Using threat image projection data for assessing individual screener performance

F. Hofer & A. Schwaninger
Department of Psychology, University of Zurich, Switzerland

Abstract

Threat image projection (TIP) is a technology of current x-ray machines that allows exposing screeners to artificial but realistic x-ray images during the routine baggage x-ray screening operation. If a screener does not detect a TIP within a specified amount of time, a feedback message appears indicating that a projected image was missed. Feedback messages are also shown when a TIP image is detected or in the case of a non-TIP alarm, i.e. when the screener indicated that there was a threat but in fact no TIP was shown. TIP data is an interesting source for quality control, risk analysis and the assessment of individual screener performance. In two studies we examined the conditions for using TIP data for the latter purpose. Our results strongly suggest using aggregated data in order to have a large enough data sample as the basis for statistical analysis. Second, an appropriate TIP library containing a large number of threat items, which are representative for the prohibited items to be detected, is recommended. Furthermore, consideration should be given to image-based factors such as general threat item difficulty, viewpoint difficulty, superposition and bag complexity. Different methods to cope with these issues are discussed in order to achieve reliable, valid and standardized measurements of individual screener performance using TIP.

Keywords: airport security, human factors, threat image projection, detection performance, reliability analysis, hit rate, false alarm rate.

1 Introduction

The task of an airport security screener is to visually inspect passenger bags for forbidden or dangerous objects. In order to perform this task effectively, a screener needs to know which items are prohibited and what they look like in x-ray images of passenger bags. As pointed out by Schwaninger [1], some threat objects look very different in an x-ray image than in reality. Other prohibited items are difficult to identify in an x-ray image because they look similar to harmless objects. Because of these and other reasons, training and visual experience are essential in order to achieve and maintain a high level of detection performance (Schwaninger [2]; Schwaninger [3]; Schwaninger and Hofer [4]).

In addition to such knowledge-based factors of expertise and training, there are several image-based factors, which also influence detection performance (Schwaninger [5]; Schwaninger et al. [6]). When prohibited items are rotated they can become more difficult to recognize (effect of viewpoint). Superimposition by other objects in the bag can also affect detection performance (effect of superposition). In addition, the number and type of other objects can affect the visual search for prohibited items (effect of bag complexity). Interestingly, comparable effects of these image-based factors on detection performance have been found for novices as well as for experts. Moreover, large inter-individual differences were found in the ability to cope with these image-based factors, which accounted for novices as well as for experts. Thus, these image-based factors seem to be rather related to relatively stable visual abilities than to training and visual experience.

The fact that image- and knowledge-based factors strongly influence detection performance points out that the effectiveness of aviation security technology is limited by the abilities and expertise of the humans that operate it. Therefore, reliable and valid procedures for assessing individual detection performance of screeners are relevant for quality control, risk analysis and screener certification purposes.

Large technological progress has been made in aviation security in the last two decades. One relatively new technology is threat image projection (TIP). This is a software function of state-of-the art x-ray machines that allows measuring of detection performance on the job. In TIP, virtual threat images are projected randomly on x-ray screening systems. For cabin baggage screening (CBS), fictional threat items (FTIs) are projected into x-ray images of real passenger bags in a random position. In hold baggage screening (HBS), combined threat images (CTIs) are displayed, i.e. virtual x-ray images of whole bags that can contain threat items. The use of CTIs is not possible in cabin baggage screening, because x-ray operators see the passengers and their luggage. Since in many hold baggage screening systems the operators are isolated from the passengers, it is possible to use CTIs in HBS TIP.

If a screener detects the projected threat item within a predefined time, the answer counts as hit. Missing a TIP-image counts as miss. Non-Tip alarms are registered in CBS if a screener gives a threat present response when no TIP image was shown. In some HBS systems not only threat x-ray images but also non-threat x-ray images of passenger bags are shown. In this case, false alarms as well as correctly judging bags to be harmless are also written into TIP report files. Feedback messages are always presented to the screener when a TIP-image has been shown or in the case of a non-TIP alarm (CBS) or a false alarm (HBS).

TIP data is an interesting source for quality control, risk analysis and assessment of individual screener performance. Especially for the latter purpose, reliability of measurement is of special importance. This was examined in two studies using CBS and HBS data, respectively.

2 CBS Study

2.1 Method

2.1.1 TIP Library

A standard TIP library based on FAA [7] was used, which is available on current TIP systems. A TIP: bag ratio of 1:50 was used in this study.

2.1.2 Participants

333 CBS airport security screeners took part in this study. They were all familiarized with the same TIP library using generic logins for several weeks before the study was started using individual logins. The study was conducted over a period of 7 months.

2.1.3 Analyses

There are different ways to estimate reliability. The most common procedures are test-retest, split-half, alternate forms and internal consistency analyses (for an overview see for example Kline [8]; Murphy and Davidshofer [9]). Because in both CBS and HBS current TIP software selects the threat items on a purely random basis, there can be quite substantial differences in repeated exposure to different items between screeners. It is therefore not possible to run the same TIP projections for every screener. This complicates reliability analyses because it implies that neither of the common reliability procedures can be applied in its pure form.

For the purpose of this study, two ways of data splitting were conducted: First, the hit rate of even days was correlated with the hit rate of odd days. Aggregated data was used, which was collected over a period of seven months. Second, the hit rate from one, two and three successive months was correlated with the hit rate of the following one, two and three months, respectively. For both ways of data splitting, some items can be in both halves, whereas some other items are only found in one of the two halves (varying across participants). Therefore, the reliability analyses in this study are a combination of split-half and test-retest reliability.

Psychophysical measures such as d’ or A’ are more valid estimates of detection performance than the hit rate alone (Hofer and Schwaninger [10]; MacMillan and Creelman [11]; Green & Swets [12]). These measures take the hit *and* false alarm rate into account. In this study, all reliability analyses were done only with the hit rate due to the following reasons: First, it is not possible to get a valid false alarm rate from CBS TIP reports because the individual non-TIP alarm rate does not completely match the individual false alarm rate. If a screener detects a real threat in a bag when no TIP image is present, this is recorded as a non-TIP alarm in the TIP report. In this case the response should count as a (true) hit and certainly not as a false alarm. Second, because correctly judging a bag to be harmless is not written into the CBS TIP report, the individual non-TIP alarm rate has to be estimated based on the averaged TIP to bag ratio, which can further reduce the internal validity of the estimates. Therefore, only the hit rate was analyzed as a measure of performance. Non-TIP alarm rates (CBS) and false alarm rates (HBS) are reported here to illustrate differences between individuals in their response tendencies.

2.2 Results

2.2.1 Correlations between hit rates of even and odd days

Figure 1 shows the correlations between the hit rates of even and odd days, aggregated over different numbers of months and averaged between the categories guns, knives and IEDs (separate results for each category (guns, knives and IEDs) are very similar to the overall result and therefore not reported here).

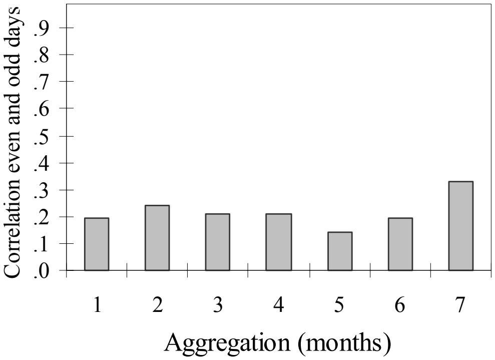

Figure 1: Correlations between even and odd days, aggregated over different numbers of months.

As can be seen in Figure 1, the correlations between the hit rate of even and odd days was relatively small and clearly below .40, even if TIP data from seven months were used. The mean even-odd day correlation for data of one month was r = .19, over seven months r = .33. It is important to note that the hit rate was very high and also very stable for all numbers of months (not shown in Figure 1). Moreover, the standard deviations were very small. These results suggest that most screeners achieved ceiling performance already in the first weeks when generic logins were used prior to the data collection for this study (see also section 2.1.2).

2.2.2 Correlations between the hit rate of consecutive months

Figure 2 illustrates the correlations for data of two, four and six successive months, calculated by correlating the hit rate between the first and second month, the hit rate between the first two and second two months, and the hit rate between the first three and second three months.

Correlation varied between $r = .32$ and $r = .58$. Thus, splitting the CBS data between different months resulted in higher correlations than splitting the data into even and odd days.

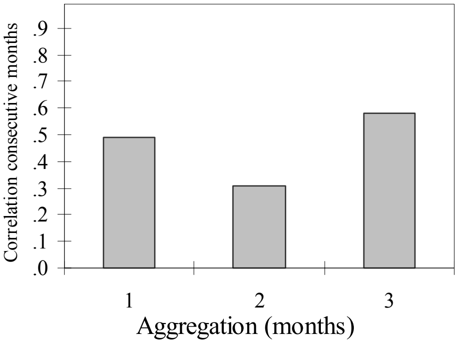

Figure 2: Correlations between the hit rate of the first and second month (1), of the first two and second two months (2), and of the first three and second three months (3).

2.2.3 Individual non-TIP alarm rates

As can be seen in Figure 3, non-TIP alarms varied substantially between individual screeners. We found 0% for the screener with the lowest and 19% for the screener with the highest non-TIP alarm rate.

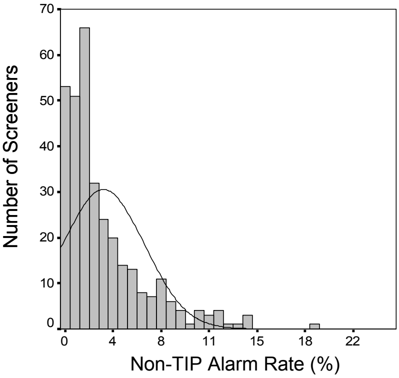

Figure 3: Distribution of individual non-TIP alarm rates (averaged over the 7 months period for each screener).

2.3 Discussion

The correlations for the CBS TIP data of the standard TIP library available on conventional x-ray screening systems were clearly below .4 when splitting the data into even and odd days. When splitting the data between consecutive months, the correlations were higher, but still too small to conclude that this data is reliable enough for individual performance assessment (all r < = .58). Note however, that the hit rate was very high and stable over different months. In addition, very small standard deviations were observed. Thus, a ceiling effect in the data and small inter-individual differences could have resulted in small reliability coefficients.

The non-TIP alarm rate varied substantially between individual screeners, which reflects differences in response bias. Since the hit rate is dependent on individual response biases, its validity for measuring detection performance in terms of sensitivity is reduced.

3 HBS study

3.1 Method

3.1.1 TIP library

The library used for HBS TIP consisted of 1028 combined threat images (CTIs). 64 improvised explosive devices (IEDs) were selected by police experts from a large x-ray image database in order to create a representative sample of different IED types. Each IED was combined with 8 bags of different difficulties rated by 8 x-ray screening experts. Each bag was also displayed without the IED. Thus, the whole HBS TIP library consisted of 64 IEDs * 8 difficulty levels * 2 (harmless vs. dangerous bags) = 1028 CTIs. A TIP: bag ratio of 1:30 was used in this study.

3.1.2 Participants

74 HBS airport security screeners participated in this study. They were all familiarized with TIP using individual logins several weeks before the individual measurement started. The TIP images used in the introductory phase were different from the ones used for the reliability analyses. The study was conducted over a period of 16 months.

3.1.3 Analyses

The same method as for CBS was used to assess the reliability of HBS TIP data. Again, correlations between the hit rate of even and odd days for different numbers of months were calculated and correlations between the hit rates of several successive months were computed.

3.2 Results

3.2.1 Correlations between data of even and odd days

Figure 4 shows the mean correlations between even and odd days aggregated over one month (r = .70) up to 16 months (r = .94). The hit rates (not shown in Figure 4) were smaller than in study 1. Moreover, much larger standard deviations were observed now.

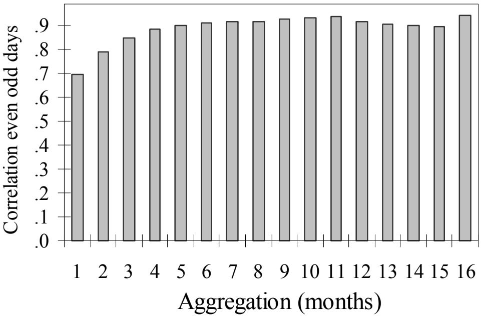

Figure 4: Correlations between even and odd days, aggregated over different numbers of months.

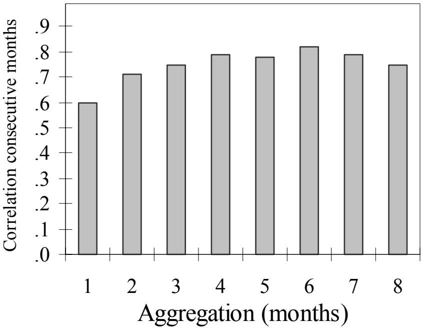

Figure 5: Correlations between the hit rates aggregated over different numbers of consecutive months.

3.2.2 Correlations between the hit rate of consecutive months

Figure 5 shows the reliability coefficients calculated by correlating the hit rate between the first and second month, the first two and the following two months, the first three and the following three months etc.

The correlation between the hit rate of the first two months was $r = .60$, the correlation between the first eight and the following eight months was $r = .75$.

3.2.3 Individual false alarm rates

As explained in the method section, in this study half of all TIPs were harmless bags whereas the other CTIs contained an IED. Figure 6 shows the distribution of false alarm rates based on TIP trials containing a harmless bag (averaged for each screener over the 16 months period).

3.3 Discussion

Compared to the CBS TIP data, much higher correlations were revealed for the HBS TIP library used in this study. The correlation between even and odd days was .70 for data aggregated over one month and > .8 when data were aggregated over three months. For data aggregated over 6 or more months, the correlation between hit rates of even and odd days was > .9. When correlating the hit rate between different numbers of consecutive months, correlations were still quite high, although less high than when splitting the data into even and odd days. Again, as for the CBS data, screeners vary substantially in their response bias, which was reflected by the variation in the false alarm rate of the HBS TIP data. Since the hit rate is affected by response bias, this result questions the validity of the hit rate for measuring detection performance in terms of sensitivity.

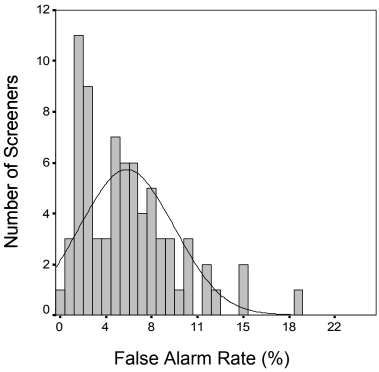

Figure 6: Distribution of individual false alarm rates from non-threat TIP trials (averaged across all 16 months for each participant).

4 General discussion

Threat image projection is a technology of current x-ray machines that allows exposing screeners to artificial but realistic x-ray images during the routine baggage x-ray screening operation. Because TIP allows realistic on the job measurement, it could be a useful tool for assessing individual screener x-ray detection performance. To this end, the measurement has to fulfil international standards of testing, i.e. the method used needs to be reliable, valid, objective and standardized. In this study, we analyzed TIP data from CBS and HBS in order to investigate reliability. Current TIP software allows only random projection of images. Since common reliability procedures need a standardized and controlled item set, they cannot be applied in their pure form. Therefore, we used a mixture between split-half and test-retest methods to estimate TIP data reliability. Data splitting was done in two different ways: The hit rate of even and odd days was correlated, while aggregating data over different numbers of months. Second, reliability was estimated by computing the correlation between the hit rate of consecutive months, e.g. the correlation between the first and second, the first two and second two, the first three and second three months, etc.

In the first study, the standard CBS TIP library was used, which is available on conventional x-ray screening systems. We found very low reliability values (all r <=.58), even for data aggregated over seven months. This is true for both data splitting methods used in this study. Although in general, splitting the data into different numbers of successive months resulted in slightly larger correlation coefficients as when splitting data into even and odd days. It is important to note that the hit rate was very high and only a small inter-individual variance was observed.

In the second study, a more difficult image library was used in HBS. The correlation between even and odd days is already .70 for data aggregated over one month. For data aggregated over 6 or more months, the correlation is > .9. When correlating the hit rate between different numbers of consecutive months, correlations are still quite high, although lower than when correlating even and odd days. Compared to the CBS data of study 1, the hit rate of the HBS TIP data was not at ceiling, and larger standard deviations could be observed.

What reasons could account for the fact that reliability of CBS data was so low while reliability coefficients were much higher for HBS data? One reason for the low reliability of CBS data could be a ceiling effect and the small inter-individual differences. When using TIP for individual performance assessment a large image library containing a representative sample of items of varying difficulty should be used. At least from a testing psychology standpoint it could also be considered to eliminate the most easy and most difficult items. Another reason for the higher reliability of HBS TIP data could be related to the differences in the TIP system. In HBS, combined threat images are used, i.e. the threat item is shown always with the same bag, embedded at the same position. Therefore, any effects of superposition and bag complexity are kept constant. In CBS, only the threat items are projected into x-ray images of real passenger bags. This induces much additional variance of image difficulty because luggage varies in terms of bag complexity. Moreover, depending on the randomly selected location of the FTI, large differences in superposition are found. Therefore, it remains to be seen, whether reliable data can be obtained using CBS TIP. We are currently investigating this in a study using a larger CBS TIP library that contains more difficult threat items.

This study also showed that there are substantial inter-individual differences in non-TIP alarm rates (CBS) and false alarm rates (HBS). This indicates differences in response bias, which also affects the validity of hit rates as a measure of detection. It certainly would be desirable to design TIP systems in which a valid measure of false alarm rate can be obtained so that more valid detection measures such as d’ and A’ can be derived from hits and false alarms (Hofer and Schwaninger [10]; MacMillan and Creelman [11]; Green and Swets [12]). This is already possible today in HBS TIP because CTIs are used and harmless bags can be projected in order to obtain valid false alarm estimates. In CBS this is not possible because FTIs are projected into x-ray images of real passenger bags. When a screener detects a real threat item, a non-TIP alarm is recorded, even though in this case it is a (true) hit and certainly not a false alarm. By separately recording these cases more valid hit and false alarm rates could be calculated. The true hits would simply be added to the hits obtained in TIP, the false alarm rate would equal the non-TIP alarms minus the true hits. Based on corrected hit and false alarm rates it would be possible to calculate d' or A' scores, which are more valid detection measures than the hit rate alone.

Acknowledgements

We are thankful to airport and police authorities that provided TIP data and funding for conducting this study. We also thank Judith Riegelnig for her help in analyzing the data.

References

[1] Schwaninger, A., Increasing effectiveness and efficiency in airport security screening, *WIT Transactions on the Built Environment*, this volume.
[2] Schwaninger, A., Training of airport security screeners. *AIRPORT, 05,* pp. 11-13, 2003.
[3] Schwaninger, A., Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International, FEB/2004,* pp. 31-36, 2004.
[4] Schwaninger, A. & Hofer, F., 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, 2004.
[5] Schwaninger, A., Evaluation and selection of airport security screeners. *AIRPORT*, 02, pp. 14-15, 2003.
[6] Schwaninger, A., Hardmeier, D., & Hofer, F., Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* pp. 258-264, 2004.
[7] Federal Aviation Administration, Functional requirements for threat image projection systems on X-ray machines, *DOT/FAA/AR-97/67*, August 1997.
[8] Kline, P., *Handbook of Psychological Testing*, Routledge, 2nd edition, 2000.
[9] Murphy, K. R. & Davidshofer, C. O., *Psychological testing: principles and applications*. New Jersey: Prentice-Hall, 2001.
[10] Hofer, F. & Schwaninger, A., Reliable and valid measures of threat detection performance in X-ray screening, *IEEE ICCST Proceedings, 38,* pp. 303-308, 2004.
[11] MacMillan, N. A. & Creelman, C. D., *Detection theory: A user's guide,* University Press: Cambridge, 1991.
[12] Green, D. M. & Swets, J. A., *Signal detection theory and psychophysics,* Wiley: New York, 1966.