Reliable Measurements of Threat Detection

Airport security technology has evolved remarkably over the last decades, which is especially evident when state-of-the art detection systems are concerned. In many cases however, such systems will be only as effective as the personnel who operates them. Indeed, the importance of human factors has gained much attention recently and it has become clear that effective selection, evaluation and training of airport security personnel are crucial factors for increasing airport security and efficiency. Since June 2000 researchers from the University of Zurich have conducted several studies at Zurich Airport, which revealed important insights for the following issues: (1) reliable measurements of threat detection, (2) screener evaluation and selection, (3) training of screeners, and (4) pre-employment assessment. The scientific studies were conducted in close collaboration with Zurich State Police, Airport Division and were funded by Zurich Airport. In AIRPORT 3/2002 an overview of these studies was presented (page 20-21). In this article the first topic is discussed in more detail, i.e. how threat detection can be measured reliably.

Measuring threat detection: an example

Evaluation and certification are important topics in this year's airport security agendas, in Europe as well as in the US. Reliable measures of threat detection are essential in order to determine the performance of individuals, companies and airports. But measuring threat detection is not so simple. Consider the following example: two screeners take a threat detection test in which 200 x-ray images are shown and half of them contain a threat item (e.g. a gun, knife, dangerous good or bomb). Both screeners achieve the same hit rate, i.e. they both detect threat items in about 90 % of the cases. Intuitively, one is tempted to conclude that both screeners are comparable in terms of their threat detection performance. But unfortunately, the hit rate alone does not tell you much. The reason is easy to understand: at test, a participant could simply judge each x-ray image as being not ok and thereby achieve a hit rate of 90 % or more. In order to find out whether a high hit rate just reflects such a "liberal" response bias and not a good detection performance we have to consider the false alarm rate, too. In the current example 200 images were used at test and only half of them contained a threat item. The hit rate is the percentage of times a bag was judged to be not ok of the 100 bags that did contain a threat. The false alarm rate is related to the other 100 bags, which did not contain forbidden objects, i.e. the percentage of times a harmless bag was scored as being not ok. As you can see in Figure 1, screener A and B have the same hit rate of 90 % but they differ remarkably in their false alarm rate. Instead of having a good detection performance, screener B just has a strong tendency to judge any bag as being not ok, which is indicated by the high false alarm rate of 78 %. Such a behaviour would result in long waiting lines at the checkpoint, trading efficiency for security. In contrast, a good screener is able to detect threats very well (high hit rate) and will also report reliably when a bag is ok (low false alarm rate). Screener A in Figure 1 is such a screener, this person achieved a hit rate of 90 % while having a false alarm rate of only 11 %. A high level of security as well as efficiency are the results. But how can we identify such screeners? Which are valid indicators of detection performance?

Signal detection measures

Signal detection theory provides methods for calculating detection measures that are independent of subjective response biases and thereby provide valid indicators of threat detection performance. This allows identifying screeners who can detect forbidden objects very well and at the same time are good in correctly identifying harmless bags. The curves in Figure 1 are called receiver operating characteristics, or simply ROC curves. They represent a graphic description of how the hit rate of an observer changes as a function of changes in the false alarm rate. Each ROC

AIRPORT 01/2003, S. 22-23, GIT VERLAG GmbH & Co. KG, Darmstadt, Germany, www.gitverlag.com/go/airport

curve is related to a different detection performance, which is indicated by the measure d' (or sensitivity). This measure is calculated by the formula d' = z(hit rate) – z(false alarm rate) and has to do with the distance of an ROC curve from the diagonal. In the formula z denotes the z-transformation, i.e. the hit rate and the false alarm rate are converted into z-scores (standard deviation units). For example Person B in Figure 1 has a high hit rate but also a high false alarm rate. Consequently, d' = z(hit rate) – z(false alarm rate) is relatively small and the person is on a ROC curve with a low d' value, namely d' = 0.5. In other words, this person has a very low detection performance and achieved a high hit rate just by judging most bags as being not ok. Security is achieved at the at the expense of efficiency. In contrast, what you are looking for in order to increase airport security performance is someone like person A. This screener has a high hit rate and a low false alarm rate. Security is achieved without sacrificing efficiency, which is reflected by a high d' value. As you can see in Figure 1, person A is on the ROC curve which corresponds to d' = 2.5, indicating a much better detection performance than the one of screener B.

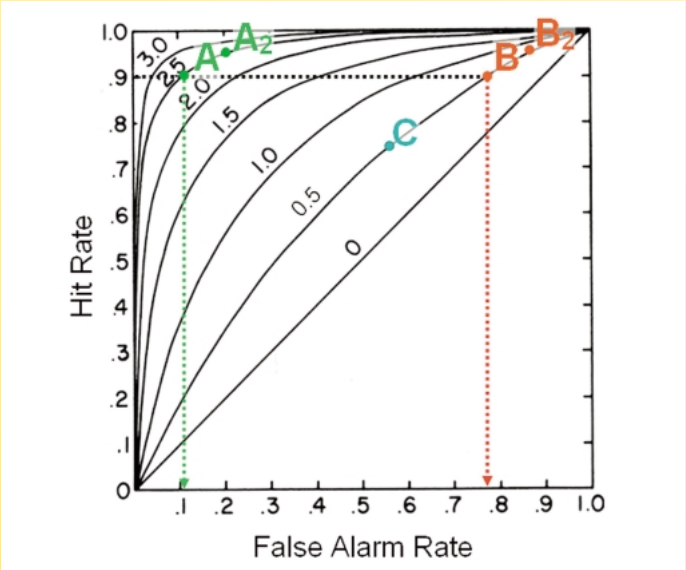

Fig.: Seven ROC curves, each of which correspond to a different d' value (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0). The higher the d' value, the higher the detection performance. For example screener A has a detection performance of d' = 2.5, which represents a much better detection performance than screener B and C with d' = 0.5. A2 is the same screener as A but with a more liberal response bias. The same is true for B2 and B.

A very useful property of signal detection theory is the fact that the detection measure d' is independent of subjective response biases. This is very important, because response biases influence the hit rate and they are dependent on a variety of factors such as the subjective probability of occurrence of certain threat objects, expected costs and benefits of the response, personality, and job motivation. For example the subjective probability that weapons, knifes and other forbidden objects could occur in cabin baggage has increased immediately after September 11, 2001. Of course detection performance (d') could not change from one day to the other. Person A in Figure 1 remained a better screener than person B. What changed immediately is the response bias. Most screeners shifted their response bias towards responding more often with not ok, which is illustrated in Figure 1 by the changed positions A2 and B2. Note that the detection performance d' remained the same, both screeners remained on their own ROC curve. Subjective response biases also differ from one person to another. For example screener C in Figure 1 has a more "conservative" response bias than person B, which results in a lower false alarm rate. But because the hit rate is also much smaller, screener C has the same low detection performance as screener B (note that in Figure 1 both screeners are located on the ROC curve corresponding to a relatively low d' value of 0.5). A second reason for the shift in response bias as a reaction to September 11, 2001 is the fact that the subjective costs produced by long waiting lines became relatively small when compared to the subjective costs of missing a threat object, and this was realized by anybody suddenly, including passengers. Last but not least, hand searching bags after screening them is time consuming and can be stressful if passengers do not cooperate well. Therefore, the personality of the screener and its job motivation are other factors, which can influence subjective response biases. Whereas response biases can change rapidly and can also be influenced by external factors like screener incentive programs, an increase of true detection performance (d') is more difficult to achieve and requires training. Depending on the threat type such training needs to be more or less intensive, an aspect which is of special importance for training bomb detection, which will be discussed in more detail in a separate article on topic (3) training of screeners.

Last but not least it should be mentioned that signal detection theory is often used to measure the detection performance of machines. Simply imagine that the letters A, B, and C in Figure 1 were automatic explosive detection systems from different vendors. Because machine A has the highest detection performance (d') you would invest in this technology. Especially if you knew that the position on the ROC curve ("response bias") can be changed by adjusting a detection threshold.

Adrian Schwaninger
University of Zurich
aschwan@
allgpsy.unizh.ch