Covert Testing at Airports: Exploring Methodology and Results

Olive Emil Wetter  
Zurich State Police  
Airport Division, Security Control  
Quality Management  
P.O. Box  
CH-8058 Zürich-Flughafen  
SWITZERLAND

Dr. Diana Hardmeier  
Zurich State Police  
Airport Division, Security Control  
Quality Management  
P.O. Box  
CH-8058 Zürich-Flughafen  
SWITZERLAND

Dr. Franziska Hofer  
Zurich State Police  
Airport Division, Security Control  
Quality Management  
P.O. Box  
CH-8058 Zürich-Flughafen  
SWITZERLAND

Abstract - Large efforts are made in order to enhance human factors in airport security, but one very important area, namely covert or infiltration tests, are often neglected. Covert tests are most important not only for quality control and risk analysis, but also for training purposes, such as for example the training of appropriate reactions in dangerous situations. Therefore, conducting covert tests at airports is a very useful approach for assessing the effectiveness of the security control and offers a possibility to uncover weaknesses in the security process by simulating the case of emergency. Its main benefit is the high ecological validity almost unparalleled by other methods. The present scientific exploratory applied study is one of the first of its kind worldwide, which analyses the results of covert tests collected at a large European airport over a period of fourteen calendar months. Statistical analyses were performed using Mann-Whitney U-Tests, Chi-Square tests and Binary Logistic Regression analysis. Results will focus on differences in detection performance between different test objects and different places of concealment. Additional factors included in the analyses will be time taken for pat-down search, sex, age, work experience, as well as the security officer’s reaction to weapon detection. Methodological problems will be discussed and suggestions made for further research in this area.

Index Terms — Covert testing, covert tests, infiltration tests, aviation security, security control, airport security, pat-down search.

I. INTRODUCTION

During the last decade, large efforts have been made to make civil aviation more secure. The attacks on September 11th, 2001, have strongly boosted this trend both inside and outside the United States. One way of enhancing security is to invest into high-tech equipment such as X-ray scanners, explosive detection systems or just recently body scanners at airports. However, this is only one side of the coin. It is obvious that the most sophisticated machines are useless if the operating personnel is inattentive or even incompetent. Thus, it was realized that at least equal investments have to be made into the human factor. As a result of this, computer-based training systems for X-ray screeners have been developed (e.g. X-ray Tutor [1, 2]). As X-ray equipment offers easy assessment of screener performance, the field of X-ray image interpretation has been quite widely researched by several authors from various countries (e.g. [3-5]).

One of the most obvious tasks of a security officer at an airport is the interpretation of an X-ray image in order to decide whether the bag contains a forbidden item or not. Another one of equal importance is to perform an effective pat-down search, which includes the correct operation of the hand-held metal detector (HHMD), or to closer screen a passenger if suspicion is aroused in order to make sure that the passenger has no forbidden item attached to his or her body. In spite of all these security requirements, a customer-friendly atmosphere has to be maintained since the vast majority of travelers are upright and decent people. Scientifically speaking, the base rate of terrorists or other criminals passing an airport’s security control is extremely low. This represents an additional challenge for the security officers. Because the base rate is that low, they might expect never to come into contact with a criminal subject. Nevertheless, they have to stay alert all the time since they absolutely have to react correctly in the rare case mentioned [6, 7].

A correct and appropriate reaction in case of a threat can only take place if the security officer has internalized the procedures which again can only be achieved by thorough training. In a first step, security officers learn, discuss and try out procedures and measures. In a second step, they have to have the possibility to test their abilities in a situation of high ecological validity. Covert tests are well suited to fulfill these requirements. Covert tests at airports are suitable to simulate the case of emergency, to train security officers, and to assess their performance. Moreover, results from covert tests can be used for quality control purposes or for conducting risk analysis.

Up to date, there exists no published scientific study which analyzes data from covert tests at airports. The present study is based on data collected over a period of fourteen calendar months. It is of exploratory nature insofar as ways of
conducting, analyzing and interpreting results of covert tests are discussed. Sharing experiences as well as identifying practical and methodological problems should foster future research in this area and contribute to enhance job performance of security officers.

II. METHOD

A. Definition

Already at the very beginning of the 17th century, Bacon urged the necessity of a common definition of terms in science: “So as it is almost necessary, in all controversies and disputations, to imitate the wisdom of the mathematicians, in setting down in the very beginning the definitions of our words and terms, that others may know how we accept and understand them, and whether they concur with us or no.” [8] Since frequent covert testing at airports is a relatively new methodology with little research having been done up to the present, there is no binding definition of the concept of “Covert Testing” available.

Publicly available documents from the United States [9-12] as well as personal communication with experts in the field reveal the following characteristics: Covert testing aims at simulating as closely as possible the situation in which a criminal subject attempts to take a threat object into the sterile area. Scientifically speaking, its goal is to achieve maximal possible ecological validity. In practice, a tester, unknown to the security officers on duty, appears and behaves as a traveler, staff or crew member while trying to pass the security control carrying one or more forbidden items with him/her. Based on these characteristics, the following short definition of “covert testing” at airports is proposed: “Covert testing” is a method of testing the security control’s ability to detect threat objects carried by a tester in a situation of high ecological validity.

B. Covert Tests

In this paper, data of covert tests carried out between April 2007 (when this new methodology was introduced) and May 2008 are analyzed. For a detailed description of the procedure of covert testing, see section “Procedure and Materials”. Some covert tests had to be aborted because the person pretending to be a passenger or staff member was recognized as testing person by the security officers. During others, failure of equipment or slipping out of the hidden threat object of its place of concealment occurred. Tests including such events have been judged as invalid trials. One test was judged invalid since the tester did not carry a threat object but a false ID badge, which is not consistent with the above definition of covert testing. Of course, this situation could happen in reality and could pose a threat to civil aviation, but, according to the previous definition, this paper focuses on smuggling a dangerous item past security control. Out of a total of 405 covert tests conducted, 372 valid ones (92%) remained for analysis.

C. Participants

372 covert tests were conducted, 79 employees happened to be tested more than once. 35% of the tested employees were female. As a regulation, searching passengers manually (pat-down search) is only allowed for security officers of the same sex. All of them are certified security officers on duty. Due to a shortage of personnel, a very small proportion of retired police officers assisted the security officers for a limited period of time. Age varied between 20 and 64 years ($M = 39.0$ years, $SD = 12.5$ years). The security officers are trained in X-ray image interpretation with an individually adaptive computer-based training (CBT, X-ray Tutor [1, 2]). The training system was installed in 2005 and focuses on cabin baggage screening. Instruction in pat-down search as well as in the general knowledge of procedures and responsibilities was provided and supervised by professional instructors of the airport police forces. Mean working experience of the tested security officers was 4.7 years ($SD = 6.0$ years).

D. Statistical analyses

Throughout this paper, nonparametric tests have been applied due to the fact that normal distribution of the variables could not be assumed. Chi-Square tests have been used for binary coded variables, Mann-Whitney U-Tests for continuous variables. Binary logistic regression analysis (forward) has been applied in an attempt to disentangle effects.

III. PROCEDURE AND MATERIALS

The procedure of covert testing closely followed the definition given in Section II. A. A tester unknown to the security officers either carried a threat object on his body or an Improvised Explosive Device (IED) in his carry-on baggage\(^1\). Places of concealment on the body were the arms, the legs, the feet, the upper body and the pubic area. Threat objects used were small guns (e.g. Beretta Model 21) or knives carried on the body as well as fuzes and IEDs in carry-on baggage with relatively low bag complexity and superposition (for a definition of these concepts, see [13]). Testers approached the security control dressed and behaving as normal passengers or in some cases as normal staff members. The covert tests took place at different checkpoints in the airport. The selection of checkpoints to be tested was neither balanced nor random but rather reflected the testers’ and the supervisors’ spontaneous choice. Table I offers an overview of the variables analyzed.

<table>
<thead>
<tr>
<th>TABLE I COLLECTED VARIABLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable name</td>
</tr>
<tr>
<td>Bimonthly</td>
</tr>
<tr>
<td>Checkpoint</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>TesterSex</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>Concealment</td>
</tr>
<tr>
<td>Threat</td>
</tr>
<tr>
<td>Discovery</td>
</tr>
<tr>
<td>Reaction</td>
</tr>
<tr>
<td>SearchTime</td>
</tr>
</tbody>
</table>

\(^1\) In rare cases, a tester carried two threat objects. To keep analyses simple, this was coded as two independent tests.
IV. RESULTS

A. Awareness and Training

In the authors’ subjective experience, the security officers’ alertness at work has increased considerably since the establishment of this new testing procedure. Over time, they have become more and more aware of the possibility of getting tested and the need of avoiding negative results. Furthermore, all of them have been trained in a standardized procedure of pat-down search during this period. This procedure has been developed based on insights gained during the period of data collection by the instructors and consists of a sequence of eleven standardized steps which have to be checked. It aims at providing an easy to remember guideline for a complete and thorough pat-down search, which guarantees best possible security. In addition, the introduction of frequent covert tests was appreciated because these tests form an ideal possibility to practice the new pat-down procedure at work. The mentioned subjective impressions and the additional efforts in pat-down search are reflected in the data as well. Figure 1 shows the trend of the detection rate in covert tests over time. Fitting a linear trend line to the data, explaining 34% of the variance, clearly shows an upward trend over time. The significance of this trend can further be confirmed with a Binary Logistic Regression using the bimonthly period as predictor and discovery as criterion variable ($Wald = 5.124, p < .05$).

![Fig. 1: Trend of the detection rate (bimonthly) and number of tests conducted.](image)

Fig. 1: Trend of the detection rate (bimonthly) and number of tests conducted.

B. Sex

Dividing the sample in two groups according to the security officers’ sex revealed that male security officers were more likely to detect threat objects than their female counterparts. This difference is significant with $\chi^2 (1, N = 242) = 19.64, p < .001$ and is of medium effect size with $w = .28$ (regarding the calculation of $w$, see e.g. [14]). Similarly, dividing the sample in two groups according to the testers’ sex revealed that threat objects carried by a male tester were more likely to be detected than those carried by female testers ($\chi^2 (1, N = 261) = 34.05, p < .001$). This significant effect is of medium effect size ($w = .36$).

Subsequently, the whole sample was divided again according to the security officers’ sex and we looked at the estimated times taken for pat-down search and search by HHMD. This analysis revealed that male security officers spent more time checking the testers ($M = 2.38 \text{ min.}, SD = 1.47 \text{ min.}$) than their female colleagues ($M = 2.00 \text{ min.}, SD = 1.33 \text{ min.}$). A Mann-Whitney U-test shows that this difference is significant ($U = 5688.50, p < .05$). Similarly, dividing the sample in two groups according to the testers’ sex revealed that the control procedure for male testers took longer ($M = 2.41 \text{ min.}, SD = 1.48 \text{ min.}$) than for their female counterparts ($M = 1.95 \text{ min.}, SD = 1.30 \text{ min.}$). This difference is significant as well with $U = 5593.00, p < .01$.

C. Age and Experience

Successful security officers were of lower age ($M = 38.18 \text{ years}, SD = 12.34 \text{ years}$) than those who failed in covert tests ($M = 41.39 \text{ years}, SD = 12.57 \text{ years}$). This difference is significant with $U = 11490.50, p < .05$. Further, working experience of security officers who failed in covert tests did not differ significantly from working experience of the ones who succeeded ($U = 11867.00, p = .184$). To test whether the fact that younger security officers perform better counteracts the effect of working experience, an analysis of covariance (ANCOVA) with success as between-subjects factor and age as covariate was conducted. The groups (succeeded, failed) did again not differ in working experience, even after having partialed out the effect of age. Please note that the result from this ANCOVA is to be treated with utmost caution because some assumptions (e.g. normal distribution of the dependent variable) may have been violated.

D. Search Time

As reported in Section IV. A, male security officers spent more time checking the testers than their female colleagues and, similarly, the control procedure for male testers took longer than those for the females. Analyzing the effect of search time on success or failure reveals that security officers who succeeded took more time for pat-down search ($M = 2.47 \text{ min.}, SD = 1.48 \text{ min.}$) than those who failed in covert tests ($M = 1.73 \text{ min.}, SD = 1.17 \text{ min.}$). This difference was significant ($U = 4390.50, p < .001$).

E. Place of concealment

Different places of concealment (see Section III.) were used for the threat objects. Those hidden in the carry-on baggage were detected best and those hidden in the upper body were detected worst. It is not surprising to us that objects hidden in the carry-on baggage were detected best. These discoveries were made by analyzing the X-ray image. Security officers at this airport are quite skilled in X-ray image interpretation due to long lasting training. Figure 2 offers an overview of the detection performance of the different places of concealment relative to the best category. The losses in detection performance relative to the reference category (carry-on baggage) are significant with $\chi^2 (4, N = 186) = 18.11, p < .01$.

---

2Since all numbers are considered security sensitive information, they are not indicated throughout this paper.
Carry-on baggage (reference) 

Arms Feet Pubic area Legs Upper body

Place of concealment

Loss in detection rate (%)

0 20 40 60 80 100 120

# Tests

Loss in detection Number of tests

Fig. 2: Relative loss in detection rate for different places of concealment relative to the category with the highest detection rate as reference category. Number of tests conducted indicated.

Figure 3 shows the loss in detection rate for different places of concealment for females and males separately. As can be seen, the observed large loss in detection rate in the upper body is due to the rather low performance of female security officers. It is plausible that this effect mainly occurs because of the anatomical differences between the two sexes. In addition, there is also quite a large difference in performance in the pubic area between the two sexes, as well as an opposite sex effect for arms. Smaller sex effects were observed for baggage and arms. Please note that the number of tests carried out for the different places of concealment as well as the two sexes vary to a large extent. For example, no tests were done for female officers using the legs as places of concealment. Therefore, further investigation has to be done to obtain reliable data allowing for statistical analyses and conclusions.

F. Checkpoint type

Covert tests were conducted either at public checkpoints or at staff/crew checkpoints. Although detection performance was slightly lower at staff/crew checkpoints, this difference was not significant ($\chi^2 (1, N = 96) = 2.00, p = .157$).

G. Type of threat object

Different types of threat objects (see Section III.) were used. IEDs were detected best and knives were detected worst. Figure 4 offers an overview of the detection performance of the different types of threat objects relative to the best category. Again, it is not surprising to us that IEDs and fuzes were detected best since they were hidden in the carry-on baggage. Items hidden in carry-on baggage might be easiest to detect because of the security officers having received adaptive training for cabin baggage screening since the beginning of 2005. The losses in detection performance relative to the reference category are significant with $\chi^2 (2, N = 213) = 12.82, p < .01$. This effect is of medium size ($w = .25$).

H. Disentangling effects

Up to this point, it remains unclear whether male officers perform better due to their sex (e.g. slightly different procedures in pat-down search for men and women) or due to the fact that they took more time for checking testers. In order to disentangle those effects, a Binary Logistic Regression was calculated. Table II shows the results of this analysis which tests the impact of the predictors SearchTime, Concealment, Age, Sex, Checkpoint, Threat, Bimonthly and Experience on the criterion variable Discovery.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>SearchTime</td>
<td>13.430</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Concealment</td>
<td>17.509</td>
<td>5</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1</td>
<td>.368</td>
</tr>
<tr>
<td>Threat</td>
<td>3</td>
<td>3</td>
<td>.096</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>1</td>
<td>.128</td>
</tr>
<tr>
<td>Checkpoint</td>
<td>1</td>
<td>1</td>
<td>.145</td>
</tr>
<tr>
<td>Bimonthly</td>
<td>1</td>
<td>1</td>
<td>.225</td>
</tr>
<tr>
<td>Experience</td>
<td>1</td>
<td>1</td>
<td>.600</td>
</tr>
</tbody>
</table>

n = 249; $R^2$ (Nagelkerke) = .200
The above results show that only the predictors SearchTime and Concealment remain in the regression. The same pattern can be observed if a Binary Logistic Regression is run excluding covert tests with IEDs in bags (see TABLE III).

TABLE III
BINARY LOGISTIC REGRESSION ANALYSIS, EXCLUDING COVERT TESTS WITH IEDs IN BAGS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>SearchTime</td>
<td>13.063</td>
<td>1</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Concealment</td>
<td>11.202</td>
<td>4</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td></td>
<td>.100</td>
</tr>
<tr>
<td>Threat</td>
<td>1</td>
<td></td>
<td>.337</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td></td>
<td>.493</td>
</tr>
<tr>
<td>Checkpoint</td>
<td>1</td>
<td></td>
<td>.148</td>
</tr>
<tr>
<td>Bimonthly</td>
<td>1</td>
<td></td>
<td>.239</td>
</tr>
<tr>
<td>Experience</td>
<td>1</td>
<td></td>
<td>.321</td>
</tr>
</tbody>
</table>

n = 227; R² (Nagelkerke) = .22

These analyses suggest that SearchTime and Concealment are important variables in connection with success or failure in covert tests and that the other predictors are of less importance. Interestingly, the clear trend over time (see section IV. A.) loses its significance when other predictors are taken into account. Note that the two significant predictors only explain between 20% and 22% of the overall variance in the criterion variable.

I. Reaction after IED discovery

If an IED has been detected in the X-ray image, the security officer has to react appropriately. In most cases, the security officers’ reaction was correct. The distribution of correct and incorrect reactions differed significantly from equal distribution with $\chi^2 (1, N = 56) = 28.23$, $p < .001$. This represents a large effect ($w = .71$).

V. DISCUSSION

In this study, we investigated in an explorative manner the influence of different factors on detection performance in covert tests. We revealed several different factors affecting performance. SearchTime appears to be the strongest predictor thereof, indicating that it is crucial that the security officers take enough time for the control. In other words, the longer the pat-down search, the more likely the prohibited item was detected, which simply indicates that a thorough control takes its time. However, such a thorough control has an effect on the throughput. Therefore, the challenge for each individual security officer and the airport is to find a balance between a proper control and the time spent for it. On the one hand, the base rate of a terrorist attack is quite low and the probabilities might be judged small, on the other hand, the possible arising costs of a failure could be tremendous and range from human and economic costs to the reputation of the airport [15-16].

Concealment was the second significant predictor in the Binary Logistic Regression analysis. In this study, security officers were best at finding threat objects using the X-ray machine. Objects hidden on the body seem to be somewhat harder to detect. The question arises if this is either due to the advantage of long lasting computer-based training in X-ray image interpretation (CBT for cabin baggage screening since 2005), or if detecting threat objects by pat-down search and by HHMD is simply more difficult, or if both explanations are true at the same time. It is obvious that the inspection of an X-ray image does not involve manual contact with the passenger, whereas pat-down search and search with the HHMD both involve much closer or even manual interaction with the passenger. The fact that the passenger’s privacy has to be respected makes an effective control of certain parts of the body, namely the upper body in women and the pubic area in both sexes, more difficult. We suggest that the results reflect both this reason and the long lasting training in X-ray interpretation.

There were no more significant predictors in the Binary Logistic Regression. Nevertheless, we believe that the impact of certain factors on detection performance per se are noteworthy. We suggest that by regular testing, a certain effect of training and awareness comes into play. Some security officers failed their first covert test and happened to be tested again. In almost all cases, their second performance was way better than their first. Some have reported that until the experience of failure, they were not aware of the difficulty and importance of a thorough pat-down search. As a result, they afterwards conducted much more thorough controls than before. Moreover, the upward trend to be seen in Figure 1 can be regarded as additional, data-based support for this statement. Further research with more standardized covert tests is needed to reveal a possible training effect in covert tests.

Although no more significant in the Binary Logistic Regression, we have a sex effect indicating that male security officers are more effective and that threat objects carried by a male tester are more likely to be found. One possible explanation for this is to be found in the anatomical differences between men and women, which require the pat-down search of a female to be much more careful since in women, the upper part of the body is a good location for concealing dangerous items (e.g. in the bra). There seems to be a slight inverse effect of age on detection performance, which almost reached significance in the Binary Logistic Regression analysis as well. Reasons might revolve around general effects of age on domains such as fluid abilities (e.g. [17]), working memory (e.g. [18, 19]), or vigilance (e.g. [20]). Experience on the job had no influence on detection performance which might be regarded as obvious because the security officers’ usual tasks are standardized routine procedures where experience counts less compared to thorough training. The different threat objects used reveal that those hidden in the carry-on baggage are detected best and small items hidden on the body (e.g. knives) are detected worst. In our opinion, however, consistent and correct use of the HHMD would in almost all cases indicate the (metallic) threat object. Currently, we are analyzing data focusing on this question.

It is wrong to think that the discovery of a threat object alone suffices and that the focus should be solely on that. If a threat object has been detected in a real situation, the security officer has to react appropriately. The next important step to be taken is a correct reaction and response to the threat detected. As a first approach to this question, it was analyzed whether security officers reacted correctly after an IED has
been detected in the X-ray image. In most cases, the reaction was correct, indicating that the security officers have internalized the procedures they were taught and that they are also applied in case of stress and perceived danger of life. At this point, we would like to encourage discussions about the ethic aspects of covert testing. The security officers’ job is already one of the most supervised of all jobs, at least in Switzerland, where yearly re-certification of a screener is legally anchored. Also the use of Threat Image Projection (TIP) gives the possibility to monitor performance continuously. With covert testing, one more powerful instrument for assessing performance is at hand. Especially when testing with IEDs, the question arises if it is ethically correct to put employees in a situation of perceived danger of life. We would like to encourage future researchers and practitioners in this area to think about this aspect.

Another question to think about revolves around habituation effects. If covert tests with IEDs are becoming more and more a standard procedure to monitor performance of security officers and are therefore introduced in high frequencies, habituation effects could occur. It could occur that a security officer’s first thought when faced with an X-ray image of an IED in a bag was that he was subjected to a covert test. The same is true for the use of TIP, where further research has to be done as well in order to find out more about long-term effects of such habituation aspects. It seems to be most important to find an optimal number of covert tests in order to make sure that security officers are aware of the potential of a terrorist attack all the time but that habituation effects can be ruled out or at least minimized.

Covert tests can serve as an instrument in order to reveal main weaknesses and are, therefore, one very powerful method to improve security sustainably. Our study shows that pat-down search and search by HHMD still somewhat lack behind in performance compared to performance in X-ray image interpretation and could therefore be improved further. We believe that standardized procedures and systems of pat-down search as well as consistent and correct use of the HHMD should be the next measure. Obviously, consistent and thorough controls take their time and we would like to urge authorities to provide enough time (and money) for an effective control, since time is the top factor affecting detecting performance. Frequent covert tests are in our view not only an instrument for assessing performance, but can also be used to train and motivate security officers and to keep them alert.

It remains to be shown if our findings are true for other airports as well since other airports have different procedures, training programs and employees. Therefore, replications at other locations are needed. We suggest conducting studies with a more standardized approach than the one used for this exploratory study. If possible, a design should also offer a possibility for detecting causality. In order to share our experiences, we would like to urge that it is necessary to conduct enough tests in order to get reliable results. In our view, our sample of 372 tests appears to be rather at the low end compared to the multitude of variables involved. Further methodological problems revolve around certain variables’ level of measurement (binary). For working with correlations, regression analyses and maybe even more sophisticated statistical procedures, a higher level of measurement would be desirable. We hope that the results discussed and the hints given can prepare the soil for future studies in covert testing in order to improve aviation security.

VI. REFERENCES

