Evaluation of CBT for increasing threat detection performance in X-ray screening

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

Abstract

The relevance of aviation security has increased dramatically in recent years. Airport security technology has evolved remarkably over the last decade, which is especially evident for state-of-the-art X-ray screening systems. However, such systems will be only as effective as the people who operate them. Recognizing all kinds of prohibited items in X-ray images of passenger bags is a challenging object recognition task. In this article we present a method to measure screener detection performance based on signal detection theory. This method is applied to measure training effects resulting from individually adaptive computer based training (CBT). We have found large increases of detection performance and substantial reductions in response time suggesting that CBT is a very effective tool for increasing effectiveness and efficiency in aviation security screening.

Keywords: X-ray screening, computer based training, aviation security, object recognition, signal detection theory, training effectiveness.

1 Introduction

Working at an aviation security checkpoint is an important and demanding task. This is especially evident for the X-ray screener who has only a few seconds of inspection time to decide whether an X-ray image of a passenger bag is OK or needs to be manually searched (NOT OK). The X-ray screening task can be described as a signal detection situation in which prohibited items represent the signal and the remaining visual information in the X-ray image of the bag represents noise. Screener detection performance can be calculated using sensitivity measures from signal detection theory such as d', Δm or Az (Green & Swets [1], MacMillan & Creelman [2]). These measures are based on hit rates and false alarm rates and are relatively independent of response biases. This is of

The Internet Society: Advances in Learning, Commerce and Security, K. Morgan & M.J. Spector (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-712-4

special importance for measuring detection performance in X-ray screening tests. If the false alarm rate is not considered, it is not possible to distinguish between good detection performance and a “liberal” response bias (Schwaninger [3]). This can be illustrated by a simple example (Schwaninger [4]). Let us assume that two screeners A and B take a test in which 200 X-ray images of passenger bags are shown and half of them contain prohibited items. Both screeners detect threat items in 90% of the cases (hit rate). When the bag contains no prohibited items, screener A judges bags as being NOT OK in only 11% of the cases (false alarm rate). In contrast, screener B has a false alarm rate of 78%. Whereas screener A has a high detection *performance*, screener B achieves a high hit rate *at the expense* of efficiency, which would result in substantially longer waiting lines at the checkpoint. This difference becomes apparent when detection performance is measured by d' = z(H) – z(FA), whereas H denotes the proportion of hits and FA the proportion of false alarms. In the formula z denotes the z-transformation, i.e. H and FA are converted into z-scores (standard-deviation units). In the example mentioned above screener A would have a detection performance of d' = z(0.90) – z(0.11) = 2.51 whereas screener B has a d' = z(0.90) – z(0.78) = 0.51. In other words, detection performance of screener A is almost 5 times higher!

If a CBT system is effective, it should be expected that detection performance d' increases as a result of training. Moreover, if threat items are seen repeatedly during training it could also be expected that they become better represented in visual memory, which could result in faster response times.

However, several methodological considerations need to be taken into account in order to achieve reliable measurements of CBT effectiveness in terms of d' increases and response time decreases. Schwaninger [5] identified three image-based factors that influence X-ray detection performance. Threat items can be more or less superimposed by other objects (effect of superposition).

Second, the number and type of other objects in a bag challenge visual search and recognition processes such that threat items in more “complex” bags usually result in a lower detection probability (effect of bag complexity). Third, when objects are rotated away from the canonical view (Palmer, Rosch, & Chase [6]) they usually become more difficult to recognize (effect of viewpoint). Since these effects have been shown to affect detection performance (Schwaninger [5]), image-based difficulty of X-ray images needs to be carefully controlled in a longitudinal study designed for evaluating CBT effectiveness. Moreover, display duration could be an important variable as well and should therefore be varied.

Finally, only X-ray images of bags and threat items that have not been seen during training should be used in order to measure CBT effectiveness reliably.

These considerations were taken into account using a pilot study, a pre-selection test, and a Latin Square counterbalanced design with four tests of equal difficulty and four groups of screeners with comparable average threat detection ability.

The Internet Society: Advances in Learning, Commerce and Security, K. Morgan & M.J. Spector (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-712-4

2 Method

The CBT used in this study was X-Ray Tutor, an individually adaptive training system based on object recognition and visual cognition (for recent reviews on these topics see Graf, Schwaninger, Wallraven & Bülthoff [7]; Schwaninger [3]).

The main aim of the system is training object recognition by increasing the number and strength of view-based representations in visual memory. X-Ray Tutor is driven by software algorithms that monitor student performance and adjust images presented to provide threat types and bag difficulty needed for the student to learn and progress based on performance deficiencies. For further information see Schwaninger [8] and [9].

2.1 Pilot study

Threat images were created by combining X-ray images of improvised explosive devices (IEDs) with X-ray images of passenger bags using a customized TRX algorithm. In the pilot study, 4000 X-ray images were used, i.e. 2000 harmless bag images and 2000 threat images (125 IEDs * 16 bags per IED). Image difficulty was rated by eight expert screeners of Zurich Airport using a slider control (rating scale 0-100). Inter-rater reliability was estimated by calculating Cronbach's Alpha among raters. Alpha for IEDs (averaged across the 16 X-ray images) was .96. Alpha for X-ray images (without averaging) was .82. Images were ordered by average rated difficulty so that 16 difficulty levels were obtained per IED.

2.2 Training library

In the training system 64 of the 125 IEDs were used. Thus, the training library consisted of 1024 X-ray images containing a bag with an IED (64 IEDs * 16 bag difficulty levels) and 1024 harmless X-ray images showing the same bags without IED.

2.3 Participants

Seventy-two screeners (fifty female) at the age of 23.9 – 63.3 years (*M* = 48.3 years, *SD* = 9.0 years) took part in this study. None of them had received a special IED or computer based training before. These screeners were divided into four groups (group A: *N* = 17, group B: *N* = 18, group C: *N* = 18, group D: *N* = 19) as described in the next paragraph.

2.3.1 Grouping of participants

Prior to training, a pre-selection test was used to distribute the screeners among four groups of equivalent detection performance. To this end, 16 IEDs rated in the pilot study were used, which were not contained in the training library. Each IED X-ray image was combined with a bag image of medium and high difficulty (difficulty level 9 and 15 estimated in the pilot study as described in section 2.1). The entire pre-selection test consisted of 64 trials: 16 IEDs * 2 difficulty levels * 2 trial types (threat images vs. harmless bags). The order of image presentation was counterbalanced across screeners. The task of the screeners was to decide whether the presented luggage contained an IED or not. After each answer, they rated the difficulty of each image on a slider from 0 (very easy) to 100 (very difficult). Statistical analyses showed that the standardized ROC curve is best described by a linear trend, $R^2 = .93$, $p < .001$. Thus, the parametric detection performance measures Δm and d' (Green & Swets **Fehler! Verweisquelle konnte nicht gefunden werden.**, MacMillan & Creelman [2]) could be calculated for each screener and four groups of comparable mean detection performance were created (Table 1). Three of the 72 screeners did not participate in this pre-selection test because they were not available during the period of testing which lasted 32 days (compare the number of screeners in Table 1 and section 2.3).

Table 1: Mean Δm, d' and their correlation, listed separately for each group of screeners. Values in parentheses represent standard deviations. All correlations (r) are significant with p <.01.

Groups of screeners	Δm	d'	r
Group A (*N* = 16)	1.58 (0.76)	1.75 (0.70)	.90
Group B (*N* = 17)	1.98 (2.14)	1.87 (1.02)	.89
Group C (*N* = 18)	1.63 (0.82)	2.07 (0.91)	.88
Group D (*N* = 18)	1.58 (0.73)	1.90 (0.88)	.84

A one factor ANOVA with group as between-subjects factor confirmed that the created groups were comparable in terms of their detection performance. There were no significant differences, neither for the Δm-values, $F(3, 65) = 0.40$, $p = .75$, nor for the d'-values, $F(3, 65) = 0.38$, $p = .77$. For both measures, no post hoc pairwise comparison between groups reached a statistic significant value (all p-values >.25).

2.4 Training blocks

The 64 IEDs used for training were distributed among four blocks of 16 IEDs so that all blocks were of comparable mean difficulty according to the difficulty ratings of the pilot study. One training block consisted of 512 images, i.e. 16 IEDs * 16 bags (difficulty levels) * 2 trial types (threat images vs. harmless bag images).

Standardized measures of difficulty ratings were subjected to one-way repeated measures ANOVA with training block as within-subjects factor. This analysis confirmed that the four training blocks were of equal difficulty. There was no effect of training block, $F(1.72, 12.04) = 0.47$; $MSE = 0.004$; $p = .94$ and pairwise comparisons between the training blocks showed no significant differences for any of the comparisons (all p-values > .25).

During training each IED was first presented in its easiest difficulty level. The order of IEDs was randomized across participants. The difficulty level was increased successively for each screener based on achievements in training (for more information on X-Ray Tutor see Schwaninger [8], [9]). Each training session was automatically terminated after 20 minutes.

The order of training blocks was counterbalanced across the four groups of trainees using a Latin Square design (see Figure 1). Between each training block the detection performance was measured in testing blocks.

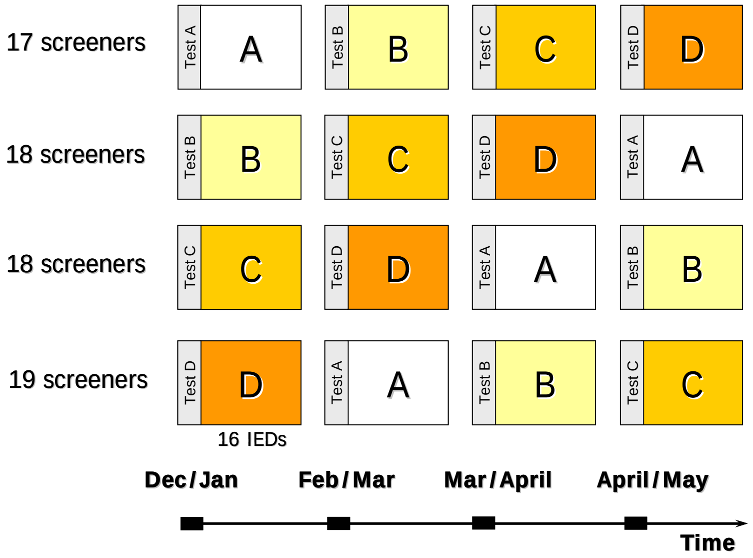

Figure 1: Latin Square design. A-D: Training blocks. Each training block consisted of 16 IEDs in 16 difficulty levels (bag images). Before each training block, a detection test containing the IEDs of the following training block was used to measure the training effects (see text for details). The study was carried out during six months starting December 2002 (see x-axis).

During training, each X-ray image was presented for a maximum of 8 seconds. Trainees had to decide whether the bag contains an IED or not by a clicking on one of two buttons. Subsequently, they judged the difficulty of the X-ray image from 0 (very easy) to 100 (very difficult) using a slider control. Screeners received immediate feedback to their answers. For X-ray images containing an IED the feedback messages were either “Threat detected” (hit) or “Threat missed” (miss). For innocent X-ray images the feedback messages were “False alarm” or “Bag OK” (correct identification of a harmless bag). In addition, an information window could be displayed which showed a labelled X-ray image and photograph of the IED.

2.5 Testing blocks

The participants were always tested using IEDs they had never seen before. This was achieved using testing blocks which contained the IEDs from the next training block (see Figure 1). At test, each IED was presented for 4 and 8 seconds in bags of the two highest image difficulty levels (15 and 16). As in training, each bag was also presented without the IED in order to obtain a better signal detection measure.

As in training, participants judged whether the presented luggage is NOT OK (contained an IED) or OK (contained no IED) and subsequently rated the difficulty of each X-ray image using a slider control.

All four tests consisted of 128 trials: 16 IEDs * 2 display durations * 2 difficulty levels * 2 trial types (threat images vs. harmless bags). The order of presentation was randomized. In contrast to the training blocks, no feedback and no additional information about the IEDs was available during tests.

3 Results

3.1 Descriptive statistics

There was a large increase in detection performance measured by signal detection d' (Figure 2a). In order to assess training effectiveness we calculated % increase values as compared to baseline measurement (first test results), averaging the two display durations. Relative detection performance d' was increased by 70.76% (Figure 2b). This is a remarkable effect if it is taken into account that on average screeners took only 28 training sessions during the six months period (*SD* = 10 TS). Moreover, for a subgroup of 52 screeners, who on average took 31 training sessions (*SD* = 8 TS), the training effect was even more pronounced; relative detection performance was increased by 84.46%!

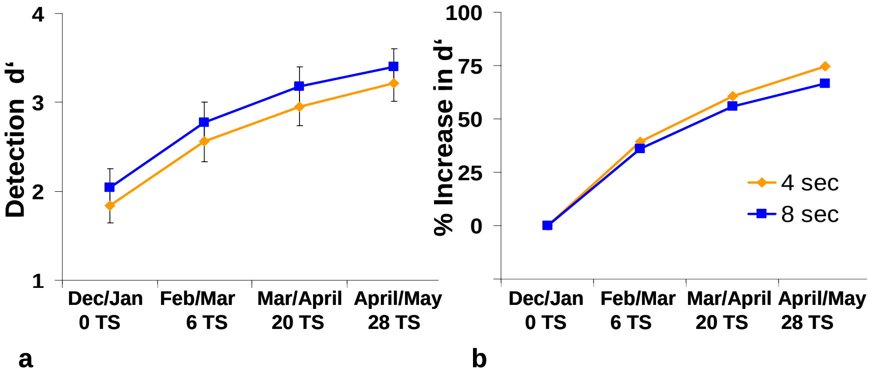

Figure 2: Absolute detection performance (a) and relative increase of detection performance (b) averaged across all 72 screeners. Display durations were 4 and 8 seconds. Error bars represent standard errors. TS = Number of training sessions.

The Internet Society: Advances in Learning, Commerce and Security, K. Morgan & M.J. Spector (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-712-4

3.2 Inferential statistics

Only significant effects are reported using the conventional cut-off of $p < .05$. Effect sizes η^2 are reported and can be judged based on Cohen [10].

3.2.1 Statistical analyses of detection performance d'

Mean detection performance d' at the four test dates of each group are shown separately for the two display durations in Figure 3.

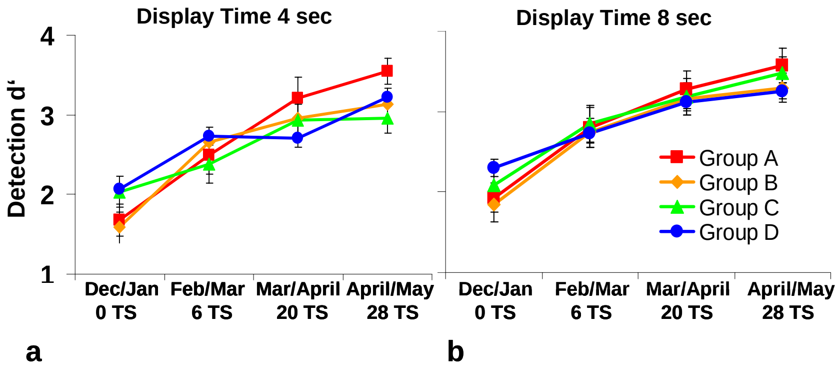

Figure 3: Detection performance d' of the four test dates for the four groups and the two display durations. Error bars represent standard errors. TS = Number of training sessions.

Again, the general training effect can be seen clearly. Detection performance d' of each group increased after each training block. A three-way analysis of variance (ANOVA) with the two within-subjects factors test date and display duration and the between-subjects factor group showed significant effects of test date, *F*(2.81, 190.54) = 124.15, *MSE* = 0.44, $p < .001$, and display duration, *F*(1, 68) = 44.15, *MSE* = 0.14, $p < .001$. With effect sizes of $\eta^2 = .65$ for test date and $\eta^2 = .39$ for display duration. The two-way interaction between test date and group was significant with an effect size of $\eta^2 = .10$, *F*(9, 204) = 2.42, $p < .05$. There was also a significant three-way interaction between test date, display duration and group, with an effect size of $\eta^2 = .09$, *F*(9, 204) = 2.18, $p < .05$.

In short, whereas the groups did not differ in their mean detection performance, there were slight differences in terms of how fast their detection performance increased across training when tested with 4 and 8 seconds of image presentation.

All Bonferroni-corrected pairwise comparisons between different test dates were significant confirming training effectiveness for the whole period of six months (all *p*-values < .001, with the exception of the comparison between test dates 3 (Mar/April) and 4 (April/May) with the *p*-value < .01).

3.2.2 Statistical analyses of reaction times

Figure 4 (top) shows reaction times for bags containing an IED separately for the four screener groups and the two display durations of 4 and 8 sec. Similarly, Figure 4 (bottom) depicts reaction times for harmless bags.

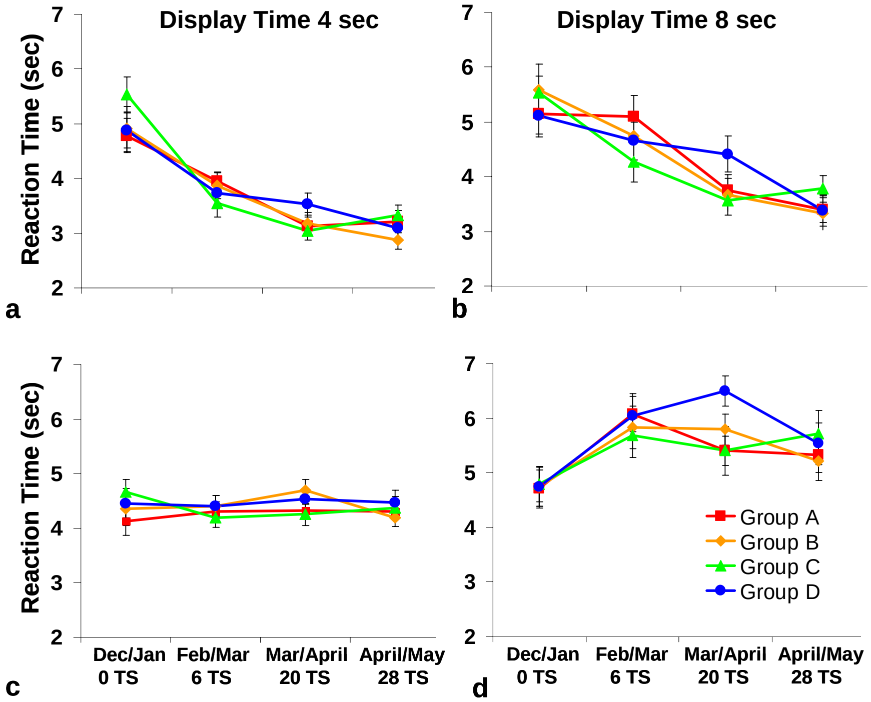

Figure 4: Reaction times of the four test dates for the four groups and the two display durations of 4 and 8 seconds. Top: Reaction times for bags containing an IED for 4 seconds (a) and 8 seconds (b). A clear decrease of reaction time was observed. Bottom: Reaction times for harmless bags for 4 seconds (c) and 8 seconds (d). Error bars represent standard errors. TS = Number of training sessions.

Only reaction times of correct responses were analysed. For threat images (bags with IED), a three-way ANOVA with test date and display duration as within-subjects factors and group as between-subjects variable showed a significant effect of test date, $F(1.71, 116.44) = 52.54$, $MSE = 1961145.60$, $p < .001$. The effect size was $\eta^2 = .44$. There was also a main effect of display duration ($\eta^2 = .57$), $F(1, 68) = 91.26$, $MSE = 476870.94$, $p < .001$. The two-way interaction between test date and display duration was also significant with $\eta^2 = .10$, $F(1.82, 123.72) = 7.30$, $MSE = 636920.93$, $p < .01$.

Bonferroni-corrected pairwise comparisons revealed significant differences for all comparisons between the reaction times of test dates (all p-values < .001, except for the comparison between test date 3 (Mar/April) and 4 (April/May) with $p < .05$). In short, response times for threat images decreased across training for all participant groups to a similar extend.

The same three-way ANOVA was used to analyze reaction times for harmless bags. Again, there was a main effect of test date, $F(1.99, 5.96) = 5.17$, $MSE = 2752802.22$, $p < .01$, with an effect size of $\eta^2 = .07$, which is much smaller than observed for threat images (see above). There was also a main effect of display duration ($\eta^2 = .74$), $F(1, 68) = 190.02$, $MSE = 901246.42$, $p < .001$, and a significant two-way interaction between test date and display duration ($\eta^2 = .28$), $F(2.49, 169.01) = 26.58$, $MSE = 463610.22.$, $p < .001$.

Except for the comparisons between test date 1 (Dec/Jan) and 2 (Feb/Mar) and 1 (Dec/Jan) and 3 (p-values < .05) no Bonferroni-corrected pairwise comparison revealed significant differences between the reaction times of different test dates. Thus, in contrast to response times for threat images, there was no substantial reduction of response times for X-ray images of harmless bags.

4 Discussion

The aim of this study was to develop a method in order to evaluate effectiveness of CBT for increasing threat detection performance in X-ray screening. Signal detection measures take the hit rate and the false alarm rate into account and provide more valid and reliable measures of detection performance than the hit rate alone (Schwaninger [3], [4]). ROC linearity analyses revealed that parametric measures d' and Δm can be computed (Green & Swets [1], MacMillan & Creelman [2]). The two measures were strongly correlated as revealed in a pre-selection test that was used to create four groups of screeners with equivalent detection performance. Four tests of equal X-ray image difficulty were created based on difficulty ratings by eight expert screeners. Inter-rater reliability was sufficient suggesting that difficulty ratings could serve as estimates of objective detection performance.

A Latin Square counterbalanced design was used to measure CBT effectiveness in a longitudinal study of six months during which each screener took about 2 training sessions of 20 minutes per week. None of them had received a special IED or computer based training before. Only new X-ray images were used in the four tests in order to measure training effectiveness in terms of generalisation to new threat items. Remarkable increases in detection performance d' were observed. Relative increase in detection performance d' as compared to the first test was 71% after an average of 28 training sessions during the six months period. For a subgroup of 52 screeners, who on average took 31 training sessions, relative increase in detection performance d' was even higher, i.e. 84%.

Image display duration at test had a small but reliable effect. When images were displayed for 4 seconds, performance was a bit worse than for 8 second display durations. This effect remained relatively stable across the four tests conducted during the six months period.

More interesting was the decrease in response time for detecting threat items as a result of training. This finding is consistent with the assumption that individually adaptive CBT increases the number and strength of view-based representations of threat items in visual memory and thus could explain a reduction of detection time. Since no response time reduction was observed for harmless bag images, the learning effect indeed seems to be more related to visual memory representations than to increased general visual processing capacities.

In sum, the results of this study suggest that individually adaptive CBT is a powerful tool for increasing threat detection performance in X-ray screening of passenger bags.

Acknowledgements

We are thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study. This research was financially supported by Zurich Airport Unique.

References

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

[2] MacMillan, N. A. & Creelman, C. D., *Detection theory: A user's guide,* University Press: Cambridge, 1991.

[3] Schwaninger, A., Object recognition and signal detection. *Praxisfelder der Wahrnehmungspsychologie*, eds. B. Kersten & M.T. Groner, Huber: Bern, in press.

[4] Schwaninger, A., Evaluation and selection of airport security screeners. *AIRPORT,* **02**, pp. 14-15, 2003.

[5] Schwaninger, A., Reliable measurements of threat detection. *AIRPORT*, **01**, pp. 22-23, 2003.

[6] Palmer, S.E., Rosch, E. & Chase, P., Canonical perspective and the perception of objects. *Attention and Performance IX*, eds. J. Long & A. Baddeley, Erlbaum: Hillsdale, N.J., pp. 135-151, 1981.

[7] Graf, M., Schwaninger, A., Wallraven, C. & Bülthoff, H.H., Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme, Cognitive Vision Systems – CogVis; IST-2000-29375*, 2002.

[8] Schwaninger, A., Training of airport security screeners. *AIRPORT,* **05**, pp. 11-13, 2003.

[9] Schwaninger, A., Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International*, in press.

[10]Cohen, J., *Statistical power analysis for the behavioral sciences,* Erlbaum: Hillsdale, New York, 1988.