ONE YEAR LATER: HOW SCREENER PERFORMANCE IMPROVES IN X-RAY LUGGAGE SEARCH WITH COMPUTER-BASED TRAINING

Schwaninger, A.[1], Wales, A.W.J. [1]

1 *University of Zurich, Zurich, Switzerland*

Abstract

A total of 273 luggage search operators from eight airports performed an x-ray competency assessment test on two occasions, with a median of 427 days apart (s.d. = 172.06 days). These screeners are part of an employment setup where recurrent computer-based training and on-the-job assessment are in action, resulting in a large performance ability improvement between the two tests. Using Drury's inspection model (1975) accompanied by signal detection measures (Green and Swets, 1966), it was found that screeners became faster for both search and decision time, as well as becoming more accurate in their detection of threat items. It is concluded that longitudinal benefits are found when screeners are part of an organised training environment that is scientifically based and individually adaptive.

Introduction

Airport baggage security screeners are increasingly being trained with sophisticated software such as X-Ray Tutor that individually adapts its difficulty to compensate for their improvements in detection performance (Schwaninger, 2004b; Koller *et al.*, 2008; Michel *et al*., 2007; Schwaninger *et al.*, 2008). At several airports, screeners are continually being tested on-the-job through a system known as "threat image projection" (TIP; e.g. Schwaninger, 2004a; Hofer and Schwaninger, 2005) that superimposes fictional threat items into actual x-ray luggage images. The basic principles behind the TIP system actually predate airport security screening, where Broadbent (1964) notes "it has often been suggested that in practical situations the efficiency of radar monitors or industrial inspectors could be improved by the insertion of artificial signals interspersed amongst the real ones" (p. 18). With modern advances in computer systems (Schaller, 1997) it is now readily possible to examine the performance of screeners while they work and provide a system of certification relating to performance ability.

Having competent workers is important in any industry, but particularly so in airport security. Since certain visual abilities are essential to become a good x-ray screener (Hardmeier, Hofer and Schwaninger, 2005; Hardmeier & Schwaninger, 2008), pre-employment testing needs to form a staple part of screener employment procedures. Training that is individually adaptive has the benefit of seeking weaknesses and addressing them with increased load into those areas (Schwaninger, 2004b). While screeners need to

demonstrate improvements they also need to fall within accepted working norms which can be deduced from retrospective analyses of TIP data and standalone performance testing. As has been shown in previous research (i.e. Hofer and Schwaninger, 2005), there are wide variations in operator ability but with appropriate training the entire distribution should move as each screener makes their own improvements (Schwaninger, Hofer, & Wetter, 2007). Given that terrorist threats are "both productive and diverse" (Lui *et al.*, 2007, p.301), training systems must be adaptive, recurrent and attempt to predict the mindset of potential terrorists in a changeable climate. Security screening is therefore both demanding and important, requiring research-led training systems capable of quantifying real-life performance.

Detection performance in terms of sensitivity is typically viewed as being a function of hits (correctly identifying an object that contains a threat item), false alarms (incorrectly stating a threat item is present), and in certain cases confidence ratings (Green & Swets, 1966; MacMillan & Creelman, 1991). A vital aspect of performance is the speed in which an operator can perform bag searches while maintaining optimal performance levels, which signal detection theory cannot accommodate as it assumes a fixed sampling interval (Smith, 2000). Thus, Drury's Two-Component Model (TCM; Drury, 1975) can be used to approximate the speed taken to search for items, and the average decision time used. This model is a useful complement to signal detection theory and provides insight into whether search or decision time improves with on-the-job training.

With the goal of applying SDT and TCM estimates to a large dataset, we have collated the data from a longitudinal project to evaluate performance changes in screeners who are part of a battery of training systems developed by the VICOREG research group (Schwaninger, 2004a, 2004b; Schwaninger, Hofer and Wetter, 2007). While it is impossible to eliminate effects generated by work days during the year, confounding variables or other sources of nuisance variance the advantages of being able to gauge actual real-life performance in screeners far outweigh the detractions. The benefits of utilising scientifically based training can therefore be deduced by comparing the two test performances. With a median date interval exceeding one year between the two tests it can be seen whether screener performance changes using a large sample of workers in a standardised test.

Methods

Screeners And Task

273 screeners from eight airports performed the X-Ray Competency Assessment Test (X-Ray CAT, Koller *et al.* 2008) on two separate occasions. The X-Ray CAT is a simulated luggage search task that presents screeners with images of bags of various difficulties with threat items placed at a base rate of 50%. Subjects only have two choices to make ("ok" or "not ok"), and the images are presented onscreen for up to four seconds. Threat categories include variants of knives, guns, improvised explosive devices and other threats such as tazers or sprays.

Analyses

As subjects were selected based on whether they had performed the test on multiple occasions, a within-subjects analysis can be performed that partials out variance caused by inter-subjects factors (Landauer *et al.*, 2008). This is important because of the lack of

experimental control that is found when making deductions based on data that is remotely collected through an autonomous computer network. Several dependent variables can be gleaned from simply evaluating data by subject, reaction time and response. Detection performance, as measured by A' (Pollock and Norman, 1964) and subjective bias (B''; Grier, 1971) are two readily-prepared statistics that are analogues to those used in signal detection theory (Green and Swets, 1966). These statistics provide an overview of performance irrespective of time and makes less assumptions about the distributions of the data than their signal-detection counterparts. All performance metrics have been multiplied by an arbitrary constant to protect security-sensitive data.

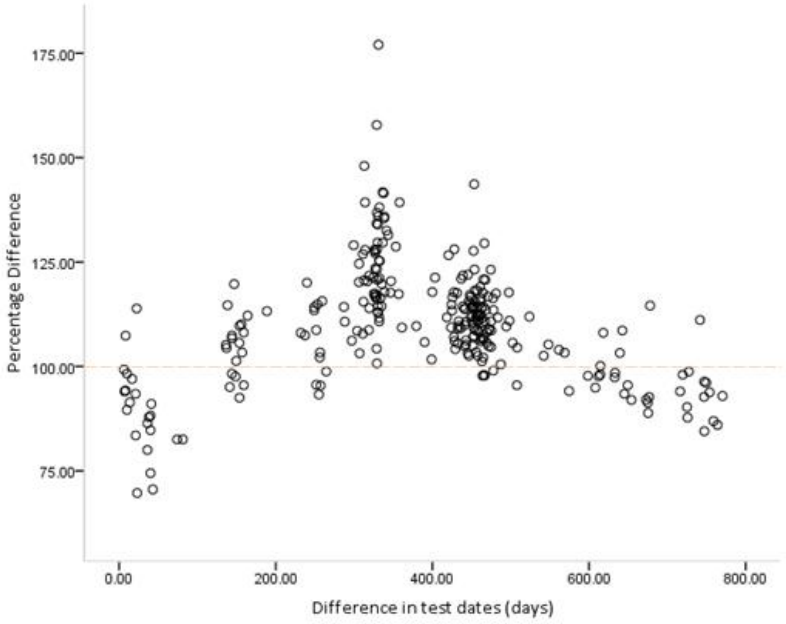
Further to this, an approximation of time taken to search for potential threat items and decision time can be made from these three parameters listed above (Drury, 1994). Several papers in the field of aviation have provided complimentary evidence to support Drury's model when applied to x-ray luggage search data (i.e. Ghylin *et al*., 2006). In brief, Drury's model allows estimation of search (STh for hits and STfa for false alarms) and non-search time (NSTh for hits and NSTfa for false alarms), the latter of which is largely made up of decision time, by stepwise linear correlation of performance as judged by Drury's p(hit/fa) formulation to reaction time by using the model's formulas. Effectively, the quickest reaction time in a trial tends to become the determinant of pure non-search time as it is assumed to mostly be free of the decision component. This process is calculated for each individual and each CAT test, and then averaged to provide overall model parameters (Table 1).

Results

As an indicator of ability, screeners are rated on a scale from 0-12 based on previous experience with adaptive training systems (Michel *et al*., 2008). Before conducting the X-Ray CAT for the first time, their average performance level was at level 0 (s.d. = 0.15), but by the time screeners had begun their second tests, they had risen to level 8 on average (s.d. = 3.65). The mean date difference (mean = 386.03 days, s.d. = 172.06 days) between testing days masks the actuality of the performance date differences, as 215 subjects performed the tests more than 300 days apart while 15 screeners took the test less than a month apart resulting in a wide variance (see Figure 1). Therefore the median date (427 days, s.d. = 172.06; range = 5 – 802) provides a more reliable estimate. Only screeners who had performed the test twice were chosen for analysis, thereby eliminating 244 candidates.

Figure 1 indicates that the subset of screeners who performed the test less than a month apart actually tended to decrease in their performance, whereas the vast majority of screeners who took the test after more than a month's interval performed better than on their first occasion. After 600 days, there is a trend towards diminished performance, but a lack of cases make this inference tenuous. Screeners also get progressively faster for correct decisions for threat and non-threat items, as shown by the scatterplot in Figure 2. Correlations between hit reaction time and probability of a hit were low for the first test ($r = 0.10$, $p > 0.05$) but statistically significant for the second test ($r = 0.30$, $p < 0.05$), while correlations between correct rejection reaction time and probability of a false alarm were high for the first test ($r = 0.38$, $p < 0.05$) but lower for the second test ($r = 0.14$, $p < 0.05$).

Figure 1: Percentage change in A' by test date difference.

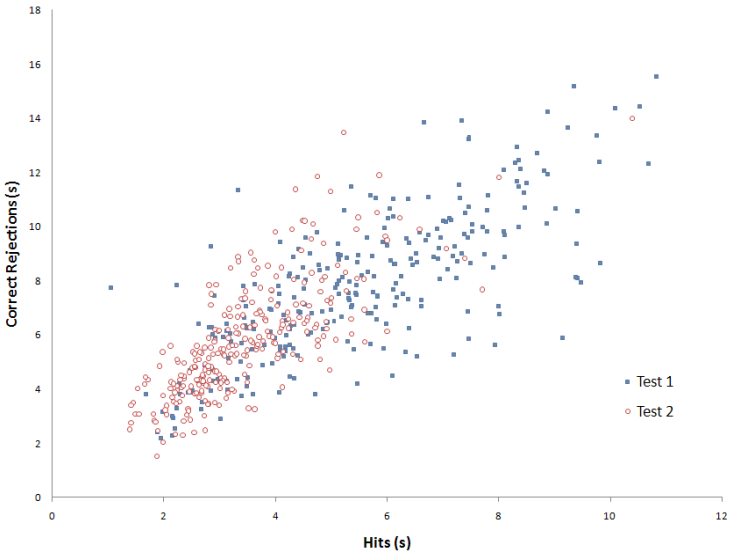


Paired-samples t-tests were performed for each metric and effect sizes were generated according to Cohen (1988) between the two tests. Table 1 shows that every performance indicator achieved experimental significance except for bias, which remained highly insignificant. Figure 3 shows histograms for each performance indicator as measured by pooled z-score values so that the effect sizes of different scales can be directly compared. R^2 values, separated by pipes, indicate the adjusted model fit for the first test and second test respectively. All tests were found to follow an approximately normal distribution, or more specifically a Gaussian exponential curve, with high r values. Although high, the r^2 values indicate that some discrepancies between the actual values and the models exist. By means of comparison, all curves should have the same area underneath them were r to equal 1, but the largest differences are between NSTh 1st and 2nd sessions (165.3 units² to 125.0 units²) and the smallest between STfa 1st and 2nd sessions (147.97 units² to 139.00 units²).

Table 1: Inferential statistics and mean differences (standard deviations)

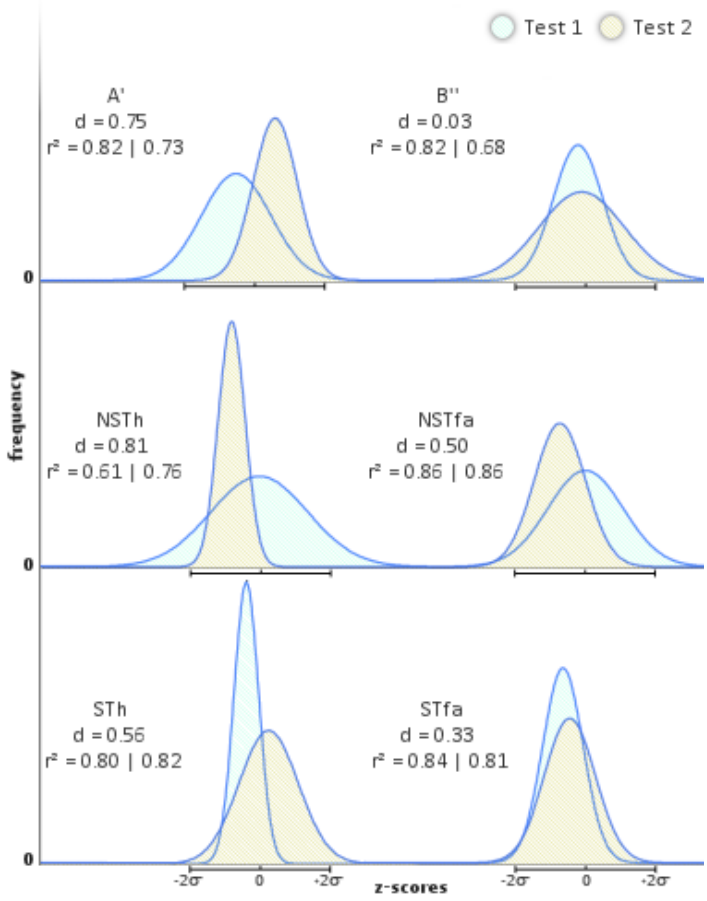
	Test 1	Test 2	Significance Within-Groups
A'	0.793 (0.08)	0.864 (0.06)	$t(273) = -11.667$, $p < 0.05$, $d = 0.75$
B''	0.290 (0.27)	0.118 (0.37)	$t(273) = 0.818$, $p > 0.05$, $d = 0.03$
STh	0.362 (0.19)	0.567 (0.38)	$t(273) = -9.343$, $p < 0.05$, $d = 0.56$
STfa	0.315 (0.15)	0.396 (0.25)	$t(269) = -5.059$, $p < 0.05$, $d = 0.33$
NSTh	2.206 (1.13)	1.249 (0.55)	$t(273) = 13.362$, $p < 0.05$, $d = 0.81$
NSTfa	3.668 (1.87)	2.603 (1.39)	$t(269) = 8.215$, $p < 0.05$, $d = 0.50$

Figure 2: Stopping time policy for Hit RT and Correct Rejection RT



With practice and on-the-job experience, the second tests all show greater kurtosis than the first tests for non-search time and detection, but not for search time. This manifests itself as lower standard deviations (see Table 1 and shown in Figure 3), indicating that a ceiling effect approaches with performance tending towards perfection with practice for non-search time and detection performance, although notably never achieved. All graphs are shown on the same frequency scale and z-score distances are all directly comparable. As expected, the greatest performance effects are seen from those screeners who were poorest in their first tests leading towards the distribution shift evident in Figure 3. Negative t-values in Table 1 show as a directional change in Figure 3, where a decrease in value is desirable (i.e non-search time) in some metrics and an increase in others (i.e. A'). Although the coefficient of search-time becomes larger with practice this is a desirable outcome as it indicates improved performance when taken as a unit of change in Drury's overall model equation.

Figure 3: Z-Score distributional changes between tests 1 and 2. R^2 values indicate the percentage of variance that each model explains for tests 1 and 2 respectively. A', B'', NSTh and NSTfa (Non-Search Time for hits and false alarms), STh and STfa (Search Time for hits and false alarms) are given as well as estimates of effect size (d).



Discussion

A substantial improvement in detection performance is seen in absolute terms (Table 1) and relative terms (Figure 1), as well as improvements in response time (Figure 2), search time and non-search time (Figure 3) with training. Of all the pairwise comparisons made, only bias failed to achieve experimental significance (Table 1), although as a standalone measure bias has no explanatory worth pertaining to performance.

Distributional changes using standardised scores make various dependent variable alterations directly comparable (Figure 3), with accompanying effect sizes providing an empirical determinant of treatment magnitude. If the aim of systemic training is to shift the distributions of performance then the enhanced kurtosis indicated by reduced standard deviations and abscissa drift provides strong evidence that performance is progressively tending towards operator performance limits for detection performance

and non-search time. Although search time performance did improve the increased standard deviations indicate that more training would be required to bring all the screeners up to the same standard as a large number did not improve to the same extent that others did.

The biggest improvements are seen for hits rather than false alarms, indicating that screeners become more adept at spotting threat items in luggage than merely correctly identifying non-threat items that can sometimes be the explanation for A' improvements, which is not explicitly split into threat and non-threat items. As there is no change in bias it cannot be concluded that what changes with training is a tendency to click "bag not ok," which is a strategy some screeners employ thinking that it will improve their detection results. Table 1 and Figure 3 show in conjunction that there are wide differences in screener ability, while the scatterplot in Figure 1 confirms that screeners vary greatly in the speed at which they examine items, with many taking less than four seconds to make their deductions and others taking up to fifteen seconds to correctly reject bags.

Both signal detection and two-component model measures show a marked improvement with training in the competency assessment test when taken over a year apart. Previous studies using a control group have confirmed that CAT performance is not a function of exposure to the task (Koller *et al*., 2008), and the median date interval of 427 days would make memorisation of the task unlikely. Therefore, we can conclude that overall x-ray image interpretation competency does genuinely improve with training, increasing the accuracy and efficiency of actual screeners' work, ultimately increasing the security of airline operations.

References

Broadbent, D.E. 1964, Vigilance, *British Medical Bulletin*, **21**:1, 17-20

Cohen, J. 1988, *Statistical power analysis for the behavioral sciences*, 2nd edition, (Lawrence Erlbaum, Hillsdale, New Jersey)

Cohen, J. 1990, Things I have learned (so far), *American Psychologist*, **45**:12, 1304-1312

Cohen, J. 1994, The earth is round ($p < 0.05$), *American Psychologist*, **49**:12, 997-1003

Drury, C.G. 1994, The speed-accuracy tradeoff in industry, *Ergonomics*, **37**:4, 747-763

Drury, C.G. 1975, Inspection of sheet metal materials: model and data, *Human Factors*, **17**, 257-265

Green, D.M., and Swets, J.A. 1966, *Signal detection theory and psychophysics*, New York: Wiley

Grier, J.B. 1971, Nonparametric indexes for sensitivity and bias: Computing formulas *Psychological Bulletin*, **75**, 424-42

Ghylin, K.M., Drury, C.G., and Schwaninger, A. 2006, Two-component model of security inspection: application and findings, *16th World Congress of Ergonomics, IEA 2006,* Maastricht, The Netherlands, July, 10-14

Hardmeier, D., & Schwaninger, A. 2008, Visual cognition abilities in x-ray screening. *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008,* Fairfax, Virginia, USA, June 1-4, 311-316.

Hardmeier, D., Hofer, F., and Schwaninger, A. 2006, Increased detection performance in airport security screening using the X-Ray ORT as pre-employment assessment tool. *Proceedings of the 2nd International Conference on Research in Air Transportation,*

ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 393-397

Hardmeier D, Hofer F, and 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. *IEEE ICCST Proceedings*, **39**, 189-192

Hofer, F. and Schwaninger, A. 2004, Reliable and valid measures of threat detection performance in X-ray screening, *IEEE ICCST Proceedings*, **38**, 303-308

Hofer F., and Schwaninger, A. 2005, Using threat image projection data for assessing individual screener performance, *WIT Transactions on the Built Environment,* **82**, 417-426

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

Landauer, A.A., Harris, L.J., and Pocock, D.A. 2008, Inter-subject variances as a measure of differences between groups, *Applied Psychology*, **31**:4, 417-422

Lui, X., Gale, A., and Song, T. 2007, Detection of terrorist threats in air passenger luggage: expertise development, *Proceedings of the 41st Annual IEEE International Conference*, 301-306

Michel, S., Koller, S.M., Ruh, M., and Schwaninger, A. 2007, Do "image enhancement" functions really enhance x-ray image interpretation? In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*, 1301-1306

N.A. MacMillan and C.D. Creelman, 1991, Detection theory: A user's guide. Cambridge: University Press.

Pollack, I., and Norman, D.A. 1964, A non-parametric analysis of recognition experiments, *Psychonomic Science*, **1**, 125-126.

Schaller, R.R. 1997, Moore's law: past, present and future, *IEEE Spectrum*, **34**, 53

Schwaninger, A. 2006, Airport security human factors: from the weakest to the strongest link in airport security screening, *The 4th International Aviation Security Technology Symposium*

Schwaninger, A., Hofer, F., and Wetter, O.E. 2007, Adaptive computer-based training increases on the job performance of x-ray screeners, *Proceedings of the 41st Carnahan Conference on Security Technology,* Ottawa, October, 8-11

Schwaninger, A. and Hofer, F. 2004a, 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,* 147-156 (Wessex: WIT Press)

Schwaninger, A. 2004b, Computer based training: a powerful tool to the enhancement of human factors, *Aviation Security International*, 31-36

Schwaninger, A. 2004a, Increasing efficiency in airport security screening. *Proceedings of AVSEC World*, November 3-5, Vancouver, B.C., Canada.

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., and Hay L. 2008, The impact of image based factors and training on threat detection performance in X-ray screening, *Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008,* Fairfax, Virginia, USA, 317-324

Smith, P.L. 2000, Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, 44:3, 408-463.