or in cargo. Furthermore, different types of bags are transported in the cabin, the hold, and cargo. Usually, small suitcases or bags serve as hand baggage, whereas big suitcases and traveling bags are transported in the hold of the aircraft. The certification of hold baggage screeners should be conducted using X-ray images of hold baggage. Similarly, cargo screeners should be tested using X-ray images of cargo items.

Screeners should be kept up to date regarding new and emerging threats. In order to verify whether this is consistently achieved, it is recommended that a recurrent certification should be conducted on a periodical basis, typically every 1-2 years. The minimum threshold that should be achieved in the tests in order to pass certification should be defined by the national appropriate authority and should be based on a large and representative sample of screeners (see also the section on standardization for more information on this topic).

Measurement of Performance on the Job Using Threat Image Projection (TIP)

Threat image projection (TIP) is a function of state-of-the-art X-ray machines that allows the exposure of aviation security screeners to artificial but realistic X-ray images during the process of the routine X-ray screening operation at the security checkpoint. For cabin baggage screening (CBS), fictional threat items (FTIs) are digitally projected in random positions into X-ray images of real passenger bags. In hold baggage screening (HBS), combined threat images (CTIs) are displayed on the monitor. In this case, not only the threat item is projected but an image of a whole bag that may or may not contain a threat item. This is possible if the screeners visually inspecting the hold baggage are physically separated from the passengers and their baggage. If a screener responds correctly by pressing a designated key on the keyboard (the "TIP key") it counts as a hit, which is indicated by a feedback message. If a screener fails to respond to a projected threat within a specified amount of time, a feedback message appears indicating that a projected image was missed. This would count as a miss. Feedback messages also appear if a screener reports a threat although there was no projection of a threat or a CTI. In this case, it could be a real threat. Projecting whole bags in HBS provides not only the opportunity to project threat images (i.e., bags containing a threat item) but also non-threat images (i.e., bags not containing any threat item). This also allows the recording of false true alarms (namely, if a non-threat image was judged as containing a threat) and correct rejections (namely, if a non-threat image was judged as being harmless).

TIP data are an interesting source for various purposes like quality control, risk analysis, and assessment of individual screener performance. Unlike the situation in a test setting, individual screener performance can be assessed on the job when using TIP data. However, if used for the measurement of individual screener X-ray detection performance, international standards of testing have to be met, that is, the method needs to be reliable, valid, and standardized (see above). In a study of CBS and HBS TIP, it was found that there were very low reliability values for CBS TIP data when a small TIP image library of a few hundred FTIs was used[9]. Good reliabilities were found for HBS TIP data when a large TIP image library was available. It is suggested that a large image library (at least 1000 FTIs) containing a representative sample of items of varying difficulty should be used when TIP is used for individual performance assessment. Also viewpoint difficulty, superposition, and bag complexity may need to be considered. Finally, as mentioned above, data needs to be aggregated over time in order to have a large enough sample upon which to perform meaningful analyses. TIP data should only be used for certification of screeners if the reliability of the data has been proven, for example by showing that the correlation between TIP scores based on odd days and even days aggregated over several months is higher than .75.

X-Ray Competency Assessment Test (X-Ray CAT)

This section introduces the X-Ray Competency Assessment Test (X-Ray CAT) as an example of a computer-based test that can be used for assessing X-ray image interpretation competency. The CAT has been developed on the basis of scientific findings regarding threat detection in X-ray images of passenger bags[10]. How well screeners can detect prohibited objects in passenger bags is influenced in two different ways. First, it depends on the screener's knowledge of what objects are prohibited and what they look like in X-ray images. This knowledge is an attribute of the individual screener and can be enhanced by specific training. Second, the probability of detecting a prohibited item in an X-ray image of a passenger bag also depends on image-based factors. These are the orientation of the prohibited item within the bag (view difficulty), the degree by which other objects are superimposed over an object in the bag (superposition), and the number and type of other objects within the bag (bag complexity). Systematic variation or control of the image-based factors is a fundamental property of the test and has to be incorporated in the test development. In the X-Ray CAT, the effects of viewpoint are controlled by using two standardized rotation angles in an 'easy' and a 'difficult' view for each forbidden object. Superposition is controlled in the sense that it is held constant over the two views and as far as possible over all objects. With regard to bag complexity, the bags are chosen in such a way that they are visually comparable in terms of the form and number of objects with which they are packed.

The X-Ray CAT contains two sets of objects in which object pairs are similar in shape. This construction not only allows the measurement of any effect of training, that is, if detection performance can be increased by training, but also possible transfer effects. The threat objects of one set can then be included in the training. By measuring detection performance after training using both sets of the test, it can be ascertained whether training also helped in improving the detection of the objects that did not appear during training. Should this be the case, it indicates a transfer of the knowledge gained about the visual appearance of objects used in training to similar-looking objects.

Materials for the test

Stimuli were created from color X-ray images of prohibited items and passenger bags (Figure 2 displays an example of the stimuli).

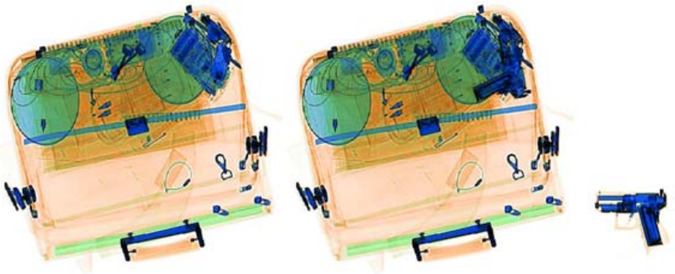


Figure 2 (Prohibited Item Identification.) Example images from the X-Ray CAT. Left: harmless bag (non-threat image), right: same bag with a prohibited item at the top right corner (threat image). The prohibited item (gun) is also shown separately at the bottom right.

On the basis of the categorization of current threat image projection systems (Doc. 30 of the European Civil Aviation Conference, ECAC), four categories of prohibited items were chosen to be included in the test: guns, improvised explosive devices (IEDs), knives, and other prohibited items (e.g., gas, chemicals, grenades, etc.). The prohibited items were selected and prepared in collaboration with airport security experts to be representative and realistic. Sixteen exemplars are used of each category (eight pairs). Each pair consists of two prohibited items of the same kind that are similar in shape. The pairs were divided into two sets, set A and set B. Furthermore, each object within both sets is used in two standardized viewpoints (see Figure 3).

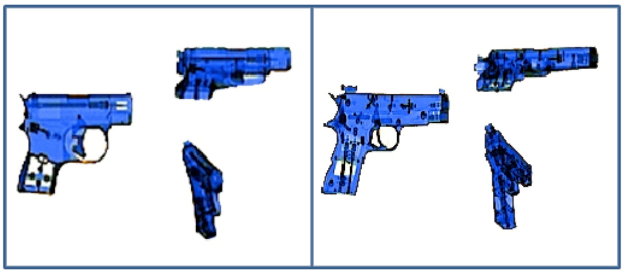


Figure 3 (Prohibited Item Screen Projection.) Example of two X-ray images of similar looking threat objects used in the test. Left: a gun of set A. Right: Corresponding gun of set B. Both objects are depicted also in 85 degree horizontal rotation (top) and 85 degree vertical rotation (bottom).

The easy viewpoint shows the object in canonical (easily recognizable) perspective[11], the difficult viewpoint shows it with an 85 degree horizontal rotation or an 85 degree vertical rotation. In each threat category, half of the prohibited items of the difficult viewpoint are rotated vertically and the other half horizontally. The corresponding object of the other set is rotated around the same axis.

In order to compare how well a prohibited object can be detected relative to its counterpart in the other image set, the two conditions should be comparable in regard to the rotation of the objects, the superposition by other objects, and the bag complexity. Furthermore, the superposition should also be the same for both viewpoints of an object. This was achieved using an image-processing tool to combine the threat objects with passenger bags of comparable image complexity, and at the same time, controlling for superposition. This tool calculates the difference in brightness between the pixels of the two superimposed images (threat object and bag) using the following formula for superposition:

$$SP = \frac{\sqrt{\sum \left[I_{SN}(x,y) - I_{N}(x,y)\right]^{2}}}{ObjectSize}$$

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

This equation calculates the superposition value of an object independent of its size. This value can be held constant for the two views of an object and the two objects of a pair, independently of the bag complexity, when combining the bag image and the prohibited item. To ensure that the bag images do not contain any other prohibited item, they were visually inspected by at least two highly experienced aviation security instructors.

Clean bag images were assigned to the four categories and the two viewpoints of the prohibited items such that their image difficulty was balanced across all groups. This was achieved using the false alarm rate as the difficulty indicator for each bag image based on a pilot study with 192 screeners. In the test, each bag appears twice, once containing a prohibited item (threat image) and once not containing a prohibited item (non-threat image). Combined with all prohibited items this adds up to a total of 256 test trials: 4 threat categories (guns, IEDs, knives, other) * 8 (exemplars) * 2 (sets) * 2 (views) * 2 (threat images v. non-threat images).

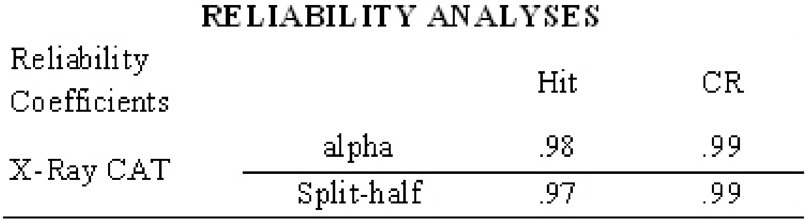
The task is to inspect visually the test images and to judge whether they are OK (contain no prohibited item) or NOT OK (contain a prohibited item). Usually the images disappear after 15 seconds. In addition to the OK / NOT OK response, screeners have to indicate the perceived difficulty of each image in a 100-point scale (difficulty rating: 1 = easy, 100 = difficult). All responses can be made by clicking buttons on the screen. The X-Ray CAT takes about 30–40 minutes to complete.

Assessing Detection Performance in a Computer-Based Test

The detection performance of screeners in a computer based test can be assessed by their judgments of X-ray images. As explained above, not only is the hit rate (i.e., the proportion of correctly detected prohibited items in the threat images) an important value but so is the false alarm rate (i.e., the proportion of non-threat images that were judged as being NOT OK, that is, as containing a prohibited item). This incorporates the definition of detection performance as the ability not only to detect prohibited items but also to discriminate between prohibited items and harmless objects (that is, to recognize harmless objects as harmless). Therefore, in order to evaluate the detection performance of a screener, his or her hit rate in the test has to be considered as well as his or her false alarm rate[12]. There are different measures of detection performance that set the hit rate against the false alarm rate, for example d’ or A’. These measures are explained in more detail below.

Reliability of the X-Ray CAT

As elaborated earlier in this chapter, the reliability of a test stands for its repeatability or consistency. The reliability of the X-Ray CAT was measured by computing Cronbach’s alpha and Guttman’s split-half coefficients. The calculations are based on the results of a study at several airports throughout Europe (see below for the details and further results of the study) including the data on 2265 screeners who completed the X-Ray CAT on behalf of the EU funded VIA project in 2007. The reliability measures were calculated based on correct answers, that is, hits for threat images and correct rejections (CR) for non-threat images (# correct rejections = # non-threat items - # false alarms). The analyses were made separately for threat images and for non-threat images. Table 5.1 shows the reliability coefficients.



RELIABILITY ANALYSES

Reliability Coefficients		Hit	CR
X-Ray CAT	alpha	.98	.99
	Split-half	.97	.99

Table 5.1 Reliabilities

As stated above, an acceptable test should reach split half correlations of at least .75 and Cronbach alpha values of at least .85. Bearing this in mind, the reliability values listed in Table 5.1 show that the X-Ray CAT is very reliable and therefore a useful tool for measuring the detection performance of aviation security screeners.

Validity of the X-Ray CAT

Regarding the different types of validity as described above, the face validity and the content validity can be confirmed instantly. In terms of face validity, the X-Ray CAT is valid as it appears to measure what it claims to measure and it reflects the relevant operational conditions according to aviation security experts. In terms of content validity, the X-Ray CAT is valid as its content is representative of the content of the relevant task. The test includes prohibited items from different categories, on the basis of the definition in Doc. 30 of the European Civil Aviation Conference (ECAC) that have to be detected by the aviation security screeners. Regarding the convergent validity of the CAT, it can be compared to another test that measures the same abilities. An example of such a test that is also widely used at different airports is the Prohibited Items Test (PIT)[13]. To assess convergent validity, the correlation between the scores on the X-Ray CAT and the scores on the PIT of a sample that conducted both tests is calculated. This precise procedure was applied to a sample of 473 airport security screeners. The result can be seen in Figure 4 (r = .791).

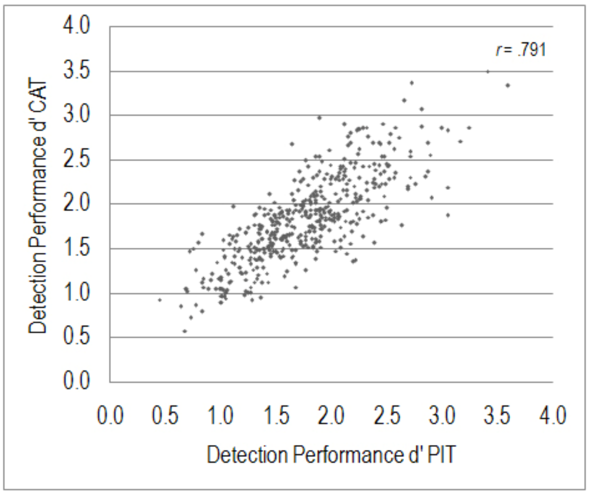


Figure 4 (CAT and PIT Detection Performance.) Convergent validity shown as the reliability between the scores of the X-Ray CAT and the PIT. The dots represent individual screeners.

Since correlation coefficients range from 0 (no correlation) to 1 (perfect correlation) (see also above), the convergent validity can be classified as quite high. This means that the X-Ray CAT and the PIT measure the same X-ray image interpretation competency. Other studies have also confirmed the concurrent validity, that is, the ability of a test to discriminate, for example, between trained and untrained screeners[14]. Figure 5 shows the results of the study. It can be seen that the detection performance increases for the trained screeners but not for the untrained screeners. This means that the test is able to discriminate between screeners who received training with the computer-based training system X-Ray Tutor and those who did not receive training with X-Ray Tutor[15]. Therefore, the concurrent validity of the X-Ray CAT can be confirmed.

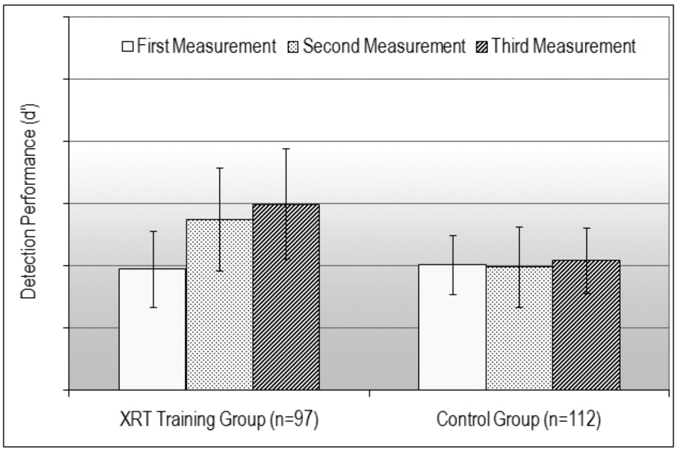


Figure 5 (Detection Performance of Groups.) Detection performance d' for trained (XRT Training Group) compared to untrained (Control Group) screeners. The concurrent validity appears in the difference of the detection performance between the two groups after one group has trained. Thin bars are standard deviations. Note: No performance measures are indicated for security reasons.

Standardization

The X-Ray CAT was standardized in regard to its development. The revisions of the test were based on data from representative samples (N > 94) of airport security screeners (more details on the revisions can be found in the following subsection). In the study described in the section on real world application, involving a large and representative sample of airport security screeners (N = 2265), a mean detection performance A' of 0.8 (SD = 0.08) was achieved. There are different approaches to the definition of pass marks. The normative approach defines a pass mark as the threshold at which a certain proportion of screeners fails the test (e.g., not more than 10 percent), based on a test measurement. That is, a screener is rated in relation to all other screeners. The criterion-referenced approach sets the pass mark according to a defined criterion. For instance, the results could be compared to the test results obtained in other countries when the test was conducted the first time or by having a group of experts (e.g., using the Angoff method[16]) rate the difficulty of the test items (in this case the difficulty of the images) and the minimum standard of performance. These approaches can of course be combined. Furthermore, the standard might be adjusted by taking into account the reliability of the test, the confidence intervals, and the standard error of measurement.

According to the Measurement Research Associates, the level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation and should not be adjusted to regulate the number or proportion of persons passing the test[17]. The pass point should be determined by careful analysis and judgment of acceptable performance. The Angoff method is probably the most basic form of criterion-based standard setting, due to the relatively simple process of determining the pass points[18]. In this method, judges are expected to review each test item and a passing score is computed from an estimate of the probability of a minimally acceptable candidate answering each item correctly. As a first step, the judges discuss and define the characteristics of a minimally acceptable candidate. Then each judge makes an independent assessment of the probability for each item that this previously defined minimally acceptable candidate will answer the item correctly. To determine the probability of a correct response for each item, that is, the passing score, the judges' assessments of the items are averaged. Then these probabilities for all items of the test are averaged to obtain the pass point for the test[19]. The Angoff method features several advantages: it is easy to implement, understand, and compute[20]. However, the

Angoff method also has disadvantages. First, it assumes that the judges have a good understanding of the statistical concepts. Second, the panelists may lose sight of the candidates' overall performance on the assessment due to the focus on individual items, as this method uses an item-based procedure[21]. Moreover, the continuum of item probabilities tends to result in considerable variability among the judges. Many judges have difficulties defining candidates who are minimally competent[22]. In the case of aviation security screeners, judges would have to focus on a person who would be just sufficiently capable of doing the job.

Revision of the test

The development of a scientifically approved test is a complex procedure. Here, the development of the X-Ray CAT is explained in order to provide an example. The first step in a test's development is the definition of what should be measured and how. It was planned that a test should be developed for the purpose of measuring the X-ray image interpretation competency of airport security screeners when they search X-ray images of passenger bags for prohibited objects. In order for the test to be face valid (see above), the nature of the items to be chosen was obvious. They should be X-ray images of passenger bags where some of these images contain a prohibited item and some do not. Careful thought should be invested in the design of the test. In this case, since it is known that several factors can influence the detection performance of an aviation security screener, the items should be constructed considering these factors. That is, the items should be constructed while controlling for the image-based factors view difficulty, superposition, and bag complexity. Furthermore, the effects that should or could be measured with the test should be considered. Depending on the initial point and the aims, the items can be developed quite differently. The X-Ray CAT is composed of two similar sets and contains prohibited items of different categories, each one in two different viewpoints. The set construction serves the purpose of measuring the transfer effects. Transfer effect means the transfer of knowledge about threat objects that is gained during training to threat objects that were not included in training but are similar to objects that were included. The X-Ray CAT can measure several effects: the effect of viewpoint, threat category, training, and transfer (see above for a more detailed description).

After the first version of the test had been constructed, it was administered to a large and representative sample in a pilot study (N = 354 airport security screeners). On the basis of the results of this pilot study, the first revision took place. First of all, a reliability analysis gave information on the quality of the test and each item (item difficulty and item-to-total correlation). Those items with a difficulty below the range of acceptable difficulty had to be revised. The range of acceptable item difficulty depends on the answer type. In this case, an item can be correct or incorrect, that is having a 50 percent chance probability. The range of acceptable difficulty was defined between 0.6 and 0.9. Furthermore, the items should possess as high an item-to-total correlation as possible. In this case, all items with a negative or very small item-to-total correlation were corrected. In order to measure any effect of threat category on the detection performance, the detection performance of a threat object should depend only on the threat object itself and not on the difficulty of the bag it is placed in. To this end, the difficulty of the bags should be balanced across all categories, across both viewpoints of the test, and also across the two sets. As a measure of difficulty for the bag images, the false alarm rate was consulted (i.e., how many times a bag was judged as containing a threat item although there was none). Then, the bags were assigned to the four categories in such a way that their mean difficulty was not statistically different. The threat objects were built into the new bags if necessary, again considering superposition. Lastly, the items were shifted between the two sets (always incorporating the twin structure) in order to equalize the difficulty of the sets. The revised test was administered to another sample (N = 95 airport security screeners), repeating the revision steps as necessary. After a third (N = 359 airport security screeners) and a fourth (N = 222 airport security screeners) revision, the X-Ray CAT was acceptable in terms of stable reliability, item difficulty, and item-to-total correlation.

In summary, the test was revised according to the image difficulty, the item-to-total correlation, and the balancing of the difficulty of the clean bag images. The aim is to achieve a high reliability with items featuring high item-to-total correlations and acceptable item difficulty. The difficulty of a threat image (a bag containing a prohibited object) should depend only on the object itself and not on the difficulty of the bag. Otherwise, a comparison between the detection performance for the different threat categories could be biased.

Real World Application of the X-Ray Competency Assessment Test (X-Ray CAT)

X-Ray CAT was used in several studies and in a series of international airports in order to measure the X-ray image interpretation competencies of screening officers. In this section, the application of X-Ray CAT is presented along with discussions and results obtained by means of the EU-funded VIA Project.

The VIA Project

The VIA Project evolved from the tender call in 2005 of the European Commission's Leonardo da Vinci program on vocational education and training. The project's full title is "Development of a Reference Levels Framework for Aviation Security Screeners." The aim of the project is to develop appropriate competence and qualification assessment tools and to propose a reference levels framework (RLF) for aviation security screeners at national and cross-sectoral levels.

Eleven airports in six European countries were involved in the project. Most of these airports went through the same procedure of recurrent tests and training phases. This made it possible to scientifically investigate the effect of recurrent weekly computer-based training and knowledge transfer and subsequently to develop a reference levels framework based on these outcomes. The tools used for testing in the VIA project were a computer-based training (CBT) program, X-Ray Tutor[23], and the X-Ray CAT. Subsequently, the results of the computer-based test measurements included as part of the VIA project procedure are reported in detail.

VIA Computer-Based Test Measurement Results

As explained earlier, the X-Ray Competency Assessment Test (CAT) contains 256 X-ray images of passenger bags, half of which contain a prohibited item. This leads to four possible outcomes for a trial: a "hit" (a correctly identified threat object), a "miss" (a missed threat object), a "correct rejection" (a harmless bag correctly judged as being OK), and a "false alarm" (an incorrectly reported threat object).

In terms of sensitivity, the hit rate alone is not a valid measure to assess X-ray image interpretation competency. It is easy to imagine that a hit rate of 100 percent can be achieved by simply judging every X-ray image as containing a prohibited item. In this case, the entire set of non-threat items is completely neglected by this measure (the false alarm rate would also be 100 percent). In contrast, Green and Swets in 1966 developed a signal detection performance measure d’ (say, d prime), taking into account hit rates as well as false alarm rates.[24] Often, d’ is referred to as sensitivity, emphasizing the fact that it measures the ability to distinguish between noise (in our case an X-ray image of a bag without a threat) and signal plus noise (in our case an X-ray image containing a prohibited item).

d’ is calculated using the formula $d' = z(H) - z(F)$, where H is the hit rate, F the false alarm rate, and z the z-transformation. For the application of d’, the data have to fulfill certain criteria (noise and signal plus noise must be normally distributed and have the same variance). Another widely used measure is the “nonparametric” value A’ (say, A prime). The term “nonparametric” refers to the fact that the computation of A’ requires no a priori assumption about underlying distributions. The measure also meets the requirement of setting the hit rate against the false alarm rate in order to achieve a reliable and valid measure of image interpretation competency. A’ was the measure of choice for the current analyses because its non-parametric character allows its use independently from the underlying measurements distributions. A’ can be calculated as follows, where H represents the hit rate of a test candidate or group and F represents its false alarm rate: $A' = 0.5+ [(H - F)(1 + H - F)] / [4H(1 - F)$. If the false alarm rate is larger than the hit rate, the equation must be modified[25]: $A' = 0.5 - [(F - H)(1 + F - H)] / [4F(1 - H)]$.[26]

The reported results provide graphical displays of the relative detection performance measures A’ at the eight European airports that participated in the VIA project, as well as another graph showing the effect of the two viewpoints on the different threat categories as explained earlier. In order to provide statistical corroboration of these results, an analysis of variance (ANOVA) on the two within-participants factors, view difficulty and threat category (guns, IEDs, knives and other items), and the between-participants airport factor is reported as well. As part of the ANOVA, only the significant interaction effects are reported and considered to be noteworthy in the context.

Detection performance comparison between airports

Figure 6 shows the comparison of the detection performance achieved at eight European airports that participated in the VIA project. First, the detection performance was calculated for each screener individually. The data were averaged across screeners for each airport. Thin bars represent the standard deviation (a measure of variability) across screeners. Due to its security sensitivity and for data protection reasons, the individual airports' names are not indicated and no numerical data are given here.

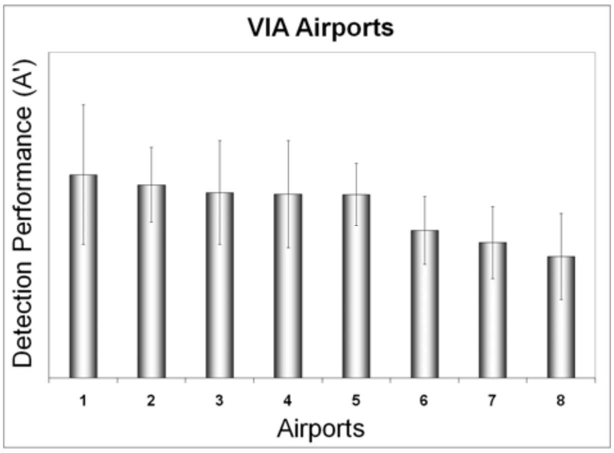


Figure 6 (Detection Performance of Airports.) Comparison between eight European airports participating in the VIA project. Thin bars represent standard deviations between screeners.

Although no numerical data is displayed in the graph, we can discern substantial differences between the airports in terms of mean detection performance and standard deviation. As described above, all VIA airports go through a similar procedure of alternation of test phases and training phases. Nevertheless, there are considerable differences between them. There were large differences in the initial positions when the project was started, and the baseline assessment test, which is reported here, was conducted at different times at different airports. The differences can be put down to differences in the amount of training that was accomplished prior to this baseline test as well as to differences in the personnel selection assessment. Some of the reported airports were already coached prior to the VIA project, though with diverse intensity and duration. Taking these differences into account, the reported results correspond fairly well with our expectations based on earlier studies of training effects.

Detection performance comparison between threat categories regarding view difficulty

Figure 7 shows again the detection performance measure A', but with a different focus. The data are averaged across the airports shown in Figure 6, but analyzed by view difficulty within threat categories. There is a striking effect on detection performance deriving from view difficulty. Performance is significantly higher for threat objects depicted in easy views than for threat objects depicted in difficult views (canonical views rotated by 85 degrees).

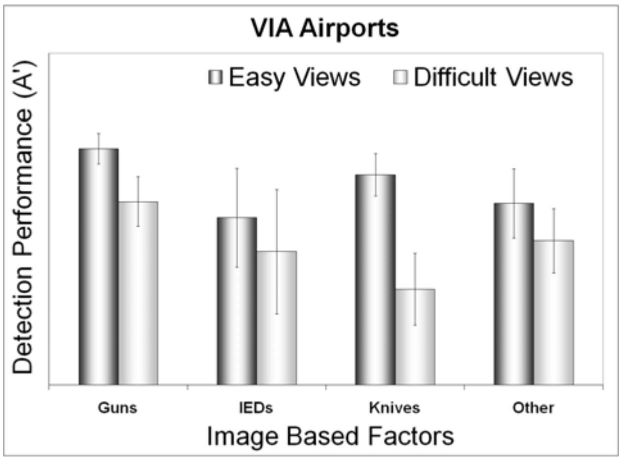


Figure 7 (Airport Detection of Prohibited Items.) Detection performance A' broken up by category and views (unrotated (easy view) vs. 85° rotated objects (difficult view)). The thin bars represent standard deviations between the eight VIA airports. Pairwise comparisons showed significant viewpoint effects for all four threat categories.

Although this effect can be found in every one of the four threat categories, there are significant differences between them regarding general differences between the mean detection performances and also between the effect sizes of view difficulty that are unequal between threat categories. Knives and IEDs, for example, differ very much in view difficulty effect size but not so much in average detection performance. As can be seen in Figure 8, the reason is quite simple: IEDs consist of several parts and not all parts are depicted in easy or in difficult view at the same time. Some parts are always depicted in easy view when others are difficult, and vice versa. Knives have very characteristic shapes. They look consistently longish when seen perpendicular to their cutting edge but very small and thin when seen in parallel to their cutting edge. This interaction effect between threat item category and view difficulty can easily be observed in Figure 7, where the difference between easy and difficult views is much larger in knives than in IEDs. Furthermore, based on earlier studies of training effects, it is important to mention here that this pattern shown in Figure 7 is also highly dependent on training[27] (interaction effects [category * airport and view difficulty * airport]).

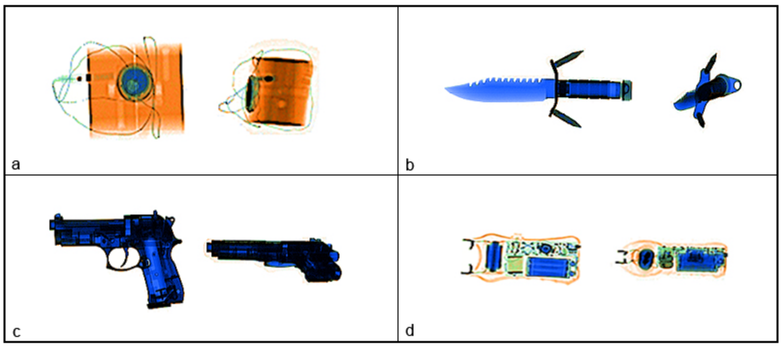


Figure 8 (Views of Prohibited Items.) Illustration of how effects of view difficulty differ between the four ECAC threat categories. 8a and b show an IED and a knife each in a frontal view and a rotated view from almost 90 degrees around the vertical axis. 8c and d show a gun and a taser each in a frontal view and a rotated view from almost 90 degrees around the horizontal axis.

Analysis of Variance (ANOVA)

The following statistics provide quantitative values for what has been reported graphically. This allows us to compare the effects of the different factors. We applied a three-way ANOVA to the two within-subjects factors, category and view difficulty, and one between-subjects airport factor on the detection performance measure A'.

The analysis revealed highly significant main effects on threat category (guns, IEDs, knives, and other items) with an effect size of $\eta^2 = .131$, $F(3, 5602.584) = 339.834$, $MSE = 2.057$, $p < .001$, on view difficulty (easy view v. difficult/rotated view) with an effect size of $\eta^2 = .47$, $F(1, 2257) = 2009.772$, $MSE = 9.031$, $p < .001$, and also on the between-subjects airport factor with an $\eta^2 = .080$, $F(1, 2257) = 28.128$, $MSE = 1.415$, $p < .001$. The following two-way interactions were also highly significant: threat category * view difficulty: $\eta^2 = .094$, $F(3, 6542.213) = 233.969$, $MSE = .931$, $p < .001$, threat category * airport $\eta^2 = .068$, $F(3, 5602.584) = 23.411$, $MSE = .142$, $p < .001$, and view difficulty * airport $\eta^2 = .159$, $F(1, 2257) = 60.953$, $MSE = .274$, $p < .001$. These results indicate different detection performance for different threat categories and higher detection performance for prohibited items in easy view than for rotated threat items (the effect of viewpoint). This is consistent with results reported in the view-based object recognition literature (for reviews see, for example, two works by Tarr and Bülthoff.[28] The effect sizes were very large according to Cohen's conventions.[29]

Discussion

Although the reported real world application consists of baseline measurement data only, some important features of the X-Ray CAT could be illustrated well. X-Ray CAT allows us to measure and to evaluate the effects of view difficulty and threat objects practically independently of each other. Furthermore, the X-Ray CAT can be used as a very reliable tool to compare the X-ray image interpretation competency of security staff at different airports and other types of infrastructure using X-ray technology for security control procedures.

Summary and Conclusions

The competency of a screener to detect prohibited items in X-ray images quickly and reliably is important for any airport security system. Computer-based tests, TIP, and to a limited extent covert tests can be used to assess individual competency in X-ray image interpretation. However, to achieve reliable, valid, and standardized measurements, it is essential that the requirements and principles detailed in this chapter are followed by those who produce, procure, or evaluate the competency assessment of the X-ray image interpretation tests of individual screeners.

This chapter introduced the competency assessment in airport security screening. In order to achieve a meaningful result the assessment has to meet the criteria of reliability and validity. Furthermore, the assessment has to be standardized to allow the evaluation of screeners' performance in relation to the population norm. Currently, there are three means for assessing X-ray image interpretation competency: covert testing, threat image projection (TIP), and computer-based image testing. Another important feature of maintaining the high level of X-ray baggage screening within aviation security is the initial and recurrent certification of screening personnel.

Threat image projection (TIP) as a means to assess X-ray image interpretation competency was discussed as well as the conditions that have to be fulfilled in order for TIP to be a reliable and valid instrument.

This chapter also focused on the computer-based X-Ray Competency Assessment Test (X-Ray CAT). It features very high reliability scores and its design allows us to measure the X-ray image interpretation competency of aviation security screeners with regard to different aspects of their ability and knowledge. The X-Ray CAT is widely used at over 100 airports throughout the world, for competency assessment and certification purposes as well as in studies assessing the fundamentals of the demands required for the job of the aviation security screener.

This chapter continued by showing how a reliable, valid, and standardized test can be used to compare X-ray image interpretation competency across different airports and countries. The results of an EU-funded project (the VIA Project) showed remarkable differences in mean detection performance across eight European airports. All these countries currently conduct weekly recurrent computer-based training. Since the X-Ray CAT will be conducted again in the first quarter of 2008, the VIA Project will also provide important insights on the benefits of computer-based training for increasing security and efficiency in X-ray screening.

Acknowledgment

This research was financially supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403). This chapter is a summary of Work Package 6: Competency assessment tests. For more information, see www.viaproject.eu. The first part of this document is based on a VICOREG draft which was the basis for the *White Paper on Principles and Requirements for Assessing X-Ray Image Interpretation Competency of Aviation Security Screeners*. We are thankful to all contributors for valuable comments regarding the White Paper.

References

1. ICAO Annex 17, 3.4.3 ("Each Contracting State shall ensure that the persons carrying out screening operations are certified according to the requirements of the national civil aviation security programme").

2. ICAO Manual on Human Factors in Civil Aviation Security Operations, Doc. 9808.

3. ICAO Human Factors Training Manual, Doc. 9683, part 1, chapter 4, and in Appendix 6—"Guidance on Recruitment, Selection, Training, and Certification of Aviation Security Staff"—and Appendix 32—"Guidance on the Use of Threat Image Projection."

4. ICAO Security Manual for Safeguarding Civil Aviation against Acts of Unlawful Interference, Doc. 8973, chapter 4, I-4–45 ("Recruitment, Selection, Training, and Certification of Security Staff").

5. ECAC Doc. 30, Annex IV-12A, "Certification Criteria for Screeners," and ECAC Doc. 30, chapter 12, 12.2.3, "Certification of Security Staff," 1.1.10.3.

6. Joshua A. Fishman and Tomas Galguera, *Introduction to Test Construction in the Social and Behavioural Sciences. A Practical Guide* (Oxford: Rowman & Littlefield, 2003); Paul Kline, *Handbook of Psychological Testing* (London: Routledge, 2000); Kevin R. Murphy and Charles O. Davidshofer, *Psychological Testing* (Upper Saddle River, NJ: Prentice Hall, 2001).

7. Neil A. MacMillan and C. Douglas Creelman, *Detection Theory: A User's Guide* (New York: Cambridge University Press, 1991); Franziska Hofer and Adrian Schwaninger, "Reliable and Valid Measures of Threat Detection Performance in X-ray Screening," *IEEE ICCST Proceedings* 38 (2004): 303–8.

8. Adrian Schwaninger, Diana Hardmeier, and Franziska Hofer, "Measuring Visual Abilities and Visual Knowledge of Aviation Security Screeners," *IEEE ICCST Proceedings* 38 (2004): 258–64; Adrian Schwaninger, "Evaluation and Selection of Airport Security Screeners," *AIRPORT* 2 (2003): 14–15.

9. Franziska Hofer and Adrian Schwaninger, “Using Threat Image Projection Data for Assessing Individual Screener Performance,” *WIT Transactions on the Built Environment* 82 (2005): 417–26.

10. Schwaninger, Hardmeier, and Hofer, “Measuring Visual Abilities and Visual Knowledge of Aviation Security Screeners”; Schwaninger, “Evaluation and Selection of Airport Security Screeners.”

11. Stephen E. Palmer, Eleanor Rosch, and Paul Chase, “Canonical Perspective and the Perception of Objects,” in *Attention and Performance IX*, ed. John Long and Alan Baddeley, 135–52 (Hillsdale, NJ: Erlbaum, 1981).

12. David M. Green and John A. Swets, *Signal Detection Theory and Psychophysics* (New York: Wiley, 1966); Neil A. MacMillan and C. Douglas Creelman, *Detection Theory: A User’s Guide* (New York: Cambridge University Press, 1991); Hofer and Schwaninger, “Reliable and Valid Measures of Threat Detection Performance in X-Ray Screening”; Hofer and Schwaninger, “Using Threat Image Projection Data for Assessing Individual Screener Performance.”

13. Diana Hardmeier, Franziska Hofer, and Adrian Schwaninger, “Increased Detection Performance in Airport Security Screening Using the X-Ray ORT as Pre-employment Assessment Tool,” *Proceedings of the 2*nd *International Conference on Research in Air Transportation*, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24–28, (Belgrade, Serbia: ICRAT, 2006), 393–97.

14. Saskia M. Koller et al., “Investigating Training, Transfer and Viewpoint Effects Resulting from Recurrent CBT of X-Ray Image Interpretation,” *Journal of Transportation Security* 1, no. 2(2008).

15. Adrian Schwaninger, “Computer-Based Training: A Powerful Tool for the Enhancement of Human Factors,” *Aviation Security International* 10(2004): 31–36; Adrian Schwaninger, “Increasing Efficiency in Airport Security Screening,” *WIT Transactions on the Built Environment* 82 (2005): 405–16.

16. William Herbert Angoff, “Norms, Scales, and Equivalent Scores,” in *Educational Measurement* (2nd ed.), ed. Robert L. Thorndike, 508–600 (Washington: American Council on Education, 1971).

17. Measurement Research Associates, *Criterion Referenced Performance Standard Setting*, 2004, http://www.measurementresearch.com/wwa/default.shtml.

18. Muhammad Naveed Khalid and Muhammad Saeed, “Criterion Referenced Setting Performance Standards with an Emphasis on Angoff Method,” *Journal of Research and Reflections in Education* 1 (2007): 66–87.

19. Measurement Research Associates, “Criterion Referenced Performance Standard.”

20. Ronald A. Berk, “A Consumer’s Guide to Setting Performance Standards on Criterion-Referenced Tests,” *Review of Educational Research* 56 (1986): 137–72.

21. Khalid and Saeed, “Criterion Referenced Setting Performance Standards with an Emphasis on Angoff Method.”

22. Berk, “A Consumer’s Guide to Setting Performance Standards on Criterion-Referenced Tests.”

23. Schwaninger, A. (2005b). Increasing Efficiency in Airport Security Screening. *WIT Transactions on the Built Environment, 82*, 407-416.

24. David Green and John Swets, “Signal Detection Theory and Psychophysics,” in *Detection Theory: A User’s Guide,* ed. Neil MacMillan (London: Erlbaum, 1966).

25. Doris Aaronson and Brian Watts, “Extensions of Grier’s Computational Formulas for A’ and B’’ to Below-Chance Performance,” *Psychological Bulletin* 102 (1987): 439–42.

26. Harold Stanislaw and Natasha Todorov, “Calculation of Signal Detection Theory Measures,” *Behavior Research Methods, Instruments, and Computers* 31, no. 1 (1999): 137–49; Green and Swets, *Signal Detection Theory and Psychophysics;* Irwin Pollack and Donald A. Norman, “A non-parametric Analysis of Recognition Experiments,” *Psychonomic Science* 1 (1964): 125–26; J. Brown Grier, “Nonparametric Indexes for Sensitivity and Bias: Computing Formulas,” *Psychological Bulletin* 75 (1971): 424–29; ICAO Security Manual for Safeguarding Civil Aviation against Acts of Unlawful

Interference, Doc. 8973, chapter 4, I-4–45 (“Recruitment, Selection, Training and Certification of Security Staff”).

27. Koller, S.M., Hardmeier, D., Michel, S., & Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent CBT of x-ray image interpretation. *Journal of Transportation Security 1(2)*, 81-106.

28. Michael J. Tarr and Heinrich H. Bülthoff, “Is Human Object Recognition Better Described by Geon Structural Descriptions or by Multiple Views? Comment on Biederman and Gerhardstein (1993),” *Journal of Experimental Psychology: Human Perception and Performance* 21 (1995): 1494–1505; Michael J. Tarr and Heinrich H. Bülthoff, “Image-Based Object Recognition in Man, Monkey and Machine,” in *Object Recognition in Man, Monkey and Machine,* ed. Michael J. Tarr and Heinrich H. Bülthoff, 1–20 (Cambridge, MA: MIT Press, 1998).

29. Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (New York: Erlbaum, Hillsdale, 1988).