Transportation Research Part A 200 (2025) 104640

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Reliability and validity of threat image projection data as a measure of performance in X-ray baggage screening

D. Buser [a], A. Schwaninger [a,b], V. Rehor [c], Y. Sterchi [a,b,*]

[a] University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Olten, Switzerland
[b] Center for Adaptive Security Research and Applications (CASRA), Zurich, Switzerland
[c] Czech Technical University in Prague, Faculty of Transportation Sciences, Department of Air Transport, Prague 1, Czech Republic

ARTICLE INFO

Keywords:
Aviation security
Baggage screening
Quality control
Threat image projection
Visual search
X-ray image inspection

ABSTRACT

Passenger baggage is screened using X-ray machines at airports worldwide to ensure transportation security. Many airports use a technology called threat image projection (TIP) to measure detection performance of airport security officers (screeners). TIP projects prerecorded X-ray images of prohibited items into X-ray images of passenger baggage being screened, and each time a TIP image is displayed (a *TIP event*), the TIP system records whether the screener detected the prohibited item. Because the prohibited items and the location of their placement in bags are randomly selected, the resulting TIP images vary substantially in difficulty and do not always look realistic. It is therefore not clear whether TIP data provides a good measure of screener performance, despite the technology's long-standing and widespread use at airports. To address this research gap, we conducted a study to estimate TIP's psychometric properties of reliability and validity by analysing a large set of TIP data from cabin baggage screening of a European airport (1,199,838 TIP events from 728 screeners over four years). We found the reliability of performance measurement to increase with the number of TIP events in accordance with the Spearman–Brown prediction. Approximately 100 TIP events were sufficient to achieve a minimum reliability value of 0.7 when TIP was challenging enough (mean hit rate below 0.9). TIP performance predicted the covert test results (wherein instructed people tried to smuggle real prohibited items through the checkpoint; 1,184 covert tests from 474 screeners), indicating that TIP is a valid measure of detection performance in X-ray baggage screening. The results imply that TIP data provides a reliable and valid performance measure if the TIP images are challenging enough and about 100 TIP events are considered per screener.

1. Introduction

Searching for threat items in baggage using X-ray technology is an important component of ensuring security in many areas, for

Abbreviations: CBS, cabin baggage screening; CI, confidence interval; CTT, classical test theory; EU, European Union; FTI, fictional threat items; GEE, generalized estimation equations; HBS, hold baggage screening; IED, improvised explosive device; TIP, threat image projection.

* Corresponding author at: University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Institute Humans in Complex Systems, Olten, Switzerland.

E-mail addresses: daniela.buser@fhnw.ch (D. Buser), adrian.schwaninger@fhnw.ch (A. Schwaninger), rehorvac@fd.cvut.cz (V. Rehor), yanik.sterchi@fhnw.ch (Y. Sterchi).

https://doi.org/10.1016/j.tra.2025.104640
Received 25 February 2025; Received in revised form 13 August 2025; Accepted 13 August 2025
Available online 28 August 2025

example in transportation, malls, prisons, customs, or sporting events. In transportation, baggage screening is standard practice at airports, but it is also used in other modes of transportation, e.g. at train stations (Eurostar, n.d.; Yu et al., 2025). To ensure a high level of security, many airports use a technology called threat image projection (TIP). With TIP, pre-recorded X-ray images of prohibited items (bombs, guns, knives, etc.) are projected onto the X-ray images of passengers' baggage (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Skorupski and Uchroński, 2016). The TIP system records the screeners' responses each time a TIP image is displayed (*TIP event*), thereby allowing to measure how well they detect prohibited items. The recorded TIP events are frequently used for quality control by airports, governments, and security companies (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Riz à Porta et al., 2022; Skorupski and Uchroński, 2016). Although TIP is considered established by many practitioners, there is a lack of research showing that TIP data provides a reliable and valid performance measure in the psychometric sense, i.e. that TIP data provides an accurate measure of detection performance that reflects how well screeners detect real threats. We therefore examined the reliability of TIP as a performance measure by analysing a large set of TIP data from an international airport of four years. We also assessed the validity of TIP performance by analysing whether it can predict how well screeners detect prohibited items in covert tests, wherein instructed people attempted to smuggle prohibited items in their baggage (e.g., knives, bombs, or guns) past the checkpoint (Walter et al., 2021; Wetter et al., 2008).

2. Background and theory

TIP is one of possible measures included in national regulation (National Aviation Security Plan, NASP) to ensure the detection of threats in passenger baggage, alongside, for example, a specified percentage of manual hand searches. More specifically, with so-called FTI TIP, pre-recorded images of threats, known as fictional threat items (FTIs), are projected onto 1 %–4 % of all passenger baggage during screening (Cutler and Paddock, 2009; Hofer and Schwaninger, 2005; Meuter and Lacherez, 2016; Skorupski and Uchroński, 2018). When screeners suspect a prohibited item, they press a designated button and the TIP system immediately provides feedback on whether an FTI is present or not (Bassetti et al., 2015; Schwaninger, 2006). The screeners' responses to such TIP images are recorded by the TIP system to evaluate the screeners' detection performance, typically calculated as the hit rate, that is the percentage of projected TIP images detected by the screener, hereafter referred to as *TIP performance* (Hofer and Schwaninger, 2005; Meuter and Lacherez, 2016). The European Union prescribes additional training in case a screener's TIP performance lies under a certain threshold (Bassetti, 2017). Some airports also reward or penalize screeners depending on their TIP performance (Bassetti, 2017; Di Cato, 2021). Researchers analyse TIP data to answer various research questions related to visual search and human factors in baggage screening tasks (Buser et al., 2023; Meuter and Lacherez, 2016; Skorupski and Uchroński, 2016). Other than performance measurement, TIP also addresses the challenge that most actual threats, such as real bombs or guns, are extremely rare in daily screening. Research has shown that people are less likely to detect rare targets, a phenomenon known as the target prevalence effect (Godwin et al., 2010; Menneer et al., 2010; Wolfe et al., 2005, Wolfe et al., 2007). By increasing the number of threats to be detected, TIP aims to mitigate this effect while also increasing motivation by providing feedback on detection performance (Cutler and Paddock, 2009; Harris, 2002; Riz à Porta et al., 2022; Schwaninger, 2009).

TIP images vary in difficulty and do not always look realistic (Riz à Porta et al., 2022). Therefore, the difficulty of the TIP images presented to each screener varies and it is unclear how many TIP images are needed for the random variation in difficulty to even out (due to the law of large numbers) and for the performance assessment to become precise and consistent, i.e., psychometrically reliable. Further, it is unclear to what degree TIP-based performance assessment reflects the screeners' performance in detecting actual threats, i.e., how psychometrically valid TIP is. In the remainder of this chapter, we will provide a short introduction into the core concepts related to reliability and validity as well as their empirical estimation based on Murphy and Davidshofer (2014). Reliability refers to the extent to which a measurement is precise or consistent, e.g. a person achieves similar scores when tested repeatedly or with two alternate forms of a test. Ensuring reliability is of particular importance when the measurement of a skill or construct has consequences for the tested subjects (Murphy and Davidshofer, 2014). The reliability of a measurement method can be quantified by adopting a statistical model, like the classical test theory (CTT). CTT was originally conceptualized with focus on questionnaires and tests. It assumes that a single underlying dimension is being measured and that every person has a single true score (T) on that dimension (Murphy and Davidshofer, 2014; Rindskopf, 2015). Therefore, all test items (in case of TIP: all TIP images) should measure the same construct (threat detection). It is assumed that a person's observed score, X, is equal to the sum of their true score, T, and the measurement error, e:

$$X = T + e \tag{1}$$

CTT assumes that the errors have a mean of zero, are independent of the true score, and that errors on different measures are independent. A consequence of this model is that the variation across individuals or the variance in the observed test scores is the sum variance of true scores (true variance), σ_T^2, and the variance in the error (error variance), σ_e^2.

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \tag{2}$$

Based on this equation, reliability is defined as the proportion of total variance in scores attributable to true variance, σ_T^2, rather than error variance, σ_e^2.

$$Reliability = \frac{\sigma_T^2}{(\sigma_T^2 + \sigma_e^2)} = \frac{\sigma_T^2}{\sigma_X^2} \quad (3)$$

Values close to one indicate a reliable test, whereas a reliability of zero indicates that the measured scores are purely random and not reflective of the measured construct. This statistical model leads to several possible methods for estimating reliability (for an overview see: Murphy and Davidshofer, 2014; Rindskopf, 2015); however, in principle, two (or more) tests are used and the correlation between the scores of the tests is calculated, providing an estimate of reliability. To determine the reliability of TIP as a performance measure, we employed the split-half reliability method (Murphy and Davidshofer, 2014). In this method, items of a test (in this case, responses to TIP events over a certain period) are split into two groups and the correlation between the test scores of both halves is calculated to indicate the consistency between the two halves. A limitation of this correlation is that it indicates the reliability for only half of the available items, which is lower than the reliability for the full set as reliability increases with the number of considered items. Under the CTT assumptions, changes in reliability based on the number of items can be estimated using the Spearman–Brown prediction (Brown, 1910; Spearman, 1910). This formula is commonly used to correct the split-half reliability for full tests or to calculate how long a test would have to be to achieve a certain reliability.

$$r_2 = \frac{kr_1}{1 + (k-1)r_1} \quad (4)$$

hereby, r_1 is the reliability of the original test and r_2 is the predicted reliability of a test that is longer than the original test by factor k.

It is suggested that reliability should reach a minimum value of 0.7 (Kline, 2000; Murphy and Davidshofer, 2014). This applies if the test results are used as a first indication (e.g., dividing the screeners into two performance groups) or for group diagnostics. However, if test results have consequences for the individual (e.g., not getting hired or having to undergo remedial training), it is highly recommended to achieve higher reliability values of at least 0.8, and ideally above 0.9 (Brough, 2018; Murphy and Davidshofer, 2014).

As introduced above, reliability coefficients indicate the proportion of the variance in score that is attributable to true variance as opposed to random error (e.g., a reliability of 0.8 indicates that 80 % of the variance between people's score stems from true differences between the people and 20 % stems from error), and therefore, provides a measure of precision relative to the score's variance. If the precision of a measurement in absolute units is more of interest, the standard error and confidence intervals (CI) can be derived from the reliability coefficient of the test, r, and the standard deviation of test scores, σ, under CTT assumptions.

$$SE = \sigma_X \bullet \sqrt{(1 - r)} \quad (5)$$

The standard error indicates the variability in test scores attributable to measurement errors in absolute terms (i.e., in the units of the score; Murphy & Davidshofer, 2014). The standard error can then be used to compute CI, which indicate the range in which the true score of an individual lies with a certain level of confidence in a normal distribution. Fig. 1 illustrates the 95 % CI (the interval that includes the true TIP performance with a 5 % probability of error) of the hit rate for reliability values of 0.7 and 0.8, based on the standard deviation observed in our study.

How well CTT assumptions can be applied to TIP data is unclear because of the differences between TIP and standardized tests. With FTI TIP, every image that serves as a test item is different because FTIs are always projected onto a different X-ray image of passengers' baggage at a random position. Furthermore, different collections of FTIs (so-called TIP libraries) can be used at one airport, and for each library, TIP images can change when updated. In addition, screeners differ in the number of TIP images they analyse, and responses to TIP events are collected over a long period of time (six months), during which the performance of screeners can change. Therefore, it should be investigated whether reliability estimates based on CTT can be applied to TIP data.

To our knowledge, only one study has examined the reliability of TIP data as a performance measure (Hofer & Schwaninger, 2005) for different screening tasks, using data from cabin baggage screening (CBS) and hold baggage screening (HBS). CBS refers to the

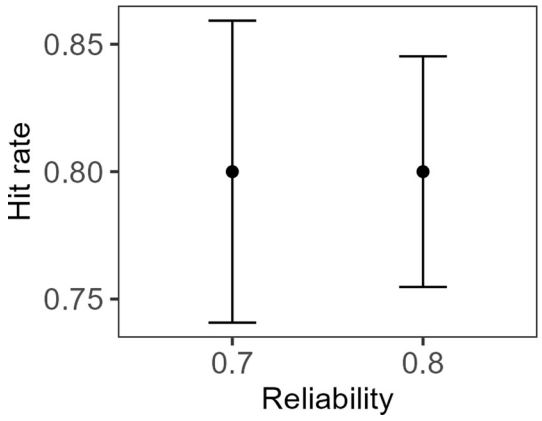

Fig. 1. 95 % CI for an average hit rate of 0.8 and a reliability of 0.7 (left) or 0.8 (right), based on the standard deviation observed in our study.

screening of passengers' carry-on baggage at airport security checkpoints, whereas HBS involves the screening of checked baggage that is transported in the hold of an aircraft. In their study, the researchers split TIP data into two groups based on odd and even days and calculated multiple split-half reliabilities using different data aggregations. They found that performance measurement was reliable for HBS TIP data (reliability >0.7), but not for CBS TIP data (reliability ≤ 0.58). The authors offered two possible explanations for this difference. First, high TIP hit rates in the CBS data caused ceiling effects in performance and the resulting small inter-individual differences in performance could have led to the low reliability. Second, the HBS TIP system used combined threat images (CTI), that is, pre-recorded X-ray images of passenger baggage containing a prohibited item were shown randomly in-between the X-ray images of real passenger baggage. The CBS TIP projected FTIs into X-ray images of passenger bags being screened. When using TIP with CTI, a higher quality and realism of TIP images can be achieved by visual inspection of the CTI before using them and by removing TIP images that are too easy or difficult. Moreover, when using TIP with whole bag images, also the false alarm rate can be measured by showing images to screeners that do not contain prohibited items (non-threat images, NTI). A limitation of the study by Hofer and Schwaninger (2005) was that it did not directly consider the number of TIP events, and the two halves of the test were unequal in size. Further, it did not correct the reduced number of TIP images due to splitting the data.

For TIP to be an effective performance measure, it must provide results that represent how well a screener would detect real threats. Therefore, it should not only be reliable, but also a valid measure of the detection performance. Validity is defined as the extent to which a test measures what it claims to measure (e.g., Murphy & Davidshofer, 2014). In case of TIP, that means TIP data should reflect how well screeners detect real prohibited items. A study by Riz à Porta et al. (2022) showed that screeners at an international airport considered approximately one-third of all CBS-TIP images to contain artifacts and look unrealistic. From this perspective, the validity of TIP performance seems uncertain. A quantitative method to evaluate validity is to analyse the degree to which an indirect performance measure can predict performance on the actual task. If the TIP performance is a valid measure of threat detection, it should be able to predict how likely a screener is to detect real prohibited items in an X-ray image. This can be done by comparing TIP performance with performance in covert tests, wherein instructed people attempt to smuggle real prohibited items past the checkpoint. These tests serve as a quality control measure and are often conducted by governments, airport, and police staff (Schwaninger, 2009; Skorupski & Uchroński, 2016; Walter et al., 2021; Wetter et al., 2008). As part of evaluating how well a mobile game can be used to assess screeners' competency in threat detection, Mitroff et al. (2017) examined TIP and covert test data among other measures. The researchers found a correlation of 0.31 in one analysis, but they did not find a correlation in a second data set and concluded that the validity of TIP data to assess screener performance remains unclear.

To investigate whether TIP data provides a reliable and valid measure of screener performance, we first assessed whether the Spearman-Brown prediction accurately describes how the reliability changes with the number of TIP images used to estimate screener performance with TIP data from an international airport. We then examined how the reliability developed over time and analysed how many screeners had completed a sufficient number of TIP events to allow for a reliable assessment of individual TIP performance. Finally, our study systematically analysed whether the TIP performance is a predictive measure of real threat detection by analysing the association between TIP and covert test performance.

3. Materials and methods

3.1. Sample

For our study, FTI TIP and covert test data from an international airport covering four years of CBS was analysed. The data included all responses to the TIP events of the professional airport security screeners, who were selected, qualified, trained, and certified according to the standards set by the appropriate national authority (civil aviation administration) in compliance with the relevant regulations in the European Union (European Commission, 2015). All responses were recorded as either hits (TIP detected) or misses (TIP missed). The entire dataset included 1,199,838 TIP events from 728 screeners. To analyse the validity of the TIP performance, we estimated a predictive model with 1,194 covert test results from a subsample of 474 screeners. The study complied with the American Psychological Association's Code of Ethics and was approved by the Institutional Review Board of the School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland.

Table 1
Number of FTIs per threat category of the two libraries used at the airport.

Threat category	Number of FTIs	
	Single-view library	Dual-view library
IEDs	600	644–652
Guns	100–148	148–152
Knives	124–252	146–152
Others	48–144	52
Total	1000	1000–1008

Numbers that changed over the four-year period are provided as a range.

3.2. Procedure

Similar to other airports (Michel et al., 2014), the screeners worked at four positions at the checkpoint and rotated between these positions, usually in intervals of about 20 min. When a screener rotated to the X-ray image inspection, they logged into the workstation with a unique user ID and inspected the X-ray images of the passengers' baggage for prohibited items directly next to the X-ray machine at the checkpoint. Two types of X-ray machines (single-view machines and dual-view machines, both without automated explosive detection) were used for cabin baggage screening at the airport, each type with its own TIP library producing about half of the TIP images analysed in this study. The airport also tested several new screening systems over the course of the study, the data of which was excluded from this study. Table 1 shows the number of FTIs per threat category for the two libraries, whereby each threat item was represented by four FTIs showing the item from different angles. According to the airport, at least 10 % of the TIP images were replaced every year, and the TIP libraries were compliant with EU regulations. The TIP systems projected FTIs onto X-ray images of passengers' baggage with a target prevalence of 2.9 %. The screeners were aware that their detection performance was monitored with TIP. When screeners suspected a prohibited item, they indicated this by pressing a specific button, and the TIP system provided immediate feedback on whether an FTI was present or not. If an FTI was present, the X-ray image had to be analyzed again without the FTI. If no FTI was present, the relevant piece of baggage was further inspected (through manual search or explosive trace detection). After X-ray image inspection, screeners logged out from the workstation and continued working at another checkpoint position or took a break.

At the target airport, covert tests were conducted regularly by the airport. For this, auditors were recruited and instructed to smuggle a threat item through security control. The selection and placement of prohibited items followed the protocol defined by the airport's quality control team. The prohibited items corresponded to the same categories as those in the TIP (guns, bombs, knives, etc.) and were placed either in the baggage or on the person. For each test, the prohibited item category, the location of where the item was placed, the difficulty of the test, the checkpoint at which the test was conducted, the type of X-ray machine that was used, and the date and time were protocoled. After the covert test, the outcome was discussed with the involved screeners and the difficulty of each test was evaluated again by the quality control team by reviewing the X-ray image of the baggage recorded during the test. The outcome was documented as either a hit (item found) or a miss (item not found).

3.3. Analyses

3.3.1. Reliability of TIP as a performance measure

To investigate the reliability of TIP performance, we assessed the split-half reliability using the following procedure: TIP events were first sorted by date and time of occurrence, and every two consecutive TIP events per screener were paired. To estimate the reliability for n number of TIP events, n pairs of TIP events were randomly selected (without replacement), and the TIP events of each pair were randomly split into two groups. The sorting by date and time ensured that both groups of TIP events had a comparable distribution over time, which is desirable when estimating the split-half reliability and comparable to the odd–even split (Murphy & Davidshofer, 2014). For each screener and each of the two groups of TIP events, the TIP hit rate (proportion of detected TIP events) was calculated, and the Pearson correlation between the hit rates of the two groups across all screeners was computed. To reduce variation due to the random sampling and splitting of TIP events, these steps were repeated 10,000 times, which is a common approach (e.g., Mundform et al., 2011). The correlation coefficients across the 10,000 repetitions were averaged, and the resulting standard errors of the mean were all below 0.001, indicating a good precision.

As airports and authorities often consolidate TIP performance on a half-year basis, our analyses were conducted for half-year periods from July 2015 to June 2019. In the first step, we determined whether the Spearman–Brown prediction accurately described how the reliability varied as a function of the number of TIP events considered for TIP performance evaluation. To estimate the reliabilities for up to 100 TIP events based on the same sample of screeners, only screeners with at least 100 TIP events per six months were included. We calculated the split-half reliabilities (as described above) with 5–50 TIP events per split, in increments of five. These estimates were then compared to the Spearman–Brown prediction based on 25 TIP events (per split). In a second step, we determined the reliability of each half-year period for a fixed number of 10 TIP events per screener and split, to retain more screeners for this analysis. Consequently, screeners that did not have a minimum of 20 TIP events within the respective half-year period were excluded from this analysis.

Table 2 shows the number and percentage of TIP events and screeners excluded per half-year based on the requirement of having a minimum of 20 or 100 TIP events, respectively. According to the airport, the share of screeners with less than 20 images per half-year increased due to more screeners working part time in other areas or on other screening systems not included in our study. Based on the Spearman–Brown prediction, we then calculated the reliability for higher numbers of TIP events.[1]

3.3.2. Validity of TIP of TIP as a performance measure

To assess the validity of the TIP data, we used TIP and covert test data from the X-ray positions in CBS (i.e., all covert tests with prohibited items hidden on the person or performed at checkpoints not for CBS, like staff screening checkpoints, were excluded) ranging from July 2015 to June 2019. For each covert test, all TIP events from the same screener within half a year before or after the

[1] The data originated from two separate TIP systems. An additional analysis revealed that the reliabilities of the two systems were comparable, and that the data can be analyzed jointly. Individual analysis of the libraries resulted in a reliability value of 0.45 and jointly of 0.43 using 50 TIP events.

Table 2
Number of TIP events and screeners excluded from reliability analysis per half-year.

Half-year period	Total number of screeners	Exclusion of screeners with <20 TIP		Exclusion of screeners with <100 TIP	
		Screeners excludedn (%)	TIP events excluded n (%)	Screeners excludedn (%)	TIP events excluded(%)
1	403	15 (3.72 %)	106 (0.08 %)	75 (18.61 %)	3523 (2.56 %)
2	410	17 (4.15 %)	139 (0.10 %)	68 (16.59 %)	3076 (2.24 %)
3	426	13 (3.05 %)	96 (0.06 %)	63 (14.79 %)	3467 (2.08 %)
4	446	16 (3.59 %)	134 (0.08 %)	71 (15.92 %)	3395 (2.10 %)
5	462	24 (5.19 %)	258 (0.13 %)	72 (15.58 %)	2768 (1.44 %)
6	482	35 (7.26 %)	298 (0.18 %)	82 (17.01 %)	2568 (1.53 %)
7	500	41 (8.20 %)	373 (0.27 %)	103 (20.60 %)	3511 (2.53 %)
8	497	49 (9.86 %)	348 (0.36 %)	124 (24.95 %)	4989 (5.11 %)

covert test were averaged. In total, 1,184 covert tests (184 with guns, 929 with IEDs, 123 with Knives, and 247 with other prohibited items) from 468 screeners were considered, and on average, 1,072 TIP events (SD = 929) and 2.53 covert tests (SD = 1.80) were analyzed per screener. For the prediction, a binomial generalized estimation equations (GEE; Ballinger, 2004; Liang & Zeger, 1986) was calculated,[2] as the data included multiple covert tests for most screeners and GEEs are suited to fit generalized linear models with longitudinal and clustered data. To facilitate model estimation, the TIP hit rate was z-transformed. The model also controlled for the prohibited item category (gun, knife, bombs, etc.), different checkpoints within the airport, X-ray machine type, and complexity of the covert test. The binomial GEE was estimated using the R-package GEE (Carey, 2024), with responses clustered within screener and with an exchangeable correlation structure. All analyses were performed using R (R Core Team, 2024).

4. Results

4.1. Reliability of TIP as a performance measure

The Spearman–Brown prediction corresponded well with the empirically estimated reliabilities and, therefore, provided an accurate description of how the reliability increased with the number of TIP events, as illustrated in Fig. 2 for the first three of the eight half-year periods (the other five half-year periods showed the same level of correspondence). Fig. 3 shows the split-half reliability of the TIP performance for 20, 50, 100, or 345 TIP events per half-year (345 is the average number of TIP events inspected by a screener).

As shown, reliability decreased over time. Decomposing the reliability into true variance and standard error (Fig. 4A and B) shows that the decrease in reliability was not attributable to an increasing measurement error (which instead also decreased over time); rather, the reliability decreased because of an over-proportionate decrease in the true variance of the TIP performance between the screeners, meaning that screeners differed less in their TIP performance. As shown in Fig. 4C, the average TIP hit rate increased over time, which might have led to a ceiling effect, i.e., limited room for inter-individual differences between screeners.

Table 3 shows the necessary number of TIP events per screener to reach a reliability of either 0.7, 0.8, or 0.9 for each half-year period. Although considering 93 TIP images for performance evaluation was sufficient to achieve a minimum reliability of 0.7 in the first half-year, 189 TIP images were necessary to obtain an equal reliability in the eighth half-year. To reach a reliability of 0.9, 357 images would have been necessary in the first half year, up to 728 images for the eighth half-year, which was not reached by any of the screeners.

To illustrate how reliability translates into the precision of a single measurement, Fig. 1 shows in which 95 % confidence intervals a reliability of 0.7 and 0.8 result (95 % CI; the interval includes the true TIP performance with a 5 % probability of error) for an exemplary measured TIP hit rate of 0.8. The illustration assumes an average true variance corresponding to the average that was observed across the eight half-year periods. As can be seen, a higher reliability results in a smaller 95 % CI (at constant true variance).

4.2. Validity of TIP as a performance measure

The average covert test hit rate was 0.79 (SD = 0.40). As can be seen in Table 4, the GEE model revealed that TIP data predicted the covert test performance well with the odds of passing a covert test increasing by 0.48 for an increase in the TIP hit rate by one standard deviation (odd ratio = 1.481, $p < 0.001$). Fig. 5 shows the estimated relationship between the TIP hit rate and covert test hit rate. As can be seen, screeners with a higher TIP hit rate also showed a higher covert test hit rate.

5. Discussion

X-ray image inspection of passenger baggage is an integral part of today's transportation security. TIP performance data is frequently used to evaluate how well security officers detect threats in passenger baggage and screeners whose TIP performance falls

[2] Additionally, a generalized mixed model was estimated, which resulted in very similar coefficient estimates but showed singularity problems. We also repeated the GEE estimation with a subsample of screeners that had four or more covert tests, again resulting in very similar coefficient estimates.

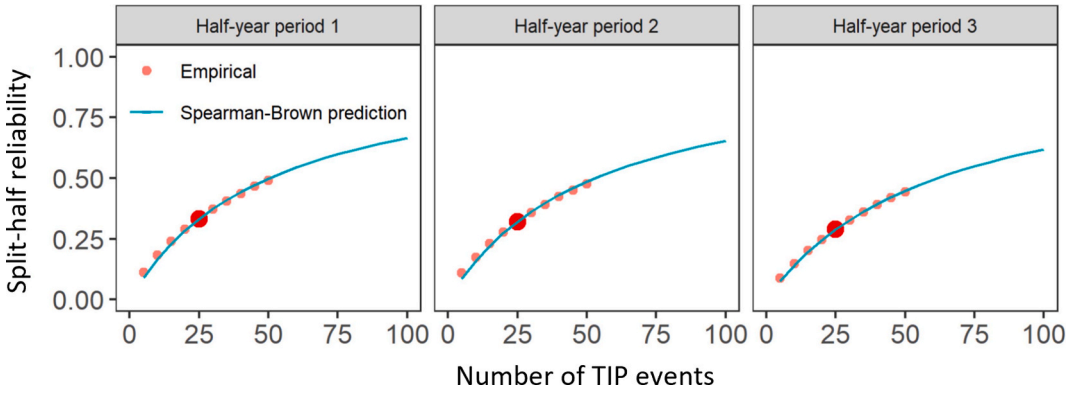

Fig. 2. Split-half reliability based on the number of TIP events considered for calculating TIP performance for the first (left), second (middle) and third (right) half-year period. The red dots indicate the empirically estimated split-half reliabilities. The blue lines show the predicted reliabilities based on the split-half reliability for 25 TIP events (large red dot). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

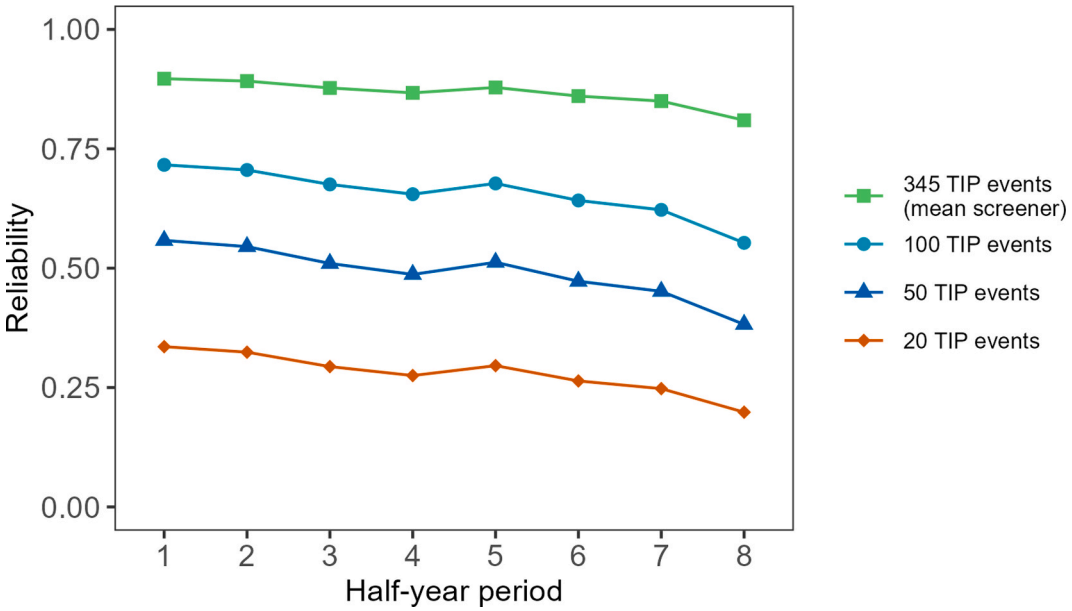

Fig. 3. Split-half reliability values for 20, 50, 100, and 345 TIP events (mean number of TIP events per screener per half-year period) for eight half-year periods.

below a defined threshold are often required to undergo remedial training and successfully complete a recertification process before they may resume screening duties (Bassetti, 2017; Riz à Porta et al., 2022). We investigated whether TIP data provides a reliable and valid measure of detection performance by analyzing data from an international airport covering four years. We found that reliability increased with the number of TIP events that are used for performance evaluation and this increase can be well described with the Spearman–Brown prediction. This finding is important because it enables the estimation of the reliability of TIP performance for a given number of TIP events per screener and the calculation of the necessary number of TIP events to achieve the desired reliability. For the first five half-year periods, during which the TIP hit rate was below 90 % and thus less affected by a ceiling effect, we found that approximately 100 TIP events were needed to achieve a statistical reliability of 0.7. A reliability value of 0.7 is recommended if the measure is used as a first indication or for group diagnostics (e.g. dividing screeners into two similarly performing groups to evaluate new equipment; Kline, 2000; Murphy & Davidshofer, 2014). However, if performance measures have consequences for screeners (e.g., mandatory remedial training), it is highly recommended to achieve higher reliability values of at least 0.8 (Brough, 2018; Murphy & Davidshofer, 2014). To achieve this, our results indicate that a larger number of TIP events is necessary (see Table 3).

Our analyses showed decreasing reliability throughout the four-year period. Although the error variance also decreased over time, there was a disproportionate decrease in the true variance between screeners, causing a decrease in reliability. It seems likely that the decline in the true variance was caused by an increase in the average TIP hit rate, leading to a ceiling effect. With an increasing number

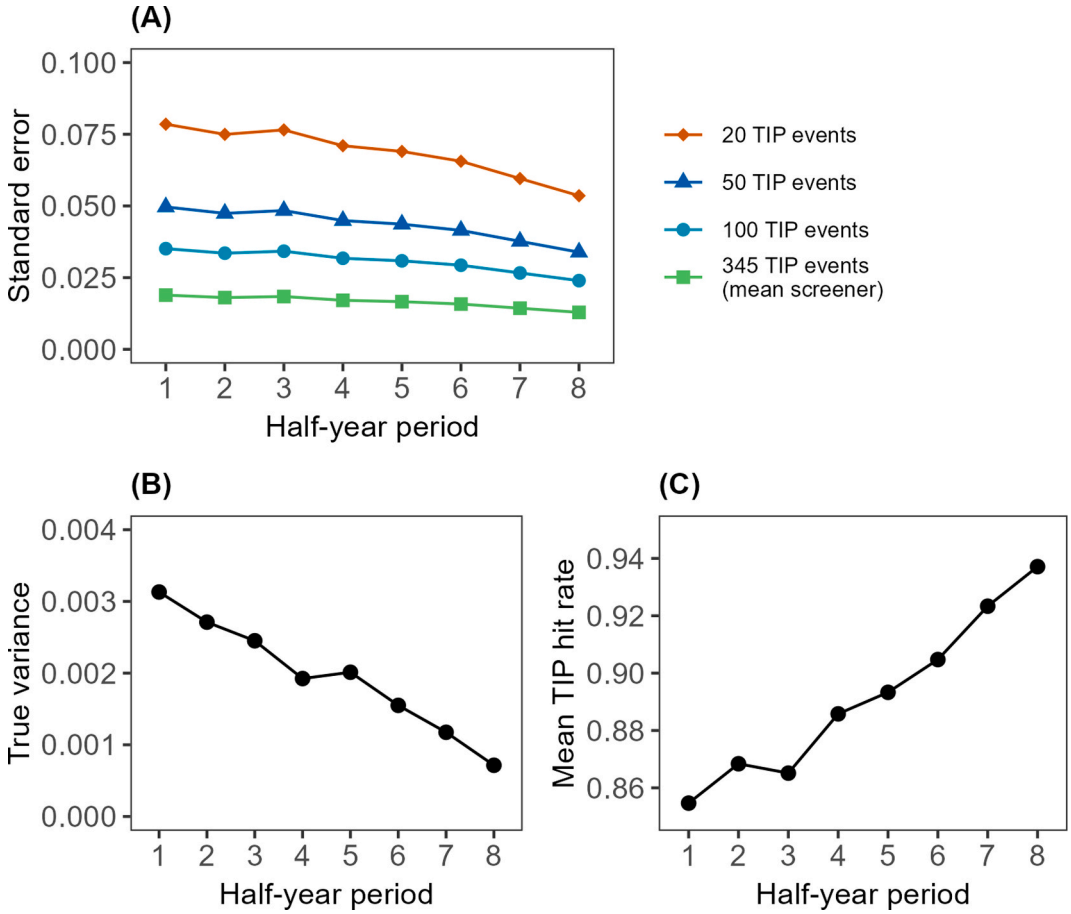

Fig. 4. Standard error (A), variance between screeners (B), and mean hit rate (C) for eight half-year periods.

Table 3
Number of TIP events required per screener to achieve a reliability of either 0.7, 0.8, or 0.9 for their performance measurement by half-year period.

Half-year period	N TIP required to achieve a reliability of 0.7 (% of screeners who did not analyze this number of images or more)	N TIP required to achieve a reliability of 0.8 (% of screeners who did not analyze this number of images or more)	N TIP required to achieve a reliability of 0.9 (% of screeners who did not analyze this number of images or more)
1	93 (16.38 %)	159 (25.06 %)	357 (51.86 %)
2	98 (15.61 %)	167 (22.93 %)	376 (57.32 %)
3	113 (16.43 %)	193 (23.47 %)	433 (53.76 %)
4	123 (18.16 %)	211 (26.91 %)	475 (65.92 %)
5	112 (16.23 %)	191 (21.43 %)	429 (49.78 %)
6	131 (19.29 %)	224 (28.63 %)	503 (72.61 %)
7	142 (25.8 %)	243 (45.6 %)	547 (92.4 %)
8	189 (52.72 %)	324 (81.69 %)	728 (100 %)
Average over all half-year periods	125	214	481

The brackets indicate the percentage of screeners who did not analyze at least the number of TIP images required to achieve the respective reliability.

of screeners reaching a TIP performance close to the possible maximum, TIP performance can no longer be distinguished between these high-performing screeners. Ceiling effects are a known reliability issue in performance measurement, for example, in competency assessment tests. The increase in the average TIP hit rate may have been caused by improvements in screeners' detection ability, e.g. due to experience (Schwaninger et al., 2010), training (Halbherr et al., 2013; Koller et al., 2008; Koller et al., 2009; McCarley et al., 2004), or personnel selection (Hardmeier et al., 2005; Mitroff et al., 2017; Rusconi et al., 2015). Another reason could be an increase in familiarity with the FTIs that were used in TIP due to repeatedly seeing the same FTIs when they are not replaced over a long period. It remains unclear how much each of these possible aspects contributed to the observed increase in TIP performance over time and the resulting drop in reliability.

Table 4
GEE results for covert test performance.

Term	β	Robust SE	Robust z	Odd ratio	p
Intercept	1.516	0.251	6.041	4.552	<0.001
TIP hit rate scaled	0.381	0.081	4.689	1.463	<0.001
Complexity: 2	−1.083	0.212	−5.113	0.339	<0.001
Complexity: 3	−2.129	0.276	−7.718	0.119	<0.001
Complexity: unknown	−0.281	0.234	−1.199	0.755	0.230
Category: bomb	0.093	0.223	0.415	1.097	0.678
Category: knife	−0.720	0.299	−2.407	0.487	0.016
Category: other	−0.715	0.250	−2.863	0.489	0.004
Checkpoint: gates	0.483	0.202	2.391	1.621	0.017
Checkpoint: staff & VIP	0.608	0.219	2.775	1.836	0.006
Checkpoint: transfer	1.326	0.276	4.799	3.764	<0.001
Machine type: 2	0.332	0.197	1.683	1.394	0.092

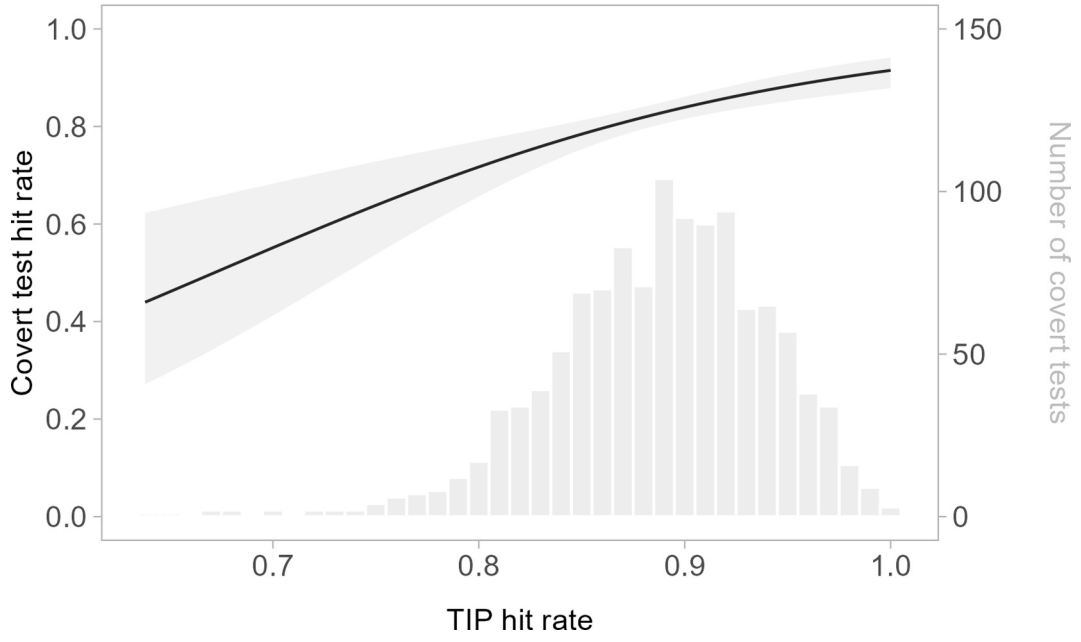

Fig. 5. Relationship between covert test and TIP hit rate (black line). 95% confidence band indicated by the grey area around the black line. The histogram shows the distribution of the TIP hit rates associated to the covert tests.

To achieve high reliability, its dependency on the difficulty of the TIP images should be considered. Our results showed a decrease in reliability over time for a constant number of TIP events, which was likely caused by an increase in the TIP hit rate. To avoid this, airports should regularly monitor the TIP hit rate. If the average hit rate of screener approaches 90 %, more FTIs should be replaced by focusing on easy FTIs (i.e., FTIs that are detected very often) to prevent ceiling effects in TIP hit rates. Furthermore, it is advisable to check the TIP images for artifacts (i.e., recognizeable differences from images of real prohibited items in the baggage; Riz à Porta et al., 2022) that can make TIP images unrealistically easy. For a given TIP system and library, reliability can be improved by increasing the number of TIP events used for performance assessment by extending the evaluation period or by increasing the TIP rate (the percentage of baggage images selected for TIP).

Reliability informs the share of variance attributable to true variance as opposed to the variance because of measurement error and, therefore, not only depends on the amount of measurement error (error variance) but also on how much individuals differ (true variance). Consequently, the reliability indicates how well the TIP performance distinguishes between high and low performing screeners, but it does not provide an indication of how precise an individual screener's TIP performance is. If one is more interested in the absolute TIP performance than in the comparison between individuals or groups, more informative standard errors and CIs can be derived from the estimated reliability and the variance across screeners. For example, Fig. 4 shows the 95 % CI for a reliability of 0.7 or 0.8, and the average true variance across all periods of our data set. These 95 % CI show the interval in which the true TIP performance lies when accepting a 5 % probability of error. As can be seen in Fig. 4, the 95 % CI decreases when reliability increases.

For TIP data to provide a useful measure of detection performance, it must be reliable and valid; TIP performance must reflect the performance in detecting real prohibited items. Bassetti (2021) reported that screeners sometimes recognize TIP images because they

appear artificial. Riz à Porta et al. (2022) found that a third of TIP images from an international airport look unrealistic. However, with two-thirds of the images looking realistic, the authors concluded that TIP performance should still achieve its purpose and largely reflect the performance in detecting real prohibited items. When validating a newly developed assessment tool, Mitroff et al. (2017) found TIP performance to moderately but significantly correlate with detection performance in covert tests in one data set, but not in a second data set. Using a large data set from four years, we found that TIP performance was significantly associated with detection performance in covert tests. In other words, screeners who performed better in detecting TIP images were more likely to detect prohibited items in covert tests. However, we found that the hit rate was higher in the TIP data than in the covert test data. This may indicate that TIP is not perfectly realistic. This is consistent with previous findings that TIP produces a share of unrealistic and easy images (Riz à Porta et al., 2022). In other words, one should not expect screeners with a TIP score of e.g. 90 % to detect 90 % of actual threats, as the latter are more difficult to detect in average. However, our results showed that TIP data has predictive validity and differentiates between screeners with high and low detection performance or can be used to compare detection performance over time.

Our finding that TIP performance can predict covert test performance does not mean that TIP should be seen as a possible replacement for covert tests. First, when screeners are exposed to threats solely through TIP, they may incorrectly assume that a bag is threat-free once the system indicates that no TIP is present (Schwaninger, 2009). Conducting frequent covert tests increases the screeners' expectation of threats occurring. Second, TIP is limited to assessing performance in X-ray image inspection, whereas covert tests can evaluate a broader range of screening activities, including manual bag searches and pat downs (Wetter et al., 2008; Schwaninger, 2009).

A limitation of our study is that we could only analyse the reliability and validity of TIP data from one airport using TIP with fictional threat items (FTI) for cabin baggage screening. It would be interesting to continue this research with TIP data from other airports. Moreover, our results showed an increase in TIP hit rate over the four-year period, which may have contributed to a ceiling effect. Whereas this pattern could reflect an actual improvement in screeners' detection skills, it may also be due to increased familiarity with the TIP images over time. With the data available for our study, we could not separate the influence of these two aspects on the reliability of TIP. Future studies could investigate how familiarity with TIP images, individual factors such as personnel selection processes, the screeners' experience, and training, as well as organizational aspects like the frequency of covert tests interact with TIP performance, to better understand variability in detection performance and improve the interpretation of TIP results.

In addition to the analysis of FTI TIP (as in our study), investigating the use of combined threat images (CTI TIP, images of baggage with integrated prohibited items; Hofer & Schwaninger, 2004) would be of interest. As such images consist of both baggage and prohibited items, they can be carefully prepared, and unrealistic images can be excluded beforehand, which should result in higher reliability and validity. By showing fully prepared images, images without prohibited items can be projected to assess the false alarm rate (Hofer & Schwaninger, 2004). In our study, we could not include false alarm data, similar to previous studies (Hofer & Schwaninger, 2005; Meuter & Lacherez, 2016; Mitroff et al., 2017; Skorupski & Uchroński, 2016), as the TIP system only provided aggregate counts of rejected non-TIP images. Many of these rejections are due to non-compliant bags (e.g., liquids exceeding the limit of 100 ml or not stored in a separate bag, and non-separated electronics), rather than actual false alarms. If the false alarm rate would be available, detection performance in terms of sensitivity and response bias could be calculated (Hautus et al., 2021), which can provide a more comprehensive view on the screeners' detection performance.

Despite these limitations, our results provide valuable information for the calculation of the reliability of TIP performance in terms of hit rate. We showed that the Spearman–Brown formula can be used to calculate the number of TIP events required to achieve the desired reliability. Our results suggest that approximately 100 TIP events are sufficient to achieve a reliability of 0.7 with a TIP library that is difficult enough (average TIP performance $<$90 %). Our study highlights the importance of avoiding ceiling effects to draw reliable conclusions about detection performance, which could be achieved by regularly exchanging TIP images. Additionally, we found clear evidence that higher TIP hit rates are associated with better covert test performance. Considering that TIP can provide a reliable and valid measure of threat detection, as shown by our study, and given its other benefits (Cutler & Paddock, 2009; Schwaninger, 2006), it should be considered employing this technology more widely wherever baggage is screened, also outside of airport security.

CRediT authorship contribution statement

D. Buser: Writing – original draft, Writing – review & editing, Methodology, Project administration, Resources, Visualization, Conceptualization. **A. Schwaninger:** Conceptualization, Funding acquisition, Methodology, Resources, Writing – review & editing. **V. Rehor:** Resources, Writing – review & editing, Conceptualization. **Y. Sterchi:** Writing – original draft, Writing – review & editing, Supervision, Validation, Funding acquisition, Methodology, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank the representatives of the airport who provided the data and Silvie Hrbková for her support in data preparation and analysis.

Data availability

The data that has been used is confidential.

References

Ballinger, G.A., 2004. Using generalized estimating equations for longitudinal data analysis. Organ. Res. Methods 7 (2), 127–150. https://doi.org/10.1177/1094428104263672.

Bassetti, C., 2017. Airport security contradictions: Interorganizational entanglements and changing work practices. Ethnography 19 (3), 288–311. https://doi.org/10.1177/1466138117696513.

Bassetti, C., 2021. The tacit dimension of expertise: Professional vision at work in airport security. Discourse Stud. 23 (5), 597–615. https://doi.org/10.1177/14614456211020141.

Bassetti, C., Ferrario, R., Campos, M.L.M., 2015. Airport security checkpoints: An empirically-grounded ontological model for supporting collaborative work practices in safety critical environments. ISCRAM 2015 Conference Proceedings 12th International Conference on Information Systems for Crisis Response and Management. Retrieved on June 17, 2025, from https://idl.iscram.org/files/chiarabassetti/2015/1258_ChiaraBassetti_etal2015.pdf.

Brough, P., 2018. Advanced research methods for applied psychology. In: Design, Analysis and Reporting, first ed. Routledge, London. https://doi.org/10.4324/9781315517971.

Brown, W., 1910. Some experimental results in the correlation of mental abilities. Br. J. Psychol. 3 (3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

Buser, D., Schwaninger, A., Sauer, J., Sterchi, Y., 2023. Time on task and task load in visual inspection: A four-months field study with baggage screeners. Appl. Ergon. 111, 103995. https://doi.org/10.1016/j.apergo.2023.103995.

Carey, V. J. (2024). GEE: Generalized Estimating Equation Solver (Version 4.13-20) [R package]. Retrieved on June 17, 2025, from https://CRAN.R-project.org/package=gee.

Cutler, V., Paddock, S., 2009. Use of threat image projection (TIP) to enhance security performance. In: 43rd Annual 2009 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 46–51. https://doi.org/10.1109/CCST.2009.5335565.

Di Cato, G. (2021). Now featuring on an X-ray monitor near you: No longer radio cassette recorders, alarm clocks and SLR cameras. *Transport Security International 2*, 28–30 [Online]. Retrieved on June 17, 2025, from https://tsi-mag.com/back-copies/summer-2021/#Transport_Security_International_Magazine/page_28-29.

European Commission, 2015. Commission implementing regulation (EU) 2015/1998 of 5 November 2015 laying down detailed measures for the implementation of the common basic standards on aviation security. *Official Journal of the European Union* [Online]. Retrieved on June 17, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R1998.

Eurostar (n.d.). Can I put photographic film through the security scanners? *Eurostar Help Center* [Online]. Retrieved on June 17, 2025, from https://help.eurostar.com/faq/us-en/question/Can-I-put-photographic-film-through-the-security-scanners.

Godwin, H.J., Menneer, T., Cave, K.R., Helman, S., Way, R.L., Donnelly, N., 2010. The impact of relative prevalence on dual-target search for threat items from airport X-ray screening. Acta Psychol. 134 (1), 79–84. https://doi.org/10.1016/j.actpsy.2009.12.009.

Halbherr, T., Schwaninger, A., Budgell, G.R., Wales, A., 2013. Airport security screener competency: A cross-sectional and longitudinal analysis. Int. J. Aviation Psychol. 23, 113–129. https://doi.org/10.1080/10508414.2011.582455.

Hardmeier, D., Hofer, F., Schwaninger, A., 2005. The X-ray object recognition test (X-ray ORT) - a reliable and valid instrument for measuring visual abilities needed in X-ray screening. In: Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 189–192. https://doi.org/10.1109/CCST.2005.1594876.

Harris, D.H., 2002. How to really improve airport security. Ergonomics in Design: The Quarterly of Human Factors Applications 10 (1), 17–22. https://doi.org/10.1177/106480460201000104.

Hautus, M.J., Macmillan, N.A., Creelman, C.D., 2021. In: Detection Theory: A User's Guide, third ed. Routledge, New York. https://doi.org/10.4324/9781003203636.

Hofer, F., Schwaninger, A., 2004. Reliable and valid measures of threat detection performance in X-ray screening. In: 38th Annual 2004 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 303–308. https://doi.org/10.1109/ccst.2004.1405409.

Hofer, F., Schwaninger, A., 2005. Using threat image projection data for assessing individual screener performance. WIT Trans. Built Environ. 82, 417–426. https://doi.org/10.2495/SAFE050411.

Kline, P., 2000. The Handbook of Psychological Testing, second ed. Routledge, London. https://doi.org/10.4324/9781315812274.

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. J. Transp. Secur. 1, 81–106. https://doi.org/10.1007/s12198-007-0006-4.

Koller, S.M., Drury, C.G., Schwaninger, A., 2009. Change of search time and non-search time in X-ray baggage screening due to training. Ergonomics 52 (6), 644–656. https://doi.org/10.1080/00140130802526935.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22. https://doi.org/10.1093/biomet/73.1.13.

McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., Boot, W.R., 2004. Visual skills in airport-security screening. Psychol. Sci. 15 (5), 302–306. https://doi.org/10.1111/j.0956-7976.2004.00673.x.

Menneer, T., Donnelly, N., Godwin, H.J., Cave, K.R., 2010. High or low target prevalence increases the dual-target cost in visual search. J. Exp. Psychol. Appl. 16 (2), 133–144. https://doi.org/10.1037/a0019569.

Meuter, R.F.I., Lacherez, P.F., 2016. When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening. Hum. Factors 58 (2), 218–228. https://doi.org/10.1177/0018720815616306.

Michel, S., Hättenschwiler, N., Kuhn, M., Strebel, N., Schwaninger, A., 2014. A multi-method approach towards identifying situational factors and their relevance for X-ray screening. In: 48Th Annual 2014 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 208–213. https://doi.org/10.1109/CCST.2014.6987001.

Mitroff, S.R., Ericson, J.M., Sharpe, B., 2017. Predicting airport screening officers' visual search competency with a rapid assessment. Human Factors: J. Human Factors Ergon. Soc. 60 (2), 201–211. https://doi.org/10.1177/0018720817743886.

Mundform, D.J., Schaffer, J., Kim, M.J., Shaw, D., Thongteeraparp, A., Supawan, P., 2011. Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. J. Mod. Appl. Stat. Methods 10 (1), 4.

Murphy, K.R., Davidshofer, C.O., 2014. Psychological Testing. Principles and Applications, sixth ed. Pearson Prentice Hall New Jersey.

R Core Team (2024). R*: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rindskopf, D., 2015. Reliability: Measurement. In: Wright, J.D. (Ed.), International Encyclopedia of the Social & Behavioral Sciences, second ed. Elsevier, Amsterdam, pp. 248–252. https://doi.org/10.1016/B978-0-08-097086-8.44050-X.

Riz à Porta, R., Sterchi, Y., Schwaninger, A., 2022. How realistic is threat image projection for X-ray baggage screening? Sensors 22 (6). https://doi.org/10.3390/s22062220.

Rusconi, E., Ferri, F., Viding, E., Mitchener-Nissen, T., 2015. XRIndex: a brief screening tool for individual differences in security threat detection in x-ray images. Frontiers in human neuroscience 9, 439.

Schwaninger, A., 2006. Threat image projection: Enhancing performance? Aviat. Sec. Int. 12, 36–41. https://doi.org/10.26041/fhnw-2096.

Schwaninger, A., 2009. Why do airport security screeners sometimes fail in covert tests?. In: 43rd Annual 2009 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 41–45. https://doi.org/10.1109/CCST.2009.5335568.

Schwaninger, A., Hardmeier, D., Riegelnig, J., Martin, M., 2010. Use it and still lose it? The influence of age and job experience on detection performance in x-ray screening. GeroPsych: J. Gerontopsychol. Geriatric Psych. 23 (3), 169–175. https://doi.org/10.1024/1662-9647/a000020.

Skorupski, J., Uchroński, P., 2016. A human being as a part of the security control system at the airport. Procedia Eng. 134, 291–300. https://doi.org/10.1016/j.proeng.2016.01.010.

Skorupski, J., Uchroński, P., 2018. Evaluation of the effectiveness of an airport passenger and baggage security screening system. J. Air Transp. Manag. 66, 53–64. https://doi.org/10.1016/j.jairtraman.2017.10.006.

Spearman, C., 1910. Correlation calculated from faulty data. Br. J. Psychol. 3 (3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x.

Walter, S., Hofer, F., Dolder, Z., Ghelfi-Waechter, S., 2021. Simulating for real: The why and how of security drills at the security checkpoint. J. Airport Manag. 15 (2), 147–159. https://doi.org/10.69554/IACW7292.

Wetter, O.E., Hardmeier, D., Hofer, F., 2008. Covert testing at airports: Exploring methodology and results. In: 42nd Annual 2008 International Carnahan Conference on Security Technology (ICCST). IEEE, pp. 357–363. https://doi.org/10.1109/CCST.2008.4751328.

Wolfe, J.M., Horowitz, T.S., Kenner, N.M., 2005. Rare items often missed in visual searches. Nature 435, 439–440. https://doi.org/10.1038/435439a.

Wolfe, J.M., Horowitz, T.S., Van Wert, M.J., Kenner, N.M., Place, S.S., Kibbi, N., 2007. Low target prevalence is a stubborn source of errors in visual search tasks. J. Exp. Psychol. Gen. 136 (4), 623–638. https://doi.org/10.1037/0096-3445.136.4.623.

Yu, X., Fan, C., Pan, J., Xiang, G., Chen, C., Yu, T., Peng, Y., Deng, H., 2025. X-ray security inspection for real-world rail transit hubs: A wide-ranging dataset and detection model with incremental learning block. Vis. Comput. 41, 1–13. https://doi.org/10.1007/s00371-024-03725-4.