APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **19**: 1157–1175 (2005)
Published online 13 September 2005 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1152

Improving Police Decision Making: General Principles and Practical Applications of Receiver Operating Characteristic Analysis

CRAIG BENNELL*

Carleton University, Canada

SUMMARY

Receiver operating characteristic (ROC) analysis is a widely used and accepted method for improving decision making performance across a range of diagnostic settings. The goal of this paper is to demonstrate how ROC analysis can be used to improve the quality of decisions made routinely in a policing context. To begin, I discuss the general principles underlying the ROC approach and demonstrate how one can conduct the analysis. Several practical applications of ROC analysis are then presented by drawing on a number of policing tasks where the procedure has been used already (bite mark identification and linking serial crimes) or where it could be used in the future (statement validity assessment and determining the veracity of suicide notes). I conclude by considering briefly some of the potential difficulties that may be encountered when using ROC analysis in the policing context and offer some possible solutions to these problems. Copyright © 2005 John Wiley & Sons, Ltd.

Police officers routinely make important decisions, many of which affect people's lives. On any given day, an officer might have to decide whether a suspect is being deceptive, whether a negotiation will end in failure, whether a bite mark was made by an adult, and so on (Taylor, Bennell, & Snook, 2002). All of these sorts of decisions are referred to as diagnostic tasks in other settings (Swets, 1996). They each require a police officer (or some other decision maker) to use the available evidence to decide between one of two alternatives. In addition, if published research is anything to go by, they are rarely, if ever, easy decisions to make (e.g. Ekman & O'Sullivan, 1991; Taylor, 2002; Whittaker, Brickley, & Evans, 1998).

According to a number of leading diagnosticians, there are two primary reasons why these tasks are difficult (Swets, Dawes, & Monahan, 2000a). The first reason is that, in the majority of cases, many different sources of evidence can be drawn on to make the decision, but only some of that evidence may actually be useful. For example, the degree to which a suspect fidgets during an interrogation might not indicate whether that suspect is being deceptive, however, the pitch of their voice might be a useful predictor. To the extent that a decision maker relies on evidence that does not reliably predict the event of

*Correspondence to: Craig Bennell, Department of Psychology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada. E-mail: cbennell@connect.carleton.ca

interest, inaccurate decisions will be made. The problem here is that accurate predictors are not always easy to identify (Swets, 1988).

The second reason why these tasks are difficult is that they typically require a judgment to be made as to how much evidence needs to be observed before deciding that the event of interest has occurred (or will occur in the future). For example, when attempting to detect deception, if a police officer were to use the suspect's voice pitch to indicate deception, that officer would have to make a judgment as to how high the pitch must be to warrant this decision. Where the police officer sets this decision threshold will have a significant impact on the utility of their decisions. The problem here is that it is not always obvious where this threshold should be placed in order to achieve optimal decision making performance (Swets, 1992).

Fortunately for the police decision maker, a general analytic procedure exists that can assist with these two problems, regardless of what the diagnostic task is. However, despite its rapidly growing popularity in other diagnostic settings (such as radiology, psychology, and engineering), the procedure has yet to find its way into the police setting (with some notable exceptions to be discussed later in this paper). The procedure is known as receiver operating characteristic (ROC) analysis (Swets, 1996), and the goal of this paper is to demonstrate how it can be used to improve the quality of decisions made in a policing context.

To accomplish this objective, I will begin by reviewing the general principles underlying the ROC procedure and demonstrate how to conduct the analysis. Second, some practical applications of ROC analysis will be presented by drawing on a number of policing tasks where the procedure has been used already, or where it could be used in the future. Third, a number of methodological difficulties that may be encountered when using ROC analysis in the policing context will be considered briefly, as will some potential solutions to these problems.

GENERAL PRINCIPLES OF ROC ANALYSIS

Each of the various policing tasks presented in the opening paragraph can be referred to as a two-alternative task (Swets et al., 2000a). Any time a decision maker is faced with such a task, there are two possible realities that need to be considered and two predictions that can be made. The event of interest can either be present or absent and the prediction can be that the event is present or absent. In an interrogation setting, for example, a suspect is either being deceptive or non-deceptive, and the police officer must predict whether the suspect is deceptive or not.

Possible decision outcomes in two-alternative tasks

The resulting combinations of these realities and predictions mean that four decision outcomes are possible when faced with such tasks. These outcomes can be presented in the form of a contingency table, as illustrated in Table 1. In this table, two of the outcomes are correct (hits and correct rejections) while the other two are incorrect (false alarms and misses). As will be demonstrated, the frequencies of these decision outcomes (or, more precisely, their probabilities) can be used to examine, and to a large extent solve, the two problems described briefly above (Swets et al., 2000a). The problem of selecting effective evidence as a way to improve decision making accuracy will be dealt with first. Following

Table 1. Possible decision outcomes in a two-alternative task

		Reality	
		Present	Absent
Prediction			
	Present	Hit	False alarm
	Absent	Miss	Correct rejection

this, setting appropriate decision thresholds as a way to improve decision making utility will be discussed.

Improving the accuracy of police decision making

When dealing with two-alternative tasks, decision making accuracy can be improved either by increasing the relative frequency of correct decisions or by decreasing the relative frequency of incorrect decisions (Swets et al., 2000a). To accomplish this, the decision maker must base their decisions on evidence that reliably distinguishes between the events of interest. In the ideal situation, this evidence will always be present when the event occurs but never present when it does not. When this happens, there is the potential for decision making accuracy to be extremely high. In the majority of cases, however, this will be rare, especially when behavioural evidence is relied upon (as will often be the case in the policing context). This is because behavioural evidence is often ambiguous in nature, in the sense that it can occur to a degree when an event is present and absent (Swets et al., 2000a).

To illustrate the situation, one can visualize two probability distributions that correspond to a set of diagnostic alternatives (Swets et al., 2000a). For example, Figure 1a represents hypothetical probability distributions obtained from deceptive and non-deceptive suspects when the degree to which they fidget in an interrogation is measured (in fidgets per minute). Figure 1b represents the same information based on the suspect's voice pitch (measured in hertz). In each of these figures, values along the x-axis occur for each diagnostic alternative (deceptive or non-deceptive) with a probability equal to the height of the distribution (Swets et al., 2000a).

As indicated in Figure 1a, the degree to which a suspect fidgets increases when a suspect is deceptive. In this case, the range of fidgets for a deceptive suspect is 6 to 14 fidgets per min, whereas the range of fidgets for a non-deceptive suspect is 5 to 13 fidgets per min. Thus, any time a suspect fidgets less than 6 times per min or more than 13 times per min a police officer can predict, in a fairly accurate fashion, whether the suspect is being deceptive. However, problems will occur when suspects fidget between 6 and 13 times per min because any values within this range are ambiguous.

The situation in Figure 1b is somewhat different. Here, the pitch of a suspect's voice also increases when the suspect is deceptive, but fewer values are ambiguous. As a result, a suspect's voice pitch is likely to be a more accurate predictor of deception than fidgets per minute and, given the choice between the two predictors, voice pitch should be relied upon more often. Indeed, the degree of discrimination accuracy that can be achieved in two-alternative tasks when relying on a particular source of evidence depends largely on the

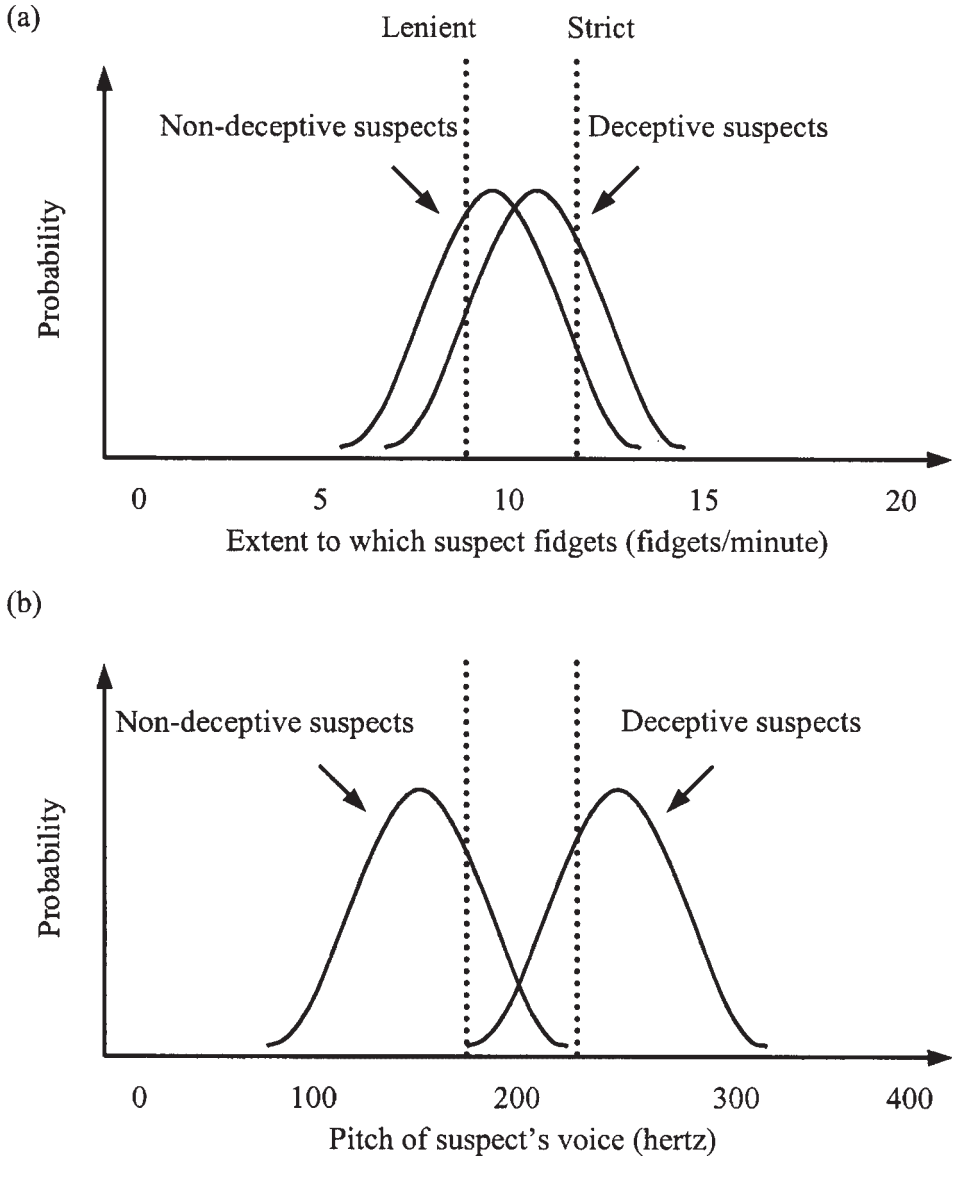

Figure 1. Hypothetical probability distributions associated with (a) fidgets per minute and (b) voice pitch, for deceptive and non-deceptive suspects. A lenient decision threshold (where a smaller amount of evidence is required in order to predict that a suspect is being deceptive) and a strict decision threshold (where a larger amount of evidence is required in order to predict that a suspect is being deceptive) are also displayed

degree of overlap that exists between the corresponding probability distributions (Swets et al., 2000a).

A primary goal for the police decision maker, therefore, is to identify and use evidence that is as unambiguous as possible for the various tasks they encounter. ROC analysis can assist with this task by providing reliable and valid measures of distribution overlap. I will demonstrate how in a moment.

Improving the utility of police decision making

If police decision makers are fortunate enough to have access to evidence that is always present when an event occurs but never when it does not, the problem of setting an appropriate decision threshold will not arise. However, when this is not the case, decision makers will have to set an appropriate threshold along the continuum of evidence. This threshold refers to a cut-off point, whereby any values above that point (or below it depending on the evidence) will result in a prediction that the event has occurred (Swets, 1992). Setting an appropriate threshold can enhance the utility of the decisions that are made (Swets et al., 2000a).

To be clear, decision making accuracy and decision making utility are not the same thing. Whereas increases to accuracy enhance '. . . the odds that any given decision will be the correct one', setting an appropriate decision threshold improves the utility of decisions, '. . . ensuring that the number of [hits] does not come at the cost of an unreasonable number of [false alarms]' (Swets, Dawes, & Monahan, 2000b, p. 82). For example, a police officer could use the same piece of evidence (e.g. voice pitch) to detect deception in serial killers and serial burglars. That evidence may be equally accurate for both purposes but it seems unlikely that a similar decision threshold would be used across the two settings, since the utility of the decisions that result from that threshold would not be the same. This is because the utility of diagnostic decisions depends on the circumstances of the situation.

Two issues in particular should be evaluated when determining the utility of a decision threshold (Swets, 1992). First, the prevalence of each diagnostic alternative in the target population needs to be taken into account. For example, killers may be much more likely to exhibit deception than burglars due to the more serious consequences that will face them if they are charged. Second, the costs and benefits associated with incorrect and correct decision outcomes, respectively, need to be considered. For example, the benefits of correctly identifying a deceptive killer may far outweigh the benefits of correctly identifying a deceptive burglar due to the extremely violent and interpersonal nature of their crimes. As a general rule, Swets et al. (2000b) suggest that

> . . . a high prevalence of a problem in a population or a large benefit associated with finding true cases generally argues for a lenient threshold [e.g., requiring a lower voice pitch to make a prediction of deception] . . . a low prevalence or a high cost of false alarms generally calls for a strict threshold [e.g., requiring a higher voice pitch to make a prediction of deception] (p. 84).

The impact that a decision threshold can have on decision making performance can be better appreciated when the threshold is represented graphically (Swets et al., 2000a). In Figure 1a, two different decision thresholds (a lenient one and a strict one) have been set along the x-axis, with each referring to a different number of fidgets per minute. The same two thresholds have been set in Figure 1b, where they refer to different voice pitches. In any case, if an observed value falls above the threshold in use, a prediction would be made that the suspect is deceptive.

From these figures, one can see that the position of a threshold determines directly the probability of making correct and incorrect decisions (Swets et al., 2000a). More specifically, the probability of a hit (pH) is equal to the area under the deceptive distribution to the right of the threshold and the probability of a miss (pM) is equal to the area under the same distribution to the left of the threshold. Because the total area

under the deceptive distribution is equal to 1, pH and pM are complementary (i.e. if you know one value you can calculate the other). Similarly, the probability of a false alarm (pFA) is equal to the area under the non-deceptive distribution to the right of the threshold and the probability of a correct rejection (pCR) is equal to the area under the same distribution to the left of the threshold (making pFA and pCR complementary).

These figures underscore the fact that decision thresholds can be placed anywhere along the continuum of evidence and that, as these thresholds are varied, the pH/pFA ratio will also vary in a predictable way (Swets, 1992). Specifically, it can be seen that if the decision threshold is made more lenient in order to increase pH, pFA will also increase. Likewise, if the decision threshold is made stricter in order to decrease pFA, pH will also decrease. Thus, decision making performance can vary drastically as a result of adopting different decision thresholds, even when relying on a source of evidence that has a constant level of accuracy.

These figures also draw attention to the fact that the degree of distribution overlap has a significant impact on the pH/pFA ratio that can be achieved at a particular decision threshold (Swets et al., 2000a). In other words, when using the same sort of decision threshold (e.g. a strict one) across two sources of evidence that differ in their discriminatory power, a greater pH/pFA ratio will occur for the more discriminatory evidence.

Another goal for the police decision maker, therefore, is to identify decision thresholds that result in the desired balance between pH and pFA for the various tasks they encounter, and to understand how the level of distribution overlap associated with a source of evidence affects this balance. ROC analysis can also assist with this task.

CONDUCTING ROC ANALYSIS

To briefly summarize, in order to effectively improve two-alternative decision making, it is necessary to have a procedure that can quantify both decision making accuracy and decision making utility. However, these two aspects of decision making performance must be quantified independently so as to avoid being biased by one another. For example, imagine a study where the researcher wishes to compare decision making performance across different decision makers, or across different diagnostic systems. If two police officers were observed to make drastically different decisions when faced with the same diagnostic task, how would one know, without having these independent measures, whether the officers differ in terms of their discrimination accuracy or whether they are simply adopting different thresholds?

Both aspects of decision making performance can be examined independently using ROC analysis. Swets (1996) has demonstrated that this can be accomplished by plotting the values of pH and pFA on a ROC graph as the decision threshold is systematically varied. As shown in Figure 2, a ROC graph is a plot of pH on the y-axis and pFA on the x-axis, when these values have been calculated across decision thresholds that range from lenient to strict.

Calculating pH and pFA

These probability values can be estimated from the cell frequencies in Table 1 (Swets, 1988). For example, for a given source of evidence and a particular decision threshold, pH can be estimated by dividing the frequency of ‘present’ predictions made when the event

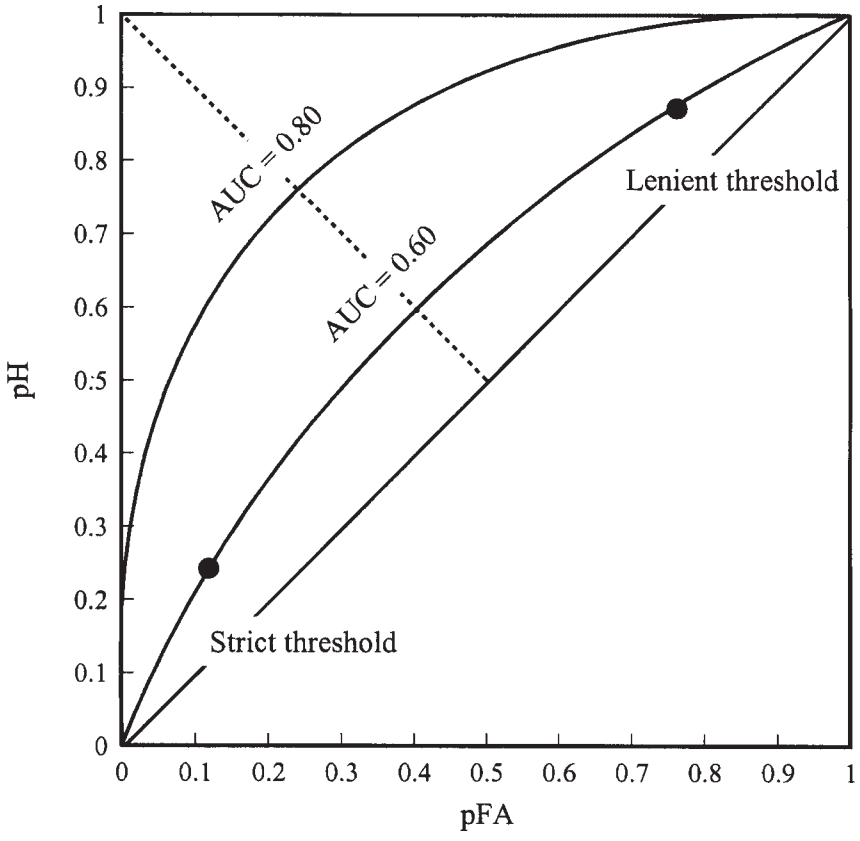

Figure 2. A ROC graph indicating two ROC curves that vary in the level discrimination accuracy they represent. A lenient decision threshold and a strict decision threshold are also displayed

was present (hits) by the total number of predictions made when the event was present (hits + misses). Likewise, pFA can be estimated by dividing the frequency of 'present' predictions made when the event was absent (false alarms) by the total number of predictions made when the event was absent (false alarms + correct rejections). As mentioned above, pH and pFA provide all the information needed to summarize decision making performance in two-alternative tasks, because once these values are known, the other two probabilities are also known (Swets, 1988). More specifically, pM = 1-pH and pCR = 1-pFA.

Constructing empirical ROC curves

When the values of pH and pFA are plotted in this fashion, the result is a concave curve, known as a ROC curve, usually made up of multiple data points, known as ROC points. A ROC curve begins in the top right-hand corner of the graph and ends in the bottom left-hand corner, with each ROC point representing the pH/pFA ratio that results when using a different decision threshold.

ROC curves can be generated empirically in a number of ways. The approach taken will depend on the question being asked. For instance, when examining decisions made by

humans, one common method is to use the yes-no procedure (Egan, 1975). Using the interrogation setting as an example, a police officer is presented with suspects, one at a time, and they have to decide, based on some sort of evidence, whether the suspects are being deceptive or not. After completing one set of trials, the police officer is then asked to change his or her decision threshold before making decisions on another group of suspects (or the police officer could be induced to change their threshold in some way). At each of the various thresholds adopted, pH and pFA are calculated and plotted on a ROC graph. To construct another ROC curve, the researcher uses the same procedure but asks the police officer to base his or her decisions on a different source of evidence. Alternatively, the researcher could test a different group of decision makers or examine the impact of various situational factors, and proceed to construct ROC curves that correspond to these conditions.

Alternatively, the more popular rating procedure could be used, where instead of providing a simple yes-no response, the police officer provides a quantitative rating of how likely he or she thinks it is that each suspect is being deceptive (e.g. on a scale from 1 to 10) (Egan, 1975). Using this procedure, the police officer is adopting multiple decision thresholds simultaneously. Values of pH and pFA are then calculated at various decision thresholds set along this scale, for decisions that exceed each threshold (Swets et al., 2000a). For example, the researcher may set thresholds at >1, >2, >3, and so on, resulting in the generation of an empirical ROC curve. In the same way as described above, the researcher could then construct different ROC curves to answer any particular questions that are of interest to him or her.

When the decision maker is not human, as is often the case nowadays, a variation of the rating procedure can be used. For example, in the interrogation setting, a statistical procedure such as logistic regression analysis could be used to provide ratings of deception. In some cases, one may be specifically interested in comparing the performance of the actuarial tool with a human decision maker on the same task to determine if one is more accurate than the other. Or, one may be interested in using the actuarial tool to identify the optimal combination of evidence that can be used to make a decision. Whatever the goal, the procedure for constructing a ROC curve using this approach is the same as for the rating procedure described above. Various decision thresholds are set along the output of the actuarial tool, and values of pH and pFA are calculated for outputs that exceed each successive threshold.

Identifying appropriate decision thresholds

As illustrated in Figure 2, ROC points falling along the top of the ROC curve occur when the decision threshold adopted is lenient (Swets et al., 2000a). In these cases, not much evidence is required to make a decision that the event of interest has occurred. In contrast, ROC points falling along the bottom of the ROC curve occur when the decision threshold adopted is strict (Swets et al., 2000a). In these cases, a lot of evidence is required to make a decision that the event of interest has occurred. Because a decision threshold can be placed anywhere along the continuum of evidence, it is desirable to have a procedure to identify appropriate thresholds. Using ROC analysis, this can be done in numerous ways (Greiner, Pfeiffer, & Smith, 2000).

One of the most effective methods is to consider the prevalence rates of the diagnostic alternatives in the target population and the costs and benefits of making incorrect and correct decisions (Peterson, Birdshall, & Fox, 1954). In the interrogation example, for

instance, we can denote the probability of encountering a deceptive suspect as $p_{deceptive}$ and the probability of encountering a non-deceptive suspect as $p_{non\text{-}deceptive}$. Furthermore, we can denote the costs of making incorrect decisions as $C_{false\ alarm}$ and C_{miss}, and the benefits of making correct decisions as B_{hit} and $B_{correct\ rejection}$. The optimal decision threshold in this case can then be defined as:

$$\frac{p_{non\text{-}deceptive}}{p_{deceptive}} \times \frac{C_{miss} + B_{correct\ rejection}}{C_{false\ alarm} + B_{hit}} \quad (1)$$

This formula provides the point on the ROC curve that leads to the optimal balance between pH and pFA. From this ROC point, one can refer back to the information used to make the decision in order to identify the threshold on the continuum of evidence that resulted in this pH/pFA ratio.

This approach obviously requires the decision maker to quantify values that are extremely difficult to quantify (e.g. how could one calculate the exact cost of incorrectly predicting that a suspect is being deceptive in an interrogation setting?). To avoid these complications, other methods can be used. One approach is to ignore specific prevalence rates, or costs and benefits, and simply estimate their ratios (Swets et al., 2000a). For example, a decision maker may not know exactly how many deceptive suspects he or she will encounter, but it may be possible to estimate that they will encounter roughly three times as many deceptive suspects as non-deceptive ones. In cases like this, these ratios can be plugged into Formula 1 to define an appropriate threshold.

Alternatively, one can assume that the prevalence rates of the diagnostic alternatives are equal and that the costs and benefits associated with the various decision outcomes are equivalent (Hilden, 1991). On the one hand, when these ratios are plugged into Formula 1, the decision maker will maximize pH while minimizing pFA (since the threshold identified will correspond to the ROC point falling closest to the upper left-hand corner of the ROC graph). On the other hand, these ratios are extremely rare and the threshold resulting from their use will not be optimal for other ratios (Greiner et al., 2000).

To bypass these problems altogether, the decision maker may wish to pre-select a value of pFA (or pH) that is deemed appropriate for the situation at hand and avoid making decisions that exceed (or fall below) this pre-selected value (Swets et al., 2000a). A police force may decide, for example, that they do not wish to exceed $pFA = 0.05$ for a particular investigative task (due to limited financial resources perhaps). An appropriate decision threshold must then be set at that point on the ROC curve that maximizes pH without exceeding $pFA = 0.05$.

Calculating measures of diagnostic accuracy

While each ROC point represents the pH/pFA ratio that occurs when using a specific decision threshold, the height of the ROC curve represents the overall level of discrimination accuracy achieved by the decision maker (i.e. higher ROC curves are characterized by greater pH/pFA ratios) (Swets et al., 2000a). Height in this case is measured by calculating the proportion of the graph's area falling under the curve, often referred to as the AUC (Swets, 1988). AUCs typically range in value from 0.50 to 1.00. An AUC of 0.50 corresponds to a ROC curve falling along the positive diagonal and indicates discrimination accuracy at the level of chance. An AUC of 1.00 corresponds to a ROC curve falling along the left and upper axes of the graph and indicates perfect discrimination accuracy.

According to criteria proposed by Swets (1988), AUCs between 0.50 and 0.70 indicate low accuracy, AUCs between 0.70 and 0.90 indicate moderate accuracy, and AUCs between 0.90 and 1.00 indicate high accuracy. Thus, the ROC graph in Figure 2 provides examples of two ROC curves that reflect low and moderate levels of accuracy. The most important point to emphasize here is that the AUC provides a measure of discrimination accuracy that is independent of any specific threshold. This is because the AUC corresponds to the position of the entire ROC curve rather than any single ROC point (Swets et al., 2000a). This is one of the primary advantages of using ROC related accuracy measures (Swets, 1986). At the same time, decisions that are made at each and every decision threshold are preserved in the analysis, allowing one to identify optimal thresholds.

Depending on what assumptions the researcher wishes to make about the data, the AUC measure can be calculated in a variety of ways. For example, methods exist for calculating AUCs based on single ROC points and for partial segments of a ROC curve (Hanley & McNeil, 1982; McClish, 1989). Non-parametric AUCs can also be calculated when one does not wish to assume anything about the underlying probability distributions associated with the diagnostic alternatives (Vida, 1993). However, the most common AUC measures are parametric AUCs. In cases where parametric AUCs are calculated, it is assumed that the underlying probability distributions of the diagnostic alternatives are normally distributed with equal variances, or that the distributions can be transformed to the normal (Swets, 1996, argues that these assumptions are often valid). Parametric AUCs can now be calculated using a range of software programs (e.g. ROCKIT; Metz, Hermann, & Shen, 1998). These programs typically calculate AUC measures based on smoothed ROC curves, generated using maximum likelihood estimation procedures (Greiner et al., 2000). A smoothed ROC curve is essentially an estimate of the ROC curve that would result if each and every possible decision threshold were examined.

CURRENT APPLICATIONS OF ROC ANALYSIS

Although ROC analysis has been used much less frequently in the policing context, as compared to other diagnostic fields, it has been applied to a number of important tasks. The first application to be discussed here is bite mark identification and the second involves strategies for linking serial crimes. Beyond obvious task differences, distinctly different issues have been examined within each of these domains.

Bite mark identification

In the area of bite mark identification, one line of research has focused on peoples' ability to distinguish between bite marks produced by adults and children. Specifically, researchers have been interested in identifying the core skills required to make accurate predictions on this task. For example, Whittaker et al. (1998) examined this skill in senior and junior forensic dentists, general dentists who had no forensic experience, dental students who had taken a forensic dentistry course, and police officers and social workers who had experience with bite mark cases. Each participant was provided with a series of 50 colour photographs of actual bite marks. A portion of these photographs depicted non-accidental biting injuries inflicted by adults while the remainder depicted accidental biting injuries inflicted by children. The task for each participant was to determine, on a scale from 1 to 7, how likely it was that the bite mark was made by an adult.

Participant ratings were compared to ground truth in each case (actual court verdicts and additional corroborative evidence) in order to construct a ROC curve for each group. Whittaker et al. (1998) found that the most accurate decisions were made by senior forensic dentists (AUC = 0.69), dental students (AUC = 0.69), and junior forensic dentists (AUC = 0.68), with significantly less accurate decisions being made by social workers (AUC = 0.63), general dentists (AUC = 0.62), and police officers (AUC = 0.62). Whittaker and his colleagues suggest these results indicate that training in forensic dentistry is important for the particular task they examined, and that this training might be more important than extensive experience (since senior forensic dentists were not significantly more accurate than forensically trained dental students or junior forensic dentists).

What is important to take away from this study, in addition to these interesting findings, is that the only way these various practitioner groups could have been compared validly was by calculating AUC measures. In the absence of such measures, it would have been extremely difficult to determine whether Whittaker et al.'s findings were due to group differences in the ability to distinguish between adult and child bite marks, or to differences in the decision thresholds adopted by participants.

In a somewhat similar study, Arheart and Pretty (2001) used ROC analysis to examine whether forensic dental experts could accurately assign a bite mark to the dentition that made the mark. Thirty-two experts were given four sets of colour photographs of bite marks along with seven dental models to compare to the marks. The task for each participant was to rate the strength of the link between each bite mark and the seven dental models using a 7-point Likert-scale. When these rating were used to construct a ROC curve the resulting AUC was 0.86, indicating that the examiners were able to correctly identify the dentition belonging to a particular bite mark in the majority of cases. Because these researchers used the AUC, we can be confident that this level of accuracy is not biased by the thresholds that were adopted to make a positive prediction.

Linking serial crimes

A second area where ROC analysis has recently been employed is in the evaluation of logistic regression models for linking serial crimes to a common offender (Bennell & Canter, 2002; Bennell & Jones, 2005). To demonstrate how ROC analysis could be used to assist with this task, Bennell and Canter examined the behaviours of 43 commercial burglars who committed two crimes each. Based on existing research, it was hypothesized that crimes committed by the same offender (linked crimes) would be characterized by different patterns of across-crime similarity compared to crimes committed by different offenders (unlinked crimes). Specifically, the hypotheses were that linked crimes would be closer together geographically compared to unlinked crimes and be characterized by higher levels of behavioural similarity in terms of target choices, entry behaviours, and property stolen.

Across-crime similarity measures were constructed for each behavioural domain and these were used to develop logistic regression models. The goal was to determine whether any of these measures, considered separately or in combination, could be used to accurately predict whether pairs of burglaries were linked. The outputs of the regression analyses, which consisted of estimated probabilities that each pair was linked, were then subjected to ROC analysis to obtain valid measures of discrimination accuracy and examine the impact of setting different decision thresholds.

The results from this study indicate that AUCs vary drastically across similarity measures, with across-crime distances being the most effective predictor (AUC = 0.80), followed by the similarity measures for target choices (AUC = 0.68), entry behaviours (AUC = 0.65), and property stolen (AUC = 0.63). While the most accurate strategy was to combine the distance and entry information into a single regression model (AUC = 0.81), it can be seen that the distance information could be used in isolation with little decrease in accuracy. Such a finding could have significant implications for how burglary data is collected and for how serial burglaries are linked.

When using inter-crime distances to carry out the linking task, the placement of the decision threshold was shown to have a very large impact on decision making utility. For example, when the threshold was set so that any crimes within 0.70 km of one another were considered linked, 52% of linked crimes and 93% of unlinked crimes were correctly classified. However, when the threshold was made more lenient, by extending it to 2.50 km, 62% of linked crimes and 68% of unlinked crimes were correctly classified. In other words, by making the decision threshold more lenient, Bennell and Canter (2002) could certainly make more hits, but at the cost of making many more false alarms.

To understand the practical significance of their results, Bennell and Canter (2002) adopted a procedure used by previous researchers (e.g. Swets et al., 2000a). First, they set a pre-determined limit on the rate of false alarms they imagined police forces would be happy with (they set this rate at pFA = 0.20 indicating that the actual rate would be determined largely by the resources police had available to investigate burglary). They then examined what pH values could be achieved across the different predictor variables at this restricted level of pFA. For example, for the ROC curve corresponding to property stolen, the least accurate predictor variable, the threshold that limited pFA to 0.20 resulted in pH = 0.40. In contrast, for the ROC curve corresponding to across-crime distances, the most accurate predictor variable, the threshold that limited pFA to 0.20 resulted in pH = 0.64. Thus, if the police force that participated in this study were primarily concerned with making additional hits when linking serial burglaries, and they wished to restrict pFA to 0.20, they could identify 24 additional linked crime pairs for every 100 pairs encountered by drawing on across-crime distances instead of property stolen (0.64–0.40 = 0.24).

FUTURE APPLICATIONS OF ROC ANALYSIS

In addition to these two current applications, there appear to be many other tasks in the policing context where ROC analysis could be used. Two such tasks are statement validity assessment and determining the veracity of suicide notes. As far as I am aware, ROC analysis has never been used to examine either of these tasks, despite the fact that its use could solve existing problems in each of the areas.

Statement validity assessment

Statement validity assessment is a technique used to establish the credibility of statements made by children who have allegedly been the victim of sexual abuse (Tully, 1999). The specific task is to discriminate between truthful and fabricated statements, largely on the basis of statement content. The technique consists of three separate components: (1) a structured, open-ended interview with the child, (2) a systematic analysis of the interview using criteria-based content analysis (CBCA), and (3) a validity checklist, which involves

an examination of other characteristics of the interview, the witness, and the investigation (Raskin & Esplin, 1991). The second component, CBCA, is considered the core component of statement validity assessment (Berliner & Conte, 1993), and it is the component focused on here (though the current discussion will apply equally well to the validity checklist).

At one point, CBCA consisted of 19 criteria that were applied to the content of the child's statement and used to estimate the statement's veracity (Ruby & Brigham, 1997). However, CBCA can now be found in a variety of forms, and most commonly consists of 18 criteria (Raskin & Esplin, 1991). Regardless of how many criteria are used, the presence of a criterion is assumed to indicate that the child is telling the truth, with the assumption being that '. . . memory of an actual experience differs in verbal quality and content from statements that are invented' (Ruby & Brigham, 1997, p. 709). Brief examples of the criteria coded for in CBCA are: quantity of details (are there descriptions of place or time?), contextual embedding (are events placed in spatial and temporal context?), and admitting lack of memory (did the child indicate lack of memory for an aspect of the incident?).

Attempts to empirically validate CBCA as a procedure for determining statement veracity have led to mixed results. There are studies carried out in the field and the lab that indicate CBCA may be a useful procedure (e.g. P. W. Esplin, T. Boychuck, & D. C. Raskin, paper presented at the NATO Advanced Study Institute of Credibility Assessment, Maratea, Italy, June 1988, as cited in Ruby & Brigham, 1997; J. Yuille, paper presented at the NATO Advanced Study Institute of Credibility Assessment, Maratea, Italy, June 1988, as cited in Ruby & Brigham, 1997). However, there are also studies that report less favourable results (Kohnken & Wegener, 1982, as cited in Ruby & Brigham, 1997). Despite what these studies say about CBCA, Ruby and Brigham (1997) suggest there are still many questions that remain unresolved: How accurate is the procedure? How many criteria should be present before deciding that the child's statement is true? Are some criteria more useful than others? Does the type of interview procedure have an impact on the results? Can the procedure be used with statements made by adults?

ROC analysis is a procedure that can provide answers to these questions. What should now be clear is that previous attempts to demonstrate the accuracy of CBCA will likely have been threshold specific. For example, if one were to use the decision threshold proposed by Landry and Brigham (1992), where the presence of five or more criteria is taken to indicate truth, the result may be quite different than if one were to adopt the seven or more criteria threshold recommended by J. Yuille (paper presented at the NATO Advanced Study Institute of Credibility Assessment, Maratea, Italy, June 1988) (as cited in Ruby & Brigham, 1997). As previously described, ROC analysis provides a way to obtain valid measures of discrimination accuracy that are not biased by the threshold adopted. In addition to solving this problem, ROC analysis could also be used to determine an optimal decision threshold for identifying truthful statements of abuse, and to make any comparisons that are of interest to the researcher. For example, separate ROC curves could be generated to compare different combinations of criteria, different types of interview methods, or different categories of subjects.

The veracity of suicide notes

In a similar way, ROC analysis could quite easily be applied to problems in determining the veracity of suicide notes. Based on the hypothesis that there exist a number of features

more typical of genuine suicide notes, numerous attempts have been made to develop discrimination procedures (e.g. Black, 1993; Edelman & Renshaw, 1982; Gregory, 1999). For example, in one of the most recent studies, Gregory (1999) examined both genuine and simulated suicide notes and determined that the content of the notes had more discriminatory power than their structure. In Gregory's study, structural variables included things like the percentage of nouns found in the notes and the average sentence length. In contrast, content variables included things like the writer leaving instructions and providing explanations for their acts.

Gregory (1999) used the content variables found in a sample of 66 suicide notes (33 genuine and 33 simulated) to construct a scale that was meant to measure the degree to which the note writer had internalized the decision to die. Consistent with his hypothesis, there was an observable trend in the results, with genuine notes falling higher up on this scale. Compared to simulated notes, genuine notes were generally characterized by higher levels of positive affect, the inclusion of instructions, fewer explanations for the act, and indications of an external locus of control. Similar results were found in a small sample of suicide notes that Gregory used to validate his findings.

Gregory (1999) did not set a decision threshold to specifically define when a suicide note should be considered genuine, though it seems from his results that a reasonable threshold could have been set. As a result, no attempt was made to measure the level of discrimination accuracy that could be achieved with his scale. In any case, if he had done this, the accuracy rates reported would have been of limited value, since they would only be valid for the threshold he adopted (and this threshold may be inappropriate in other settings). Thus, as in the case of statement validity assessment, ROC analysis could prove a useful tool to deal with these problems.

POTENTIAL DIFFICULTIES WITH ROC ANALYSIS

While the discussion above suggests that ROC analysis will be of value in the policing context, problems will no doubt be encountered when the procedure is used. Some problems affect the validity of accuracy measures, while others have an impact on setting optimal thresholds. Five difficulties will be discussed that relate to accuracy measures. These have been proposed by Swets (1996) and include: (1) establishing ground truth, (2) ensuring independence of truth determination and the diagnostic system, (3) ensuring independence of the sample and truth determination, (4) obtaining representative samples, and (5) setting up a realistic decision making environment. The major difficulties associated with selecting decision thresholds, which involve estimating the prevalence rates of the diagnostic alternatives and the costs and benefits of the various decision outcomes, have already been discussed, as have potential solutions to these problems.

Establishing ground truth

One of the main problems that will be encountered when using ROC analysis in the policing context is that it may be difficult to establish ground truth. To obtain valid measures of discrimination accuracy, the researcher must be able to do this (Swets, 1988). Consider the bite mark example described above. Under field conditions it may be difficult to know for certain whether a suspect was responsible for the bite mark in question. While researchers have devised a number of methods for establishing ground truth in such studies

(e.g. panel decisions, court verdicts, suspect confessions, etc.) each of these methods is far from perfect. The same is true for the other applications discussed above.

Unfortunately, there is no way to resolve this issue completely in many cases. So, what can the researcher do to minimize the problem? One obvious piece of advice is to use the most effective evidence available for establishing ground truth. For example, in bite mark studies, the researcher may be able to use DNA evidence to establish ground truth. When such evidence is not available, as might often be the case, the next best alternative may be to use multiple sources of less effective evidence, as some researchers have already done (e.g. Whittaker et al., 1998). If neither of these options is feasible, it may be necessary to resort to simulation based lab studies, where an accurate determination of ground truth is ensured. However, while the accuracy of ground truth will be enhanced under laboratory conditions, these studies will lack the realism of studies carried out in the field. Alternatively, one might choose a completely different approach for dealing with ground truth, by simply avoiding the need to determine it. While not the preferred method, researchers have proposed procedures for conducting ROC analysis without the need for ground truth (e.g. Beiden, Campbell, Meier, & Wagner, 2000; Henkelman, Kay, & Bronskill, 1990).

In any case, and regardless of what approach is ultimately taken, the researcher who examines diagnostic decision making will simply have to be aware of the many problems that can exist in establishing ground truth and report them to the reader (or practitioner) if they are relevant. In this way, they can ensure that an appropriate level of caution is used when interpreting their results.

Ensuring independence of truth determination and the diagnostic system

The second problem that may be encountered is that ground truth may be determined, at least in part, by the diagnostic system being evaluated. For example, the content of a child's statement of sexual abuse may indicate to a police officer that the child is telling the truth, and this may lead that officer to 'push hard' for a confession from the accused. A problem arises when the researcher uses that confession to determine ground truth, because they have not isolated the determination of truth from the diagnostic system. As Swets (1996) makes clear, this is a problem because the accuracy of the diagnostic system '. . . is scored against a determination of truth that it helped to make . . . [and therefore the system] ought to do well' (p. 113).

The only way to eliminate this problem is to ensure that ground truth is always determined independently of the diagnostic system being evaluated. In the statement validity example just described, where suspect confessions guided sample selection, the CBCA criteria should play no part in the extraction of a confession.

Ensuring independence of the sample and truth determination

The third problem that may be encountered is that the procedures used to determine ground truth may affect the selection of cases, '. . . resulting in an easier sample than is realistic' (Swets, 1996, p. 113). Such a problem could exist in many of the tasks described above. For example, consider the linking study conducted by Bennell and Canter (2002). They used arrest records as the basis for determining ground truth. Because arrests were the sole basis for determining truth, the sample of burglaries that were selected for analysis will not be entirely representative of the sorts of burglaries the police have to link.

Specifically, the truth determination procedure used in this case restricts the sample to solved serial burglaries.

Unfortunately, not all serial burglaries that are committed result in an arrest, and burglaries that do, are solved for a reason. One likely reason why some serial burglaries are solved is because they are characterized by high levels of behavioural similarity. This may not be the case for unsolved serial burglaries. If this is true, the measures of accuracy reported by Bennell and Canter (2002) are likely higher than what would be found in the field. This is because, in the field, the logistic regression models developed by Bennell and Canter would have to be applied to all burglaries, not just those that end up being solved.

To minimize this problem, one must ensure that the sample of cases selected for analysis, and the procedure used to determine ground truth, are independent. For example, in Bennell and Canter's (2002) study, they could have used some other method to determine ground truth that would have had less of an impact on the types of burglaries selected (i.e. not restricting their sample to just solved serial burglaries). One possibility would have been to determine ground truth based on arrest records and burglaries that were taken into consideration (TIC) by the police. TIC crimes are unsolved crimes that an offender admits to committing after they have been arrested by the police. In this case, TIC burglaries become 'solved', allowing them to be incorporated into research, but there is nothing about those burglaries per se that resulted in them being cleared by the police (such as high levels of across-crime similarity).

Obtaining representative samples

The fourth problem that may be encountered has to do with obtaining representative samples and determining the extent to which results are generalizable to other cases. In order to obtain valid measures of accuracy, the sample of cases drawn on must be similar to those that will eventually be applied to the diagnostic system (Swets, 1988). This is often not achieved. For example, consider the representativeness of the suicide notes collected by Gregory (1999). Gregory examined 33 genuine notes (and 33 simulated notes) that were written before 1957. Not only is the sample size relatively small, the time period when the notes were written may mean they are not representative of the range of notes that would be encountered by the police today.

Obviously, researchers hope to solve this problem by collecting reasonably large and representative samples. However, in situations where this is not possible, researchers should provide as much detail as possible about their sample so that a determination can be made about the degree to which the findings generalize to other cases. Even when providing such detail, it would usually be appropriate to perform some sort of validation test, as Gregory (1999) did, where the accuracy of the diagnostic system is determined by how well it performs on a sample of yet-to-be-observed cases. A variety of procedures are now available for this purpose (e.g. Efron, 1982; Gong, 1986). However, while cross-validation will provide some indication of generalizability, the degree of generalizability will still depend on how closely the test sample approximates reality.

Setting up a realistic decision making environment

The fifth, and final, problem that may be encountered with ROC related accuracy measures involves setting up a realistic decision making environment. According to Swets (1996), when running lab studies, accuracy measures are likely to be inflated to an unknown

degree. This is due primarily to the fact that the decision making environment in lab studies is artificial, in the sense that it does not replicate all the problems a decision maker may face in the field. For example, just because it is possible for a police officer to identify false allegations of abuse in the lab, this does not mean the same level of accuracy could be achieved in realistic situations (e.g. when a person's life is on the line, when the police officer is tired, when the media is exerting pressure, etc.).

While studies conducted under laboratory conditions can be used to '. . . test a system in a standard way and at its full potential' (Swets, 1996, p. 120), some consideration should be given to these issues. Field studies provide an obvious alternative, despite the difficulties in determining ground truth, but if these are not possible an attempt can be made to approximate reality. For example, researchers routinely use monetary costs and benefits with their laboratory subjects to simulate real world consequences and gains (Getty, Swets, Pickett, & Gonthier, 1995). While such practices certainly do not capture fully the costs and benefits experienced by police officers in the field, they provide researchers with an opportunity to calculate slightly more realistic accuracy measures.

CONCLUSION

ROC analysis is a procedure that graphically illustrates the balance between hits and false alarms that will occur at each and every decision threshold, and it provides an estimate of discrimination accuracy that is independent of these thresholds. As a result, the procedure can be used to improve police decision making by allowing the police decision maker to set more appropriate decision thresholds and to estimate the accuracy of their decisions in a more valid manner. More specifically, threshold setting procedures will allow the police decision maker to adopt a threshold that results in the desired balance between hits and false alarms for any particular situation, a balance that can often be in tune with the probabilities of encountering the diagnostic alternatives and the costs and benefits of the various decision outcomes. A valid accuracy measure will allow the police to identify what evidence they should use to make their decisions, to determine if their choice of evidence should differ depending on the particular situation, and to compare the accuracy of different groups of decision makers or diagnostic systems.

ACKNOWLEDGEMENT

I would like to thank Krista Richard, Brent Snook, and Paul Taylor for their helpful comments on earlier drafts of this paper.

REFERENCES

Arheart, K. L., & Pretty, I. A. (2001). Results of the 4th bite mark workshop—1999. *Forensic Science International, 124*, 104–111.

Beiden, S. V., Campbell, G., Meier, K. L., & Wagner, R. F. (2000). On the problem of ROC analysis without truth: the EM algorithm and the information matrix. *Proceedings of the International Society for Optical Engineering, 3981*, 126–134.

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: tests using regression and ROC analysis. *Science & Justice, 42*, 53–64.
Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: a new method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and Offender Profiling, 2*, 23–41.
Berliner, L., & Conte, J. R. (1993). Sexual abuse evaluations: conceptual and empirical obstacles. *Child Abuse and Neglect, 17*, 111–125.
Black, S. T. (1993). Comparing genuine and simulated suicide notes: a new perspective. *Journal of Consulting and Clinical Psychology, 61*, 699–702.
Edelman, A. M., & Renshaw, S. L. (1982). Genuine versus simulated suicide notes: an issue revisited through discourse analysis. *Suicide and Life Threatening Behavior, 12*, 103–113.
Efron, B. (1982). *The jackknife, the bootstrap and other re-sampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.
Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*, 913–920.
Getty, D. M., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: a laboratory investigation of the effects of predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1*, 19–33.
Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association, 81*, 108–113.
Gregory, A. (1999). The decision to die: the psychology of the suicide note. In D. V. Canter, & L. J. Alison (Eds.), *Interviewing and deception* (pp. 129–156). Aldershot, UK: Ashgate Publishing.
Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). Principles and practical applications of the receiver operating characteristic analysis for diagnostic tests. *Preventative Veterinary Medicine, 45*, 23–41.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.
Henkelman, R. M., Kay, I., & Bronskill, M. J. (1990). Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making, 10*, 24–29.
Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making, 11*, 95–101.
Kohnken, G., & Wegener, H. (1982). Zur Glaubwuerdigkeit von Zeugenaussagen Experimentelle Weberpruefung ausgewaehlter Glaubwuerdigkeitskriterien. *Zeitschrift fuer Experimentelle und Angewandte Psychologie, 29*, 92–111.
Landry, K. L., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. *Law and Human Behavior, 16*, 663–675.
McClish, D. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making, 9*, 190–195.
Metz, C. E., Hermann, B. A., & Shen, J. H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Medical Decision Making, 17*, 1033–1053.
Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Professional Group on Information Theory, 4*, 171–212.
Raskin, D. C., & Esplin, P. W. (1991). Statement validity assessment: interview procedures and content analysis of children's statements of sexual abuse. *Behavioral Assessment, 13*, 265–291.
Ruby, C. L., & Brigham, J. C. (1997). The usefulness of the criteria-based content analysis technique in distinguishing between truthful and fabricate allegations: a critical review. *Psychology, Public Policy, and Law, 3*, 705–737.
Swets, J. A. (1986). Indices of discrimination of diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin, 99*, 100–117.
Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.
Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000a). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000b). Better decisions through science. *Scientific American, October*, 82–87.

Taylor, P. J. (2002). A partial order scalogram analysis of communication behavior in crisis negotiation with the prediction of outcome. *The International Journal of Conflict Management, 13*, 4–37.

Taylor, P. J., Bennell, C., & Snook, B. (2002). Problems of classification in investigative psychology. In K. Jajugam, A. Sokolowski, & H. H. Bock (Eds.), *Classification, clustering and data analysis: Recent advances and applications* (pp. 479–487). Heidelberg: Springer.

Tully, B. (1999). Statement validation. In D. V. Canter, & L. J. Alison (Eds.), *Interviewing and deception* (pp. 85–103). Aldershot, UK: Ashgate Publishing.

Vida, S. (1993). A computer program for non-parametric receiver operating characteristics analysis. *Computer Methods and Programs in Biomedicine, 40*, 95–101.

Whittaker, D. K., Brickley, M. R., & Evans, L. (1998). A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis. *Forensic Science International, 92*, 11–20.

