Victoria L. Phillips,[1] *Ph.D.; Michael J. Saks,*[2] *Ph.D.; and Joseph L. Peterson,*[3] *D.Crim.*

The Application of Signal Detection Theory to Decision-Making in Forensic Science

REFERENCE: Phillips VL, Saks MJ, Peterson JL. The application of signal detection theory to decision-making in forensic science. J Forensic Sci 2001;46(2):294–308.

ABSTRACT: Signal Detection Theory (SDT) has come to be used in a wide variety of fields where noise and imperfect signals present challenges to the task of separating hits and correct rejections from misses and false alarms. The application of SDT helps illuminate and improve the quality of decision-making in those fields in a number of ways. The present article is designed to make SDT more accessible to forensic scientists by: (a) explaining what SDT is and how it works, (b) explicating the potential usefulness of SDT to forensic science, (c) illustrating SDT analysis using forensic science data, and (d) suggesting ways to gain the benefits of SDT analyses in the course of carrying out existing programs of quality assessment and other research on forensic science examinations.

KEYWORDS: forensic science, signal detection theory, SDT, ROC, measurement, research

Signal Detection Theory (SDT) constitutes both a body of knowledge and a set of analytical methods designed to rigorously examine decision making by machines and people alike. SDT is a product of the marriage in the 1950s of mathematical statistics and advances in electronic communications (1).

Radar is the archetypal technology that precipitated SDT's development and provided its principal metaphor. The advent of radar technology presented air traffic observers with the challenge of detecting the "signal" returned from aircraft amid the din of instrumentation disturbances, echoes, and other "noise" that closely resembled the signal itself. (Note that we already are encountering the terminology of SDT. To assist the reader, Appendix 1 provides a list of major terms used in Signal Detection Theory and their definitions.) Errors could lead to such tragedies as mistaking a foe for a friend (or vice versa), or allowing two planes into a vector where only one could safely fly. SDT provided a means to quantify and analyze such decision problems, and in so doing enable the decisions to be optimized.

Since ts origins, SDT methodology has been adopted for use in a wide variety of fields, among them:

- Military applications. Given SDT's origins in the testing of various technologies and procedures useful to the military, it is not surprising that SDT has been employed to optimize the use not only of radar but of sonar, seismic detectors, and laser radar, as well as in evaluating human performance in activities ranging from visual observation techniques to air combat (2–6).
- Psychology and Human Factors. SDT is a familiar tool in psychological research, where it has been applied to a broad range of psychological processes, including perception (7–10), memory (11–14), forecast accuracy (15), human vigilance (16,17), and even the detection of social cues (18).
- Medical diagnosis. In radiology, SDT has been utilized both to assess the practitioner's interpretation of radiographic output and to evaluate the diagnostic accuracy of various radiographic technologies (19–23). For instance, through the implementation of SDT methodology, computed tomography was determined to be diagnostically superior to radionucleide scanning for tumor detection (22)—a finding that revolutionized medical diagnosis.
- Psychiatric diagnosis. SDT methodology has been used extensively in the assessment and development of tests and measures to facilitate psychiatric diagnosis (24–28).
- Dentistry. SDT methodology has been used to evaluate endodontic techniques as well as to compare the effectiveness of different imaging and diagnostic procedures (29,30).
- Clinical chemistry. Chemists utilize SDT to identify and assess various chemical parameters. For example, SDT has been used to identify chemical indicators of organ rejection (31) and to distinguish cancerous lesions (32).
- Engineering. SDT has found important uses in engineering, in such areas as acoustical engineering and electrical engineering, and especially in telecommunications (33–36).

SDT also would be helpful, and perhaps ideal, for analyzing the decision making involved in forensic examinations. As in other fields, SDT methodology could contribute to deepening our understanding of decision making in forensic examinations and to improving the quality and accuracy of the conclusions resulting from those examinations. Using SDT, it is possible to quantify and analyze the diagnostic capabilities of forensic examiners, their instruments, and their procedures. Indeed, in a few instances, SDT already has had forensic science applications, though most of that work has not been conducted by forensic scientists. For example, psychologists have used it to assess the accuracy of polygraphic lie detection (37,38). Communications experts and engineers have applied SDT to forensic voice spectrography (39). And engineers and cognitive scientists have used SDT to develop pattern recognition procedures that may be useful in a variety of areas of forensic science (40).

[1] Visiting professor, Department of Psychology, Arizona State University, East Campus, Mesa, AZ.
[2] Professor of Law and Psychology, Arizona State University, Tempe, AZ.
[3] Professor of Criminal Justice, University of Illinois at Chicago, Chicago, IL.

Received 17 Dec. 1999; and in revised form 24 March 2000; accepted 24 March 2000.

Very recently, two studies by forensic dentists have employed SDT. Whittaker et al. (41) compared the performances of senior and junior forensic odontologists, final year dental students, general practitioners with no forensic experience, police officers, and social workers in distinguishing the bitemarks of children and adults. (The first three of those groups performed equally well and significantly better than did the latter three groups.) In the second study, Andersen and Wenzel (42) compared two dental imaging techniques: subtraction radiography and conventional bitewing films. The patients examined had no dental restorations, thus adding to the difficulty of identification. Examiners were asked to match each bitewing to its source using a four-point confidence rating scale. (The conventional bitewing films produced poor performance [three out of four observers were inaccurate], accuracy was notably improved with subtraction radiography.)

The present article is designed to make Signal Detection Theory more accessible to forensic science. The article will: (a) explain briefly what SDT is and how it works, (b) explicate the potential usefulness of SDT to forensic science, (c) illustrate SDT analysis using forensic science data, and (d) suggest ways to gain the benefits of SDT analyses in the course of carrying out existing programs of quality assessment and other research in and on forensic science.

An Overview of Signal Detection Theory

Origins and Purposes of Signal Detection Theory

As mentioned above, the theory of signal detectability (as it was termed originally) was developed in the contexts of mathematics and engineering. Signal Detection Theory was specifically designed to address problems encountered in radar identification. While the advent of radar has proven invaluable to modern commercial and military aviation, the interpretation of radar waves is not precise. Consider the task of an air traffic controller, who directs pilots into safe flight paths, that is, into unoccupied vector airways. The controller must decide whether or not a particular path is occupied by any other aircraft. The opposing aircraft can be thought of as the "signal" that the air traffic controller must detect—if it is present—in order to avoid a flight hazard. The controller makes these judgments on the basis of examining a complex array of radar signals. The risks are that the controller may erroneously conclude either that an opposing aircraft is present when it is not, or that one is not present when it is.

Prior to the inception of SDT, analytical techniques focused on a single index of judgmental accuracy, namely, the percentage of correct decisions. The accuracy of detection was considered, statistically speaking, to be a lone function of ability: accuracy was equated with the sensitivity of the observer (or the equipment) to the presence of the signal. That is, each outcome, be it safe passage or fatal crash, was presumed to reflect only the raw ability of a radar technician to perceive a true signal, discriminating this signal from mere radar disturbance, or "noise."

The technician's discrimination ability, however, is not the sole determinant of the judgment outcome. Given ambiguous input to scour, the technician must decide whether a noisy pattern is or is not a true signal. Does a radar blip indicate the presence of another aircraft or is it merely static? Given an indistinct pattern to analyze, even the most competent technician is forced to make an uncertain judgment. In order to make that judgment, the examiner must form an implicit decision threshold, a borderline that divides a decision that a "signal is present" from a decision of "no signal." The location of that threshold is affected, among other things, by the realization that some kinds of errors are more serious than others. For example, the error of deciding that a flight vector is unoccupied when in reality it is occupied is more dangerous than the error of deciding that a vector is occupied when in reality it is clear. To optimize those decisions, more needs to be known about the factors that contribute to judgment and their dynamics.

Swets (43) provides a candid reminder that scientists and diagnosticians frequently operate in a realm of imprecision and ambiguity, where decisions nevertheless must be made despite murky, indistinct data in less than clear contexts.

> Diagnostic tests and systems of many kinds are used in a host of practical settings to assist in making a positive or negative decision about the occurrence of a particular event or the existence of a particular condition. Will an impending storm strike? Is this aircraft unfit to fly? Is that plane intending to attack this ship? Is this nuclear power plant malfunctioning? Is this assembly-line item flawed? Does this patient have the acquired immune deficiency (AIDS) virus? Is this person lying? Is this football player using drugs? Will this school (or job) applicant succeed? Will a document so indexed contain the information sought? Does this tax return justify an audit? Is there oil in the ground here? Will the stock market advance today? Will this prisoner vindicate parole? Are there explosives in this luggage?
>
> These examples are a reminder that diagnostic test results do not usually constitute compelling evidence for or against the condition or event of interest, or evidence of a sort that leads directly to either a positive or negative decision (p. 522).

An Illustration

The ABC Laboratory screens skin biopsies for cancer. Tissue samples are scored on a scale that runs from 1 (normal tissue) through 7 (clearly malignant). This score measures the number and degree of abnormal cell elements.

A tissue sample from a 16-year-old girl is received and scored as a 3. A score of 3 is ambiguous: it might be an early indicator of melanoma (signal) or it might be due to normal tissue variation (noise). Because melanoma is very unusual in teenagers, the lab reports "no cancer" to the physician. But the lab is wrong, and 14 months later the girl dies of the melanoma. At a later civil trial for the missed diagnosis, laboratory technicians testify that had a sample scored as 3 come from a middle-aged or older person, they would have reported possible malignancy to the physician, since melanoma is more common in older adults. When the stakes are high, as in this case of unidentified cancer resulting in death, an error in signal detection can lead to disastrous consequences.

Unfortunately, diagnoses of this type do not involve clearly differentiated pieces of information. Many decisions, in this and other fields, are close calls made in the ambiguous, gray area where reality is less than certain. Given ambiguous input to sift through, an examiner must at some point make the cancer-or-not call.

The capability of the lab to discriminate tissue samples may not be the issue. Often of greater importance is the threshold set for interpretation. The lab described above may have required a rating of 4 or higher before it would declare a teenager's biopsy to indicate malignancy. With equal proficiency at perceiving and scoring biopsy samples, the lab could have set 3 as the minimum threshold for reporting samples from teenagers as cancerous. An undesirable trade-off for doing that, of course, is that more noncancerous samples would then be reported as malignancies, leading to unnecessary and potentially harmful treatment.

An Explication of SDT

Signal Detection Theory is an analytical approach specifically designed to address ambiguous decision scenarios, such as in the preceding illustration. This approach is designed to disentangle the two major components of accuracy inherent in fuzzy-choice decision situations: the examiner's discrimination ability (or diagnostic ability, or degree of sensitivity to the evidence) and the decision threshold used by the examiner. Two examiners with equal discrimination ability will reach different conclusions if they employ different decision thresholds for determining what is and what is not a signal. In our biopsy lab illustration, diagnostic accuracy is a function of *both* the lab's ability in testing for cancer (given proper equipment and expertise) *and* the decision threshold used to determine whether or not a sample is diseased.

Thus, decision outcomes depend upon two distinct components of accuracy: discrimination ability and decision thresholds.

Discrimination Ability—Discrimination ability can be thought of as the capacity to analyze the information (or discriminate the evidence) at hand. It is the perceptual ability to discern similarities and differences among stimuli. Discrimination capacity is a function of both (a) the technician's ability and (b) the quality of the evidence. "Quality of the evidence" refers to the fact that where the signal-to-noise ratio is greater, sensitivity will be greater, and vice versa. Quality is greater where the signal-to-noise ratio is higher. Differences in ability no doubt exist across labs, examiners, and types of analyses being performed.

Decision Threshold—But even assuming that all medical radiology labs, for example, had equal analyzing ability (due to similar experience, knowledge, skill, and technology), and therefore were equally proficient in discriminating the evidence, they still could differ in their decision outcomes because they use different decision thresholds. These differences in decision thresholds result in some labs making more and others making fewer mistakes in their ultimate case decisions. The decision threshold or decision point is, figuratively speaking, the line which, when perceived to be crossed, turns a presumed negative into a positive. The decision point represents the confidence threshold of the examiner for affirming the presence of diseased tissue (for one example, or matching latent and known fingerprints, for another). The decision threshold set for examining evidence often varies across different laboratories, examiners, and circumstances. Two factors are directly relevant to setting the decision threshold:

Prior probability of a positive—In the example of diagnosis of disease, is the disease frequent or infrequent? If infrequent, a strict (or conservative) criterion typically is set for concluding that disease is present (i.e., a higher scale rating is needed). If frequent, a lenient criterion typically is set (i.e., a lower scale rating is necessary to conclude that disease is present).

Utilities associated with each possible outcome—The utilities, or costs and benefits, of a decision can include life-and-death consequences, time, money, and numerous other considerations. For example, false positives can lead to the dangers of chemotherapy and associated costs of treatment. False negatives can lead to worsening of the disease and the risk of death.

An example of the impacts of context and motivation on the setting of subjective decision points is provided by a military example. The USS *Stark* was attacked by an enemy aircraft that had not been detected (43). In response, observers' criteria for detection lowered as they became more vigilant. The consequence was an erroneous attack on a civilian airliner mistakenly judged to be an attack aircraft. As different as these two outcomes are, they can be the product of equally good stimulus discrimination, accompanied by shifting decision criteria. The accuracy of decisions is thus biased by the decision threshold, and does not depend only on an examiner's detection ability.

The biopsy laboratory example will enable us to see how both factors, discrimination ability and decision tendency, affect decision outcomes. First of all, some kinds of cancers are easier to discern than others, that is, the quality of information varies across cancer types. Moreover, obviously, some lab technicians and some labs are more skilled than others in discriminating whatever evidence is submitted. But, within a given lab or examiner, discrimination capacity remains fairly constant. Thus, regardless of the patient's age, the lab utilizes the same procedures to test samples. For both a 16-year-old and a 60-year-old, the lab's competency (i.e., its discrimination ability) is identical. What does differ between these two cases is the decision criterion used to make the binary cancer-or-not decision. The decision threshold represents the point at which the lab is willing to declare "cancer." There is no magic formula for setting decision thresholds; they vary as a function of the psychological factors of expectancies (perceived prior probabilities) and motivation (perceived utilities).

For example, given the possibility of a rapidly progressing cancer such as melanoma, an erroneous diagnosis of cancer is a more "desirable" or less costly mistake than is overlooking a cancer that could result in death. Considering only these costs, labs would be motivated to adopt a very lax decision criterion: ambiguous scores would be classified as cancerous. The *likelihood* of having cancer, however, is also an issue. Since, for example, older people are much more likely (than younger ones) to develop cancerous melanoma, criteria will tend to be set lower for older patients and higher for younger patients. Accordingly, on identical evidence, different decisions will be made. The lab, trying to establish the best criterion for choice, likely takes into account *both* the prior probabilities and the utilities of each possible outcome. If labs do not address these issues directly and consciously, then the setting of decision thresholds is, in effect, delegated to the varying psychology of individual examiners.

With the biopsy example in mind, we can begin to formalize and then quantify what we can know about such decision-making. There are two alternative states of "reality": disease (signal presence) vs. no disease (signal absence). An examiner can make one of two possible determinations: disease present vs. disease absent. Thus, one of four joint outcomes is possible for this or any diagnosis: the lab could be in error either by (a) overlooking a cancerous sample (a "miss" in the language of SDT) or (b) labeling a cancer-free sample as malignant (a "false alarm"). Alternatively, the lab could be correct either by (c) accurately diagnosing cancer (a "hit") or (d) correctly classifying a sample as benign (a "correct rejection"). These four outcomes are depicted in Fig. 1.

Costs and benefits are associated with each of these decisions, and the choices involve tradeoffs (43). A fundamental point is that trade-offs exist among the proportions, or probabilities, of the four outcomes. Exactly where the criterion is set determines the balance among these proportions. Where the decision threshold is low or "lenient," more hits will occur, but more false alarm errors also will occur. Where the threshold is high or "strict," more correct rejections will occur, but so will more misses. A technician can be con-

sistently competent, but vary in detection accuracy as a function of using different decision criteria.

The Relative (or Receiver) Operating Characteristic Curve

One of the greatest advantages of SDT is that it permits researchers to assess "accuracy" as an *independent* function of discrimination ability. That is, SDT allows one to disentangle raw discrimination capacity from the effects of varying decision thresholds. This is accomplished by the use of a specific analytic tool associated with SDT, the ROC curve, shown in Fig. 2. ROC is an acronym for "relative operating characteristic" (also often referred to as "*receiver* operating characteristic"). It is a graphical representation of decision factors.

The ROC graph is a plot of the proportion of hits against the proportion of false alarms for a specified number of decisions. Note that although four outcomes are possible for each decision, as discussed above, only two independent response probabilities exist. Thus, for example, given a true diseased tissue sample, the response can be either a hit or a miss; the sum of hit and miss probabilities equals 1.00. Given a true nondiseased sample, the response outcome will be either a false alarm or a correct rejection; the sum of these two response probabilities is also equal to 1.00. As a result, knowing the probability of hits provides information about the probability of misses. Likewise, given either a false alarm or correct rejection probability, one can infer its complement. Consequently, it is necessary to plot only two response proportions, one corresponding to the truly diseased samples and one corresponding to the truly disease-free samples. (It is customary to plot hits and false alarms, though plotting misses against correct rejections would be conceptually equivalent and equally informative.)

	Reality	
Decision	**Disease**	**No Disease**
Yes, diseased	**Hit**	**False Alarm**
No, not diseased	**Miss**	**Correct Rejection**

FIG. 1—*Four possible decision outcomes.*

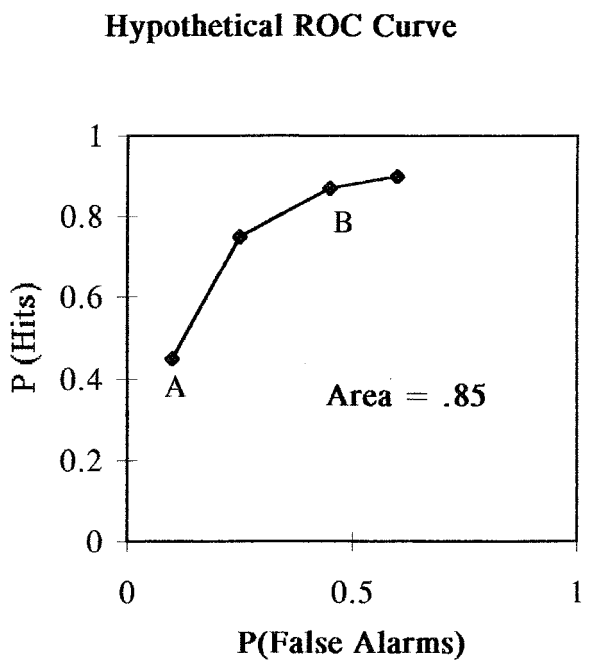

FIG. 2—*An illustrative ROC curve with four plotted outcome proportions: (0.10, 0.45), (0.25, 0.75), (0.45, 0.87), (0.60, 0.90).*

The illustrative ROC curve plotted in Fig. 2 will help us to explicate the information provided by an ROC curve. The X-axis represents the proportion of responses that are false alarms. On the Y-axis are the proportion of responses that are hits. Both conceptually and graphically, the ROC curve provides a distinction between discrimination ability and decision thresholds, and presents information on both, separately. Each single plotted point, i.e., a hit and false alarm proportion pair, represents one decision threshold. The ROC curve is created by connecting all available decision threshold points. The measured area under the ROC curve provides an index of accuracy that reflects pure discrimination ability—unaffected by varying decision thresholds. Area can take on values between 0.50, representing chance performance (where false alarm and hit rates are equal), and 1.00, representing perfect discrimination between truly diseased and truly nondiseased samples.

Importantly, the ROC curve in Fig. 2 represents all possible decision thresholds for a fixed discrimination ability of 0.85. Thus, for each decision, the lab(s) may be regarded as having good discrimination ability, but the particular decision outcomes vary as a function the different decision thresholds. To clarify, though discrimination ability is constant for a particular lab, a distinct threshold is used for each decision. Different decision thresholds alter the relations among the decision outcomes, such that threshold changes would move points along the curve. More conservative thresholds, reflecting a bias against affirmative decisions, are represented by the points toward the lower left corner. Point A in Fig. 2 represents a conservative threshold where false alarms are few (0.10), but a lower hit rate (0.45) is the tradeoff. That is, slightly fewer than half of truly diseased samples are classified as such, but few non-diseased samples are erroneously classified as cancerous. By contrast, though discrimination ability is constant at 0.85, Point B represents a relatively uncautious decision threshold where the observer is more likely to respond positively and declare a sample to be cancerous. Again, note the tradeoff among outcomes: the hit rate for cancer detection is higher at 0.87, but because the threshold has become more lenient, the false alarm rate has increased to 0.45. Thus, the ROC graph quantifies and displays both pure discrimination ability (the area under the curve) and the subjective "threshold" points employed for each decision.

The "Value-Added" of SDT

In sum, what does SDT and its ROC analysis add to the study of the kinds of decision-making we have been discussing, over and above counting correct versus incorrect answers? According to Swets and Pickett (44), SDT provides three distinct and unique advantages in analyzing decision outcomes.

First, ROC provides a pure index of accuracy, or raw diagnostic ability. This index is independent of decision biases, incentives, and so on. SDT allows the two decision parameters, diagnostic ability and decision tendency, which are two distinct sources of error, to be disentangled and quantified.

Second, ROC analysis produces an index of the decision threshold. This threshold is a product of the expectancies (prior probabilities) and utilities (costs and benefits) associated with decision outcomes. Given more precise knowledge of outcome probabilities

and the specific utilities for correct decisions and errors, this index can facilitate the specification of an optimal decision threshold.

Third, and relatedly, in using SDT it becomes possible to examine the covariation among decision outcomes as thresholds vary. That is, it answers this question: what is the impact of changing one's decision threshold (for concluding that the sample or evidence contains a true "signal") on the number of hits, misses, false alarms, and correct rejections? Decision outcomes stand in direct relationship to one another such that changes in subjective decision rules necessarily affect the balance among outcome probabilities.

While typical measures of accuracy confound discrimination ability with judgment and decisional factors, SDT permits the calculation of pure measures of accuracy that measure discrimination ability alone. Without an unaffected index of accuracy, inconsistent or variable accuracy rates could be interpreted as reflecting either poor proficiency or changing decision incentives. Without independent measures, there would be no way to disentangle such interpretations. Through ROC analysis, SDT allows for the calculation of an unbiased measure of discrimination ability, a measure uncontaminated by varying decision thresholds. Thus, SDT gives researchers the tools to avoid mistaking the effects of different decision tendencies and differing levels of competency for each other. This, in turn, opens the door to: (a) more efficient and effective discovery of factors associated with greater competence, (b) discovery of factors that affect decision thresholds, and (c) optimizing decision making by gaining control over the decision thresholds.

Beyond these three principal benefits of SDT, ROC produces a comprehensive picture of decision outcomes. This method provides measures and a graphical display of decisional accuracy across distinct threshold points. Using ROC, the abilities of forensic scientists as human measuring instruments, and the conditions that give rise to the greatest accuracy, can be more easily and clearly identified.

Potential Usefulness of Signal Detection Theory in Forensic Science

The distinction between discrimination capacity and decision thresholds, the factors that affect those, and the benefits of measuring them, can be as important in forensic science as they have been in the other fields that have used Signal Detection Theory to gain greater understanding of and control over their decision-making.

Forensic examinations of evidence commonly attempt to classify and quantify unknown substances, assist investigators in reconstructing events surrounding the commission of a crime, and associate (or disassociate) offenders and victims of crimes and their environments. Forensic science is similar to the scientific disciplines discussed in the foregoing sections of this paper in that forensic scientists often encounter ambiguous and murky decision-making situations. Examiners do not have the luxury of preparing their own evidence samples but must work with physical clues, often of limited quantity and inferior quality, left in the aftermath of crimes. The challenge to the forensic examiner is great, particularly when called upon to *individualize* evidence; that is, to compare evidence of unknown origin with a standard of known origin to determine if they share a common source. In some situations, the signal-to-noise ratio in a forensic comparison may be relatively high, as where a pristine fingerprint is found on a clean glass counter at a crime scene. In other situations it may be low, as where a firearms examiner compares striation marks on a badly damaged bullet with a bullet test-fired in the laboratory.

Thus, in many of the same ways that it has proved useful to other fields, SDT might be useful to the study and improvement of work in the forensic sciences, such as by evaluating the effectiveness of alternative technologies and procedures; testing the discrimination capacity of examiners separately from the decision thresholds they use; identifying the factors that lead to superior discrimination ability; identifying the influences on decision thresholds; and by setting explicit and optimal decision thresholds.

Sources of Uncertainty in the Process of Forensic Science Examination

The quest of the forensic examiner to individualize evidence sets him or her apart from other natural scientists. To do this, the examiner tries to find something(s) unique about an object that distinguishes it from all other similar objects. Samples may be subjected to a succession of measurements and placed in more and more restricted categories of similar objects. In order to graduate to a higher level of certainty, and ultimately attain an *individuality,* the examiner searches for specific characteristics that distinguish that evidence from material with the same class characteristics and, therefore, would make it unique. The forensic examiner must be able to distinguish more general class characteristics of materials from individual characteristics common to the evidence, its source, and no other. Conversely, when items are found to possess *dissimilar* class or individual characteristics, the examiner may conclude they do not share a common origin.

Rather than making a firm statement of common source, it is more likely the examiner will conclude the items *could have* shared, *probably* share, or are *consistent with* sharing a common origin. DeForest et al. (p. 7, 45) refer to such qualified conclusions as "partial individualizations". There are varying levels of individualization that may be reached as an examiner proceeds through a review and notes points of similarity or dissimilarity. However, with the exception of such areas as biological fluids where data are recorded on the distribution of an array of genetic markers in different populations, the forensic sciences possess little empirical data to assist examiners in interpreting the meaning of their test results and affixing a probability or confidence level to their findings. There are also specialties like fingerprints where professional organizations allow their members to offer nothing less than absolute identifications.

The ability of evidence to be individualized is determined, initially, by the *nature* of the evidence itself, and whether it bears distinctive features that can be identified, measured, and interpreted. Some objects may produce patterns that are easily decipherable while other mass-produced products may produce none. Fingerprints, for example, possess features (minutiae) of such intense variation that, singly or in combination, they are more readily distinguished from each other. Other kinds of evidence, such as handwriting, are less distinctive, so efforts to individualize are fundamentally more difficult. In addition, handwriting shows variability not only between individuals but within individuals (that is, each person's writing varies each time that person writes), further complicating the task of individualization. Still other evidence like paint, glass, soil, new tools, weapons, or other objects may possess few or no decipherable individualizing characteristics that can be used to differentiate that item from other similar ones or to associate that item with its source, and affords only a collection of similar class characteristics that show the evidence to be consistent, but not necessarily individualistic.

Independent of the inherent variability of a given type of evidence, the quality of the evidence varies. For example, a latent fin-

gerprint may be complete and clear, or it may be partial, smudged, and overlapping with other prints. The quality of evidence depends also upon steps the offender may have taken to eradicate the evidence that was created—wiping down surfaces at the crime scene to remove fingerprints or forcing a rape victim to bathe. The quality of evidence may also be compromised by inclement weather, curious onlookers, or by police investigators who fail to preserve it properly or contaminate it through faulty packaging. Delicate markings on bullets may be damaged upon removal from walls, trees, or human bodies. The examiner is faced with the daunting task of determining if certain markings were the result of the offender's actions, or if they occurred later, possibly inadvertently, by an investigator or evidence collector. Thus, even for evidence that has the advantage of intense variability, the "signal's" strength may be quite weak. Consequently, decisions must be made under conditions of uncertainty.

The potential for evidence to yield useful information is also influenced by the sensitivity and reliability of laboratory techniques. In effect, these techniques enhance the quality of the evidence. For example, the field of forensic serology has been greatly advanced in recent years by methods capable of characterizing the DNA of biological samples, and thereby capitalizing on the variability across individuals' DNA. The ability to match blood, semen, and other body tissue has also been bolstered by the building of population databases that have laid a statistical foundation that quantifies the likelihood of finding various DNA types throughout the human population. Other forensic methods for examining pattern evidence, including firearms and toolmarks, shoe prints, and handwriting, still rely principally on microscopic methods of analysis, although computerized imaging technology is now being used to store and retrieve information on bullets and cartridge cases, but is not used in making ultimate comparisons (p. 478, 46). Once these examiners are satisfied that both the questioned and standard evidence possess similar class characteristics, they seek out individual characteristics that appear unique to the particular gun, tool, or individual that produced the mark in question. Here, the examiner is challenged to visually (microscopically) compare the similarity of patterns and form an opinion about the individuality of the evidence. Ordinarily, the examiner does not have access to a database that assists in quantifying the rarity of the marks, or which even records them, but must rely on memory of other samples viewed in the past.

Because there are no standardized curricula in forensic science, much depends upon the training program at the particular laboratory, the person under whom the forensic examiner trained as an apprentice, and the trainer's particular approach and philosophy toward forensic comparisons. Because many of these examiners lack advanced academic or research training, and because the discipline as a whole lacks a common academic core, examiners do not share a similar scientific or theoretical basis for examining and interpreting the evidence in question. Subtle or not so subtle lessons are learned about expectations (prior probabilities) or utilities associated with correct associations versus false alarms, and correct versus erroneous exclusions.

Ultimately, evidence comparisons depend on the analyst finding a sufficient number of points of concordance (and no points of unexplained discordance) to satisfy some matching criteria that the examiner has established. Examiners often employ different criteria that are not published or, perhaps, not even articulated. Usually, the greater the number of points of correspondence, the greater the confidence in the match. Fingerprints are a good example where the counting of corresponding minutiae between the questioned evidence and known standard forms the basis for an individualization (or absolute identification). However, the statistical foundation for establishing fingerprint individuality is weak, examiners employ their own standards in forming conclusions, and some jurisdictions have adopted various different minimum (legal) thresholds (47).

Consider, also, the situation of the firearms examiner. Even bullets fired consecutively from the same firearm will exhibit dissimilarities. Although Rowe (p. 426, 48) states that "if both bullets were fired from the same barrel, numerous matching patterns will be readily evident", at the same time, Burrad observes that, "One of the most surprising things which must strike any observer who is examining fired bullets is the astonishing differences that seem to be present on bullets which are known to have been fired through the same barrel" (p. 380, 49). The firearms examiner must test fire several bullets and intercompare them in order to form a mental image of the striation patterns that are common to the various test firings. Examiners use this mental image as a basis for evaluating the degree of correspondence among the bullets known to have been fired from the same weapon and the degree of correspondence between those and a questioned, and often distorted, bullet. Thornton and Rios advise that just as "total accord" is not to be expected in the striation markings of bullets fired from the same gun, complete "absence of accord" is not to be expected from bullets fired from different weapons (50). In their training, examiners must gain an appreciation for the extent of striation matching that will be found in bullets fired from the same versus different guns. A match is declared when the extent of agreement between the test and evidence bullets exceeds that of the best known non-match (that is, comparison of bullets fired from different guns of the same type). Other authors make reference to these "intuitive criteria" that examiners must acquire through their training and experience (51), that the criteria for a match are elusive (50), and that there are "no objective, quantitative criteria for determining the individuality of toolmarks" (p. 145, 52).

Consider the similarities and differences of the work of the document examiner in comparing handwriting. Unlike other types of evidence, such as fingerprints, which do not change throughout one's lifetime, handwriting may change substantially in both systematic and random ways over both the short and long term, depending upon the writer's health, the speed of writing, the position of the writing surface, maturation, and so on. Comparisons are further complicated where a person attempts to copy another person's signature or possibly disguise his own. From a forensic standpoint, the quality of a handwriting sample also depends upon the quality and the quantity of the sample present. Is the evidentiary sample a single credit card receipt with a lone signature, or a series of checks the subject may have written over a period of many months? Typically, examiners will seek representative known/standard writings by subjects drawn from different sources the subjects are known to have written. Absent a sufficient number of known writings, subjects may be asked or compelled through court order to provide handwriting exemplars. The document examiner tries to get a feeling for the range of variation in the subject's writing by carefully examining the standards that have been provided. The quantity of the writing samples is important to establish the range of variation in particular letters. The risk of erroneous matches is increased when the quantity of samples is limited or where one author is attempting to simulate the writing of another. Ellen (53) states that the examiner would expect to find several consistent differences in the shape and formation of letters written by different individuals, but not in the natural writing of the same person. He notes handwriting examiners experience difficulty in estimating the frequency

of occurrence of differences, however, and must keep in mind that not all differences are independent of every other. Examiners apply varying significance to characteristics found, with greater weight usually given to those features that are uncommon or peculiar. The degree of perceived peculiarity, however, is largely a function of the training and caseload experiences of the individual examiner. In order to reach a conclusion that the questioned and known writings were made by the same hand, an examiner must, among other things, judge that the variation between the questioned and the known writing is no greater than the variation among the known standards.

Forensic individualization conclusions, then, typically are the product of an exercise of an examiner's judgment in response to an array of complex stimulus patterns. At the end of the day, the examiner must decide whether or not to reject a presumption of non-signal in favor of a conclusion that a signal has been detected.

Using SDT to Answer Research Questions

Our knowledge of the relative capabilities, limits, and potential of various forensic analytical methods in examining and individualizing evidence is at a rudimentary level. Clearly, it would be useful in the forensic context to be able to assess the various components of, and factors contributing to, forensic science decision-making and the level of accuracy it achieves. Following are examples of research questions that could be asked and answered with the help of research employing SDT.

Discrimination Ability

Until now, measures of decisional accuracy by forensic scientists have consisted of simple counts of examiners' correct and incorrect responses to known matches and known non-matches. The use of SDT would permit proficiency test data to measure pure ability separate from the confounding effects of decision thresholds. This would permit more powerful research on the ability levels of different examiners, labs, procedures, and technologies.

The discrimination ability of forensic scientists appears to vary across laboratories, examiners, and types of examinations. With SDT we can begin to pin down more precisely the nature of those differences. In addition to measuring and describing those variations, it would be interesting and useful to discover the factors that facilitate the acquisition of competence or enhance it. It would be informative to obtain more complete information about the academic backgrounds, training, and work experiences of the examiners completing proficiency tests. With the increased analytic control provided by SDT, it would be possible to assess the connection between examiners' backgrounds and their competence (after removing the effects of different decision thresholds) as well as to determine if background and training produce differences in the decision thresholds employed. Do examiners with different training or experience achieve different levels of discrimination accuracy, or are apparent differences merely the product of the use of different decision thresholds? Or are apparent non-differences the result of different thresholds masking real differences in ability?

For example, the finding that FBI document examiners do no better than laypersons at finding correct associations (54) might be found, using SDT analyses, to be the product of real, but offsetting, differences in discrimination skill and thresholds between experts and laypersons. Perhaps document examiners have superior discrimination skills than laypersons, but are more cautious in declaring a match. Conversely, underlying the finding of apparent superiority of examiners over amateurs in discriminating writing from different persons (54), we may find differences in decision thresholds but an absence of differences in discrimination accuracy.

Turning from the people to the technologies and methods, evaluations using SDT would facilitate choosing among new technologies (as has been done in medical imaging) or different examination techniques. Making these choices would benefit from having pure measures of the discrimination ability resulting from the use of one or another technology or adopting one or another examination protocol.

Decision Thresholds

As discussed earlier, accuracy is not a function merely of discrimination skill, but depends also on the location of decision thresholds. With the use of SDT, the part played by decision thresholds could be determined. Do they vary from one forensic science specialty to another? For example, one would expect to find that fingerprint examiners set uniquely stringent thresholds. Do examiners trained in different specialties, or by different mentors, make characteristically different uses of decision thresholds (some higher, others lower, some more stable, others varying)? Even within the same examiner doing the same type of examination, are there conditions under which thresholds vary, or do they remain constant across circumstances?

Do different labs have explicit norms, or cultures that more subtly produce examiners who use higher or lower thresholds, reflecting the values of those labs or the police departments or prosecutors they serve? Peterson et al. (55) found that laboratories in different cities varied greatly in terms of the percent of time their written laboratory reports actually disassociated the offender from the crime. On average, fewer than ten percent of all crime laboratory reports excluded the suspect from the crime scene or connection to the victim. Whereas one laboratory's reports would include a statement that the evidence and standard did not share a common origin (an exclusion), such a finding in the three other laboratories studied would more commonly be expressed as an "inconclusive" finding. While much of this variation reflected divergent policies with respect to reporting such exclusions, or divergent examination procedures, rather than differences in examiners' actual conclusions, some of these differences may also reflect differences in the decision thresholds employed by the personnel of these labs.

Prior experience can lead examiners to have expectations about the likelihood that evidence examinations will result in inclusions or exclusions. These differences in experience can lead to differences in prior probabilities, which in turn can lead to differences in the setting of subjective decision thresholds. Some limited data were gathered in the 1980s on the percentage of actual crime laboratory examinations which *typically* result in positives (inclusions) and negatives (exclusions) (55). Firearms evidence resulted in associations between evidence and a standard (bullets, cartridge cases, weapons) far more often (40 to 80%) than did other types of evidence. Fingerprints also ranked relatively high in associating offenders to their *personal* crimes, when compared with other evidence categories, though not as often in associating offenders with *property* crimes. (These differences in the likelihood of fingerprint identifications in different crime categories were due largely to the relatively indiscriminate collection of fingerprints in property crimes and a failure or inability of investigators to gather standard or known fingerprints from suspects to compare with the evidence prints.) Similarly, the success in using bloodstains to associate suspects with their *personal* crimes was five fold greater than its ability to connect offenders with their *property* crimes. This was pri-

marily explained by the fact that investigators supplied biological samples of known origin (standards) for personal crimes at a much higher rate than for property crimes. Moreover, for both fingerprints and bloodstains, the frequency of matching likely has been increasing over the past 10 to 15 years, owing to improved laboratory procedures for examining such evidence, and the ability to query computer automated databases in search of "cold" hits. These differing frequencies of successful matches between types of evidence, or between types of crimes for the same evidence, or changes over time, might result in different and changing expectations by examiners of the likelihood of an inclusion, given certain types of evidence, which in turn may cause them to shift their decision thresholds.

It is also important to remember that if tools, guns, and suspects were brought in at random for testing and comparison with crime scene evidence, the vast majority of them would be excluded. In other words, the random match probability is very small. Because evidence standards are not collected randomly, however, but are chosen by investigators who ordinarily have other evidence or indicators that a particular person or object may be involved, much of the evidence compared by examiners does in fact correspond with standards from suspects because the investigator has identified the true perpetrator. For example, the FBI has reported that in its case experience about one-third of comparisons performed by its DNA analysis section have been found to exclude the designated suspects. Although this rate of exclusion is substantially higher than Peterson et al. (55) found for laboratories using conventional serology, examiners in both situations most likely acquire an expectation of a high prior probability of a match, certainly much higher than if they examined evidence brought in with less selectivity. In comparable decision-making in other fields, prior expectations such as these have been found to lower examiners' decision thresholds, making them more likely to announce a "signal" than may be desirable.

With regard to the utilities associated with examination results, the prevailing forensic science norm is that the examiner must be very conservative about declaring a match or common origin between questioned evidence and materials of known source unless all (or a reasonable number of other) possibilities are explored and explained. The judicial environment and its presumption of innocence also send the message that it is better to err by wrongly exonerating a guilty person than by falsely convicting an innocent one.

On the other hand, there may be forces that create utility in the opposite direction. At the extreme, Moenssens has written that many experts are tempted to "fabricate or exaggerate" results. "All experts are tempted, many times in their careers, to report positive results when their inquiries come up inconclusive, or indeed to report a negative result as positive. . . ." (p. 17, 56). There is the anecdotal evidence of the recent U.S. Department of Justice Inspector General's report showing that pressure has been applied by investigators for examiners to rewrite their reports to favor a particular position (57), and also the example of individuals like Fred Zain. James Starr's (58) treatments of errant forensic examiners also suggests that the adversarial process can exert profound influence such that experts desire to help secure a "win" for the team by which they are employed.

In more subtle ways, the adversary process and the organizational identity of forensic scientists with police agencies may lead to a tendency to resolve doubts in favor of inculpation. Or, when an examiner is working on submitted evidence from a crime, and additional information about the case is made available to the examiner, suggesting a high or low likelihood of the suspect's guilt, expectancies are created that can affect decision thresholds. On the other hand, examiners may be so well trained, even if only implicitly, to use pre-set decision thresholds, that they may remain unaffected by learning about other inculpatory or exculpatory evidence. SDT could help in the study of such issues.

In practice, we also know that forensic examiners sometimes reach conflicting interpretations of the very same evidence. This is fairly common with evidence (such as handwriting) where there are no firm scientific or empirically verified criteria for making common origin judgments and where the expertise of examiners is largely a function of the person under whom they trained. Occasionally, too, the difference in expert judgment is influenced by the side in the litigation employing the expert. SDT provides both a basis for understanding why two equally skilled experts might reach different and equally sincere conclusions, and a method for testing whether that explanation is valid. An expert for the prosecution may use a lower threshold in forming a conclusion associating a suspect with a crime victim, while another expert, working for the defense, may employ a higher threshold and judge the evidence comparison to be inconclusive.

Finally, one could test some of the effects of proficiency testing methods. Do thresholds vary as a function of whether an examiner is taking a blind test versus a nonblind test? It may be that when examiners *know* they are being tested, they adjust their decision thresholds to minimize the likelihood of the most serious type of error in the eyes of the courts or society, namely, a false alarm (erroneous inclusion), but when working a normal case, they set their decision thresholds so as to reduce the risk of a miss (incorrect exclusion).

Developing Optimal, Explicit Decision Thresholds

To date, examinations in forensic science involve essentially intuitive and subjective judgments as to when the presumption of a nonmatch should be set aside in favor of declaring a match. Even considering only "objective" points of comparison, various subfields of forensic science have varying rules of thumb concerning when enough matching points exist, though, in practice, everything rests with the discretion of the individual examiner. Quality assurance procedures in laboratories that require positive associations to be verified by another examiner, or which randomly audit or check the results of completed cases, are procedures intended to control such discretion.

In firearms and toolmarks examination, Biasotti and Murdock (p. 150, 52) concluded that among known nonmatching bullet comparisons (bullets fired from different weapons), typically no more than three consecutive corresponding striae were found. The overall percent of matching striae, which varied between 15 and 30% for nonmatching comparisons, was deemed of limited value as an identification criterion. Their conservative criteria for a matching identification were finding "at least two different groups of three consecutive matching striae in the same relative position, or one group of six consecutive matching striae." This makes plain that other criteria (some less conservative and others even more conservative) could be employed, and these variations can lead to different conclusions by different examiners.

Before a document examiner can reach a conclusion of common origin "a sufficient number of individual characteristics must be present in both questioned and known writing," and "when considered in combination with each other," are sufficient to conclude the writings were executed by the same person (p. 370, 45). Most docu-

ment examiners state that the "number of significant and consistent characteristics" that must be present "can be taught only by experience" (p. 700, 59). "There is no universally accepted number of points of similarity that the examiner must find before offering an opinion that two writings are identical" (p. 151, 60). Again, it is clear that decision thresholds that vary with the examiner can account for differences in the decisions that document examiners make.

Even in the field of fingerprints, where examiners in other countries have set varying numbers of points of agreement as the threshold amount of similarity required in order to declare a match (e.g., 16 in Great Britain, 12 in Austria), in the United States, "[t]he criteria for absolute identification . . . are wholly dependent on the professional judgment of a fingerprint examiner. When a fingerprint examiner determines that there is *enough* corresponding detail to warrant the conclusion of absolute identification, then the criteria have been met" (p. 71, 47).

Not only does the preceding discussion make clear that decision thresholds can vary considerably from one examination to another, it suggests the desirability of developing standardized thresholds. In other fields, Signal Detection Theory has been quite helpful in work aimed at developing explicit, optimal decision thresholds, which balance the desire for maximum hits against the desire to minimize false alarms. In forensic science, SDT could be an important tool in the development of guidelines that help achieve optimal decision thresholds.

To develop an optimal decision threshold, one needs to consider two important factors: the likelihood of "matching" evidence (in the case of forensic identification) and the utilities of all possible decision outcomes. In the formula below, taken from Swets (43), we see that calculation of an optimal decision point requires the specification of prior probability estimates and the assignment of numerical values to the costs of errors and the benefits of correct responses. In this formula, S_{optimal} represents the slope of the "best" ROC curve, where accuracy is high and the tradeoff between the number of hits and false alarms best suits the decision task at hand. In the first part of the equation, the expected probabilities of true non-matches and true matches are represented, with the latter value in the denominator. In the second part of the equation, costs are subtracted from benefits in cases where the evidence truly matches and truly does not match.

$$S_{\text{optimal}} = \frac{P_{\text{non-match}}}{P_{\text{match}}} \times \frac{(B_{CR} - C_{FA})}{(B_{\text{HIT}} - C_{\text{MISS}})}$$

How can these values be determined such that they are both reliable and accurate? Certainly there is no objective "right" answer. Swets argues that though determining these values is inherently challenging, any decision threshold reflects a consideration of these factors and tentative assumptions of what their respective values might be. By making the assumed values explicit, the decision maker can address his or her implicit assumptions and consider how these assumptions might affect decision outcomes.

As discussed earlier, examiners likely have a relatively high expectation of a "match," such that the perceived probability of a match will exceed the probability of a nonmatch. Examiners should also consider the utility of each decision outcome. For example, given a true "match," it is likely that the benefits of a hit exceed the costs associated with a miss. (This may, however, also depend on the severity of the crime under investigation.) Given a true "nonmatch," the greatest disutility in our legal system would be a false alarm. The cost of a false alarm is therefore no doubt much higher than the benefit gained from a correct rejection. Considerations of these kinds facilitate the development of the most optimal decision thresholds.

Illustrative Analyses of the Application of SDT to Forensic Science Data

In this section we illustrate the uses of Signal Detection Theory by applying it to some forensic science proficiency testing data. [Concerning proficiency testing in forensic science, see Peterson et al. (61) and Peterson and Markham (62,63).]

Method

Two different proficiency tests are examined, one on firearms identification (64) and one on handwriting identification (65). Both test examples involve discrimination tasks where examiners are asked to compare evidence samples and determine whether the evidence found at a crime scene matches evidence taken from suspect(s). Labs are given three options for responding to each evidence comparison: yes, it's a match; no, it's not a match; or inconclusive (that is, cannot conclude whether it is a match or a non-match). Though several examiners may be involved in an examination at a particular lab, each lab provides a single response to each comparison question.

For analysis, the two reports were considered independently, and for each report comparison decisions were tallied across all participating labs (for that test year). Each lab's responses were treated as individual decisions. So, for example, if 40 labs each responded to three comparisons, 120 decision outcomes were tallied. Thus, the total number of decisions analyzed was a joint product of the number of labs participating multiplied by the number of comparisons made by each lab. Two different indices were calculated for each report: discrimination ability and decision tendency. An ROC curve for each report's data was graphed by plotting and connecting two decision thresholds. Discrimination ability or perceptual "accuracy" is represented geometrically as the area under the ROC curve. A decision tendency was also calculated using outcome proportions. This index is an overall estimate of how the participating labs are likely to respond when evidence samples do not, in reality, match.

Two specific analytic techniques were utilized. (See Appendix 2 for a more detailed explication of how the analyses were conducted.) First, the three possible responses (to each comparison) were treated as ratings of confidence that two samples being compared shared a common origin—with "Yes" expressing the greatest confidence, "Inconclusive" taken to reflect some moderate degree of confidence, and "No" expressing the least confidence. Responses were tallied and converted into two decision points using the "Confidence Rating" procedure for ROC curve generation [see e.g., Dorfman and Alf (66) and Gescheider (67), for explication of this procedure]. One of the two decision points represents a strict threshold; this threshold produces a low false alarm rate. The other point represents a lenient threshold, which results in a higher hit rate along with an increased false alarm rate. As discussed earlier, these two decision points are used by examiners with the same degree of discrimination ability, yet they produce different rates of correct and incorrect decisions. That is, discrimination ability is the same across both of these decision thresholds.

Second, as noted, all responses were treated as individual decisions and pooled together for analysis (within each dataset). Due to potential problems with data pooling, a procedure known as "jackknifing" was also implemented. To elaborate briefly: typically, ROC curves are generated for each individual examiner responding to a variety of stimuli. Because forensic science proficiency data are conducted one test at time (annually) and the responses of any given lab cannot be linked because the labs respond anonymously,

analyses had to be adapted to the nature of the available data. We had to treat each lab as a separate observer. But with the usual analysis method, this would translate into 50 different firearms ROC curves—one for each lab. With 50 different ROC curves, each one would provide very little data. Instead, we have pooled the responses given by the 50 labs into a single ROC curve (with the number of decisions equaling the number of labs times the number of decisions made by each lab). Though pooling increases the richness of the data, pooled data run the risk of distortion (e.g., strange or unusual observations from one or two labs may distort the pooled data). The jackknifing procedure is designed to reduce such possible biases. Jackknifing involves a kind of weighting of each lab's responses so that no one lab can artificially increase or decrease the pooled variability of all participating labs. This procedure was applied to all analyses conducted herein. Dorfman and Berbaum (68) provide a technique that allows for SDT analysis of pooled, rating-method (more than two response alternatives) data. The jackknifing procedure was originally introduced by Quenoille (70) and was then utilized more generally by Tukey (71). See Dorfman and Berbaum (68) and Dorfman, Berbaum, and Metz (69) for a discussion and a thorough explanation of the procedures and benefits of the jackknifing procedure.

Firearms Identification Data from FSF Test 1985-3

In this first example, we analyze the firearms identification proficiency data. This test required that examiners compare a questioned cartridge case against three known evidence samples. The test was originally distributed to 81 laboratories, of which 50 provided responses. The instructions were as follows:

> One of your submitting agencies is investigating a murder case where an individual was shot and killed from ambush. The two bullets passed through the body and were not recovered. Two fired cartridges were recovered at the crime scene. Extensive investigation developed a suspect who is known to have had the same type of weapon as used in the murder. The weapon has disappeared but it was determined that the suspect used the weapon for target practice at two or three areas in the county. These areas were searched and additional fired cartridge cases recovered. The prosecuting attorney believes that it will be of value to the case to "link" the cartridge cases from the murder scene to those from an area where the suspect is known to have fired his weapon. You are asked to examine the evidence.
>
> Exhibit 1—two fired cartridge cases from the scene of the crime
> Exhibit A—two fired cartridge cases from the first target shooting area
> Exhibit B—two fired cartridge cases from the second target shooting area
> Exhibit C—two fired cartridge cases from the third target shooting area
> Could Exhibits A, B, or C have been fired from the same weapon as Exhibit 1?

Each of the 50 labs responded either "yes," "no," or "inconclusive" to the three comparison questions (comparing Exhibit 1 to Exhibits A, B, and C). Response outcomes were tallied for all 150 decisions. Using a conventional measure of accuracy, 137 out of 150 decisions were correct, suggesting an accuracy rate of 91%. Using ROC analysis, however, we can assess accuracy as a function of discrimination ability alone and assess whether (and how much) subjective decision thresholds vary by task. In addition, ROC provides an estimate of decision tendency that suggests how labs are likely to respond to new choices, given their past performance.

For analysis, two decision points were calculated and plotted (see Fig. 3). The first decision threshold, Point A, represents the strictest threshold for responding with an affirmative "match" response. With this conservative criterion, false alarms are kept at a minimum. Even with this strict criterion for declaring a match, the labs performed with a high hit rate of 0.95 and a very low false alarm rate of 0.01. (The observed false alarm rate in this dataset was zero. In order to calculate the necessary values for the analysis, the false alarm rate had to be set slightly above zero. See Appendix 2.) Point B represents a more lax criterion, where the labs were more likely to respond "yes." Using this threshold, the hit rate was an extremely high 0.99, but this more lenient criterion also produced an increased false alarm rate of 0.16. Note that the labs showed exceptional discrimination ability on this test at 0.99. Importantly, this high level of ability is the same for both plotted decision points. Thus, the differences between Points A and B are due to changes in subjective decision rules, not to fluctuating discrimination ability. Moreover, recall that the conventional index of accuracy (percentage of all responses that were correct responses) found a lower estimate (91%). This is because the conventional measure cannot distinguish errors produced by lack of discrimination ability from errors produced by varying decision thresholds.

An estimate of overall decision tendency also was calculated from response outcomes. Given evidence samples that are, in fact, nonmatches, participating labs were likely to provide incorrect affirmative responses 1% of the time. By contrast, these labs were likely to determine nonmatches accurately for 80% of their decisions. (Note that these percentages do not sum to 100 because a third option, "Inconclusive," was also available.)

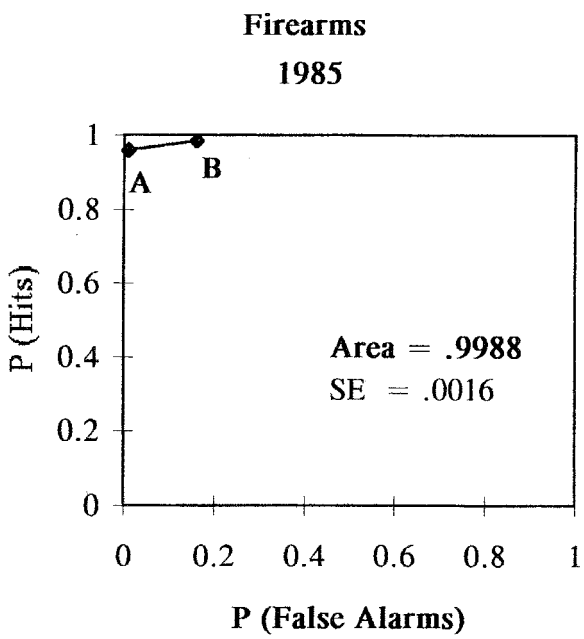

FIG. 3—*ROC graph of decision outcomes for firearms test.*

Questioned Documents Analysis Data from FSF Test 1987–5

For this test, forensic document examiners were asked to compare a questioned extortion note with known writings from four suspects, one of whom had in fact written the extortion note. According to the test manufacturer, "This test was designed to be a relatively easy and straightforward test. . . . All the writings in this test were natural and free of disguise." Each lab provided a response for each of the four comparisons. Of 55 subscribing labs, 33 returned reports of their examinations. One lab, however, gave "no opinion" responses to all questions and therefore was excluded from our analysis. As a result, 128 decision outcomes were tallied and analyzed.

Note that in contrast to the firearms test, where each comparison could have led to a decision of "yes" or "no" without affecting the other decisions, the structure of this handwriting test does not present independent decisions to the examiners. Once one suspect was judged to have been the author of the questioned document, all of the others would necessarily be judged not to be the author. Thus, the poorest possible performance—erroneously identifying a nonauthor as having been the author (thereby excluding the actual writer and inculpating an innocent writer, leaving the two remaining suspects to be scored as correct)—would be 50% correct. (In fact, only 52% of the examiners correctly identified the writer of the questioned note, while 45% reached "inconclusive" results.) Thus, treating each comparison as if it were independent of the others can be quite misleading. Nevertheless, for present illustrative purposes, we do treat these decisions as if they were independent of each other.

As seen with the firearms report, sheer discrimination ability is quite high at 0.98. Again, this level of ability is constant for the two threshold points generated (Fig. 4). Point A is the more conservative threshold, where false alarms were quite low (0.005). But the tradeoff resulting from this decision threshold is that the proportion of hits was only 0.63. Point B represents a more lenient criterion, where more affirmative responses are made. Notably, due to the change in threshold, the hit rate increased markedly to 0.97, while the false alarm rate increased to 0.25. Note how ROC analysis allows us to see how greatly decisions vary as a function of where examiners place their decision thresholds, even while their raw discrimination ability remains quite high.

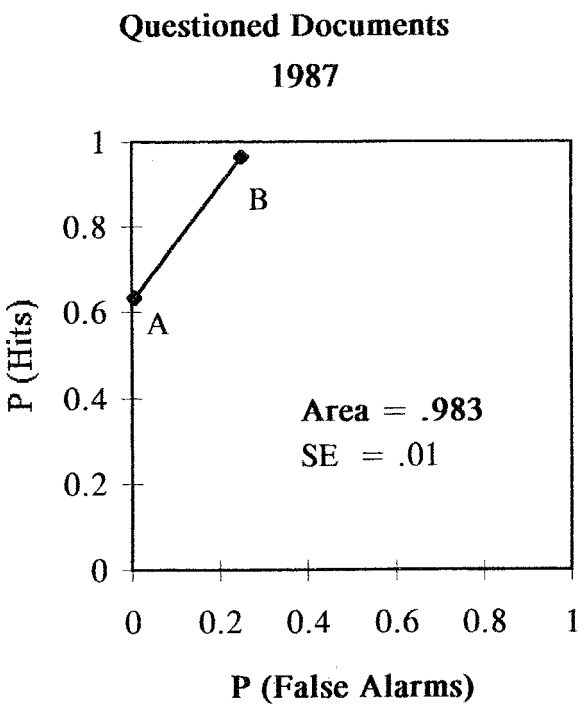

FIG. 4—*ROC graph of decision outcomes for QD test.*

The estimate of decision tendency suggests that, overall, the labs are likely to respond correctly to a true nonmatch comparison (a "No" response) on 75% of such decisions. In addition, in this problem, the labs showed virtually no tendency (<0.0001) to make erroneous affirmative responses to true nonmatches.

Using the conventional accuracy measure (the percentage of correct responses out of total responses), accuracy on this test was 71%. By removing the effects of varying decision thresholds, the estimate of raw discrimination ability increased to 0.98. Thus, in this example, fluctuating decision thresholds mask high discrimination ability. (Conversely, in other situations, differences in decision thresholds also can mask low discrimination ability.)

Comparing the Two ROC Curves

Visual examination of the two ROC curves suggests that discrimination ability is quite high in both, as indicated by similar values for the area under the curve. The apparent superiority of firearms examiners, or firearms examination tasks, in terms of discrimination ability, is quite small.

More interesting, the differing slopes of the respective ROC curves reflect the use of different decision criteria. Why the differing slopes if accuracy is nearly the same? In the previous section we introduced the idea of an optimal decision threshold: the slope of an optimal decision point is a joint function of the perceived benefits of a correct decision and costs of an incorrect decision as well as the perceived likelihood that the crime scene evidence does indeed match (or not match) evidence taken from the suspect. In these data, at least, the firearms examiners have a shallower slope (0.17), indicating a lower threshold for declaring a match, while the document examiners have a steeper slope (1.4), suggesting a higher threshold for declaring a match. These differences, in turn, suggest different perceptions of the risk of error attending their different examinations (at least under the conditions of nonblind proficiency testing).

More generally, if these two tests were representative of the performance of firearms examiners and document examiners, the data of these ROC analyses would suggest that the greater general accuracy of firearms examiners (whose hit rate was higher and false alarm rate was lower) is due not to superior discrimination ability but the use by firearms examiners of more consistent decision thresholds. This, in turn, suggests that improvements in the performance of document examiners are more likely to come from studies of factors that affect decision thresholds, such as motivation, expectations, stakes, and so on.

The discussion in the preceding paragraphs has relied on an "eyeball" examination of the current data, rather than a formal statistical comparison of the two curves. Various limitations in the data prevent the application of statistical methods that assume equal variance. For further discussion, see Conclusions and Suggestions.

Conclusions and Suggestions

Though Signal Detection Theory is widely used in research in fields as varied as aviation, psychiatry, and radiology, to date it has barely been employed in forensic science. This article has reviewed

the basic concepts of SDT, has discussed various aspects of forensic science theory and practice that likely could benefit from the application of SDT, and has illustrated SDT analysis using forensic science proficiency data.

With the exception of a few forensic specialties, such as DNA typing, there have been few sustained efforts to evaluate the accuracy of forensic laboratory methods for analyzing various types of evidence and the examiners who use them. Under pressure from several quarters, most notably the courts following the U.S. Supreme Court's decision in *Daubert* v. *Merrell Dow Pharmaceuticals* (72), other forensic science specialties are beginning to subject their methods to systematic empirical testing (73). The tools provided by SDT would expedite the development of the needed research.

SDT is particularly well suited to the study of decision-making where the presence or absence of a "signal" has to be discerned amid a complex array of other, ambiguous stimuli. The correctness of an examiner's judgment is a function of both (1) the examiner's discrimination ability (or sensitivity to the evidence) *and* (2) the decision criterion (or decision threshold) employed by the examiner. A traditional single measure of accuracy, not disaggregated into its components, is incomplete and biased by inherent psychological factors, such as expectancy and motivation. SDT allows researchers to disentangle the two factors. Generally, a technician's ability to discriminate among stimulus inputs is rather stable. The decision criterion, however, is highly variable. The ROC curve graphically presents these two essential pieces of information which determine decision outcomes: (1) the area under the ROC curve is an unbiased indicator of pure discrimination ability, and (2) the points on the ROC curve represent the decision thresholds. A useful way to make comparisons between these two fundamental components of decisions is to compare the ROC curves generated when examinations are made under differing conditions.

SDT would facilitate numerous informative lines of research in forensic science. For example, from existing forensic science proficiency data, more clear answers could be found concerning the performance of participants in those studies. Most notably, raw diagnostic skill could be separated out from the confounding effects of decision thresholds. The relative accuracy produced by different protocols and technologies could be more clearly evaluated. The abilities of examiners with different backgrounds, training, and experience could be more completely and clearly assessed. The forces that cause decision thresholds to rise, to fall, or to remain stable could be identified and optimal decision thresholds could be more rationally and effectively determined.

Few efforts, past or recent, have attempted to look at the discrimination accuracy of specific examination methods while controlling for such factors as the technical ability of the analyst, the quality of the evidence, or the decision threshold used by the examiner. Nor have studies attempted to investigate one of the most nettlesome issues of forensic science, which is how the organizational role of an expert (government examiner or defense expert) may influence the decision thresholds employed in deciding whether or not items of evidence share a common origin. In that context, or even within the same laboratory, when two experts arrive at different results, we lack an understanding of the mechanisms by which two equally technically competent examiners can arrive at different results. Consequently, questions persist about the dependability of some specialty areas. SDT may provide clarifying insights into these and other problems.

Existing forensic science proficiency testing data have a number of limitations. Because they derive from a program that is largely voluntary, all laboratories do not subscribe and provide data. Some commentators suggest that this self-selection skews the results toward an exaggerated appearance of accuracy because the poorer laboratories choose not to participate. Laboratories periodically are asked for information describing methods used, the qualifications of examiners, and time spent on examinations, but these data have not yielded particularly useful information. The level of difficulty of examinations is not controlled from test to test and (without the use of SDT) it is impossible to attribute variations in results to the difficulty of the test, the competence of the examiner, or the implicit choice of decision thresholds. While Peterson and Markham (62,63) recently reviewed 13 years of proficiency data, those data are not available in the published literature, nor are the raw data openly shared [in contrast to the traditional norm in science (74)], so it is impossible to track results reliably over time, or to conduct secondary analyses on them. The adoption of SDT obviously cannot solve all of these problems, but it could be a useful adjunct to existing or improved programs of research in and on forensic science.

To gain the most effective and efficient benefits of SDT analysis, several steps would be advisable. Research on decision-making in forensic science should begin to employ measures that lend themselves more readily to ROC analysis. One such change is to have examiners in proficiency studies rate the clarity of the "signal" they perceive on a scale, instead of (or in addition to) the dichotomous or trichotomous categories found in most proficiency studies. (For example, the technicians in the pathology laboratory illustration presented earlier in our article used a 7-point rating scale.) A confidence rating procedure such as this is advantageous because the estimates of accuracy are statistically reliable, it is sensitive to time constraints, and it enables examiners to make use of different decision thresholds simultaneously (as most decision makers do in real-world decision making) (44). Most important, the use of confidence rating procedures would help researchers to efficiently and reliably measure the varying thresholds utilized by examiners. Second, the analyses would be more informative and more powerful if data from the same individuals or labs could be linked over time, as is commonly done in similar research in other fields. This would enable research to examine decision factors across time, as techniques and technology develop and change across the sub-fields of forensic identification. Also, the data would be more amenable to comparative analysis among types of evidence, types of cases, levels of expertise, and so on. Finally, it would also be possible to change decision factors, such as perceived probability or perceived costs, and then assess the effects of threshold changes on the trade-offs among outcomes. Outcomes could be compared within and between labs given knowledge of the actual response distributions.

Although, as we illustrated in this article, somewhat heroic statistical manipulations can put existing data to use, more precise information could be obtained far more easily if scaled responses were obtained, and if the performance of the same individuals or labs could be linked and compared in response to different tasks.

Signal Detection Theory is a research tool that is so well established in so many different fields, and is so well suited to the fundamental task of forensic science, that it inevitably will come to be employed as a matter of routine by those who conduct research on forensic science decision-making. The goal of this article has been to hasten that day by introducing the fundamentals of SDT to the forensic science community.

Acknowledgments

The authors wish to thank Professor Donald Dorfman for his advice concerning the application of SDT methods to the available proficiency study datasets, Professor Lola Lopes for the cancer laboratory illustration, and Professor Jay Christensen-Szalanski for the initial suggestion of an article such as this.

APPENDIX 1

TERMINOLOGY

Signal—Information about the object of interest. For example, in radar detection, the signal might be an aircraft or a radio communication.

Noise—In the context of radar detection, the term noise referred to "white noise," or interference that could mask or even mimic a true radar signal.

Diagnostic accuracy—In the context of SDT, diagnostic accuracy is a function of two decision factors: raw discrimination ability and the threshold adopted for declaring a positive. In the text, for these two decision factors we use the terms discrimination ability and decision threshold. Synonyms for discrimination ability include: discrimination capacity, ability, diagnostic ability.

Prior probability—The perceived likelihood that a signal will be present or that the evidence will match.

Utility—The perceived benefits of correct decisions (hits and correct rejections) as well as the perceived costs of erroneous decisions (misses and false alarms).

Decision outcomes—The correctness or incorrectness of every decision made in relation to some true criterion. Every decision outcome is one and only one of the following:

Hit—correct identification; a true positive.

Miss—failure to identify; a false negative.

False Alarm—incorrect identification; a false positive.

Correct Rejection—correct non-identification; a true negative.

ROC—An acronym for Receiver Operating Characteristic or Relative Operating Characteristic (the terms receiver and relative are used synonymously). The ROC curve is an analytic tool developed in SDT, which provides graphical and statistical representations of important decision factors: discrimination ability (area under the ROC curve) and decision thresholds (points plotted on the ROC curve or slope of the ROC curve).

APPENDIX 2

ROC Analysis: The Confidence Rating Procedure

In this section we explicate our analytic strategy using the firearms data. As stated in the text, 50 labs participated in the firearms examination and each provided responses to three comparison questions: do any of the three known pairs of cartridge cases match the pair of cartridge cases from the crime scene? In actuality, two of the three unknown pairs did share a common source with the unknown pair. For analysis, we tabulated all 150 responses (three different comparisons by each of 50 labs) by the ground truth of the comparison. As shown in Appendix Table 1, below, when the cartridge did not, in reality, share a common origin ($n = 50$), no labs made an affirmative response. Eight labs provided an inconclusive response and 42 labs made the correct "no" response. For comparisons in which the cartridge did, in reality, match, 95 labs correctly provided affirmative responses, four labs provided inconclusive responses, and one lab provided an incorrect negative response.

The response frequencies were then cumulated from right to left, as shown in the Table 2. In cumulating frequencies, we were able to identify two different decision thresholds. Using the confidence rating procedure, each decision was treated as either a "yes" or "no" response. For clarification, if five response options had been utilized by the examiners, such as a scale with "definite match," "probable match," "possible match," "probable nonmatch," and "definite nonmatch" as alternatives, $n - 1$ or 4 decision thresholds could be calculated. In this case, the most strict threshold would include only "definite matches" in the "yes" response category. The next, less strict threshold, would include both "probable matches" and "definite matches," followed by a threshold that would include "possible matches," "definite matches," "probable matches," and so on.

Table 1 gives the raw firearms data. The three response options yield $n - 1 = 2$ decision points. The first decision point or cutoff includes those cells in Table 2 that are more darkly shaded. With this decision threshold, only the clear matches (as perceived by the examiner) would be identified as such. Thus, the "yes" column represents the cumulative frequency of response outcomes using a strict decision threshold. The second threshold, including all four shaded cells, represents a comparatively lenient decision threshold where "inclusives" are combined with the "clear" matches.

Cumulated frequencies were then converted to response proportions (Table 3). These proportions represent hit and false alarm rate pairs for two different decision thresholds. The hit rate is equal to the cumulated frequency of "yes" responses divided by the total number of true matches. The corresponding false alarm rate is

TABLE 1—*Response frequencies (using the firearms data).*

	Examiner Opinion Concerning Whether the Samples Share a Common Origin			
Ground Truth	No	Inconclusive	Yes	Totals
Non-match	42	8	0	50
Match	1	4	95	100

TABLE 2—*Cumulated response frequencies (using the firearms data).*

	Examiner Opinion Concerning Whether the Samples Share a Common Origin			
Ground Truth	No	Inconclusive	Yes	Totals
Non-match	42	8	0	50
Match	1	4	95	100

TABLE 3—*Response proportions: hit and false alarm rates (using the firearms data).*

	Examiner Opinion Concerning Whether the Samples Share a Common Origin		
Ground Truth		Point B	Point A
Non-match (FA)	1.00	0.16	0.01
Match (HIT)	1.00	0.99	0.95

equal to the cumulated frequency of "yes" responses divided by the total number of true non-matches (the false alarm rate for point A was adjusted using the formula $n/2\text{-}1$). Point A depicts the hit and false alarm proportions using the more strict decision criterion for declaring a match. Point B shows the hit and false alarm proportions using the more lax decision criterion. These two points were plotted and connected to form the ROC curve. (Note that in order to perform the necessary calculations, the zero false alarm rate when the criterion was at Point A had to be replaced with a small positive number, here 0.01.)

An unbiased measure of accuracy was computed by converting the above proportions to standardized scores (Z scores or the inverse of the standard normal cumulative distribution) and calculating slope and intercept values. A standardized accuracy score was then calculated using the following formula (where a = intercept and b = slope):

$$Z_A = \frac{a}{\sqrt{1 + b^2}}$$

The quotient was converted to the index A_Z (using the normal cumulative distribution). A_Z or area represents an unbiased measure of discrimination ability and is equal to the area under the plotted ROC curve. In this firearms example, discrimination ability was quite good at 0.999.

An overall decision tendency was calculated using the false alarm proportions for both decision thresholds. This measure provides an estimate of response bias in cases of a true non-match. The tendency to say "no" was calculated using the false alarm rate for the more lenient threshold minus 1. The tendency to say "yes" incorrectly was obtained directly from the false alarm rate associated with the more strict decision threshold.

Jackknifing Procedure

A jackknifing procedure was also utilized in the current analyses. Jackknifing involved computing "pseudovalue" measures of accuracy and averaging these values into a single A_Z score. The above procedures, frequency tabulation, cumulation of frequencies, and calculation of response proportions were conducted $n - 1$ times (with n = number of participating labs), to produce 49 pseudovalues of accuracy and response tendency. That is, to reduce possible biases introduced by individual labs on the pooled responses, these measures were calculated 49 times using responses from 49 labs, that is, each time with one of the labs removed. The final indices of accuracy and response bias reported are averages over 49 pseudovalues. A description of this procedure is provided by Dorfman and Berbaum (68).

References

1. Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley, 1966.
2. Eubanks JL, Killeen PR. An application of signal detection theory to air combat training. Hum Factors 1983;25:449–56.
3. Colquhoun WP. Sonar detection as a decision process. J Appl Psychol 1967;51:187–90.
4. Colquhoun WP. Effects of raised ambient temperature and event rate on vigilance performance. Aerospace Med 1969;40(4):413–7.
5. Swets JA. Signal detection theory applied to vigilance. In: Mackie RR, editor. Vigilance: relationships among theory, physiological correlates and operational performance. New York: Plenum, 1977;705–18.
6. Williges RC. The role of payoffs and signal ratios in criterion changes during a monitoring task. Hum Factors 1971;13:261–7.
7. McNicol, D. A primer of signal detection theory. London: Allen & Unwin Ltd., 1972.
8. Bonnel A, Noizet G. Application of signal detection theory to perception of differences in line length. Acta Psychol (Amst) 1979;43:1–21.
9. Swets JA. The relative operating characteristic in psychology. Science 1973;182:990–9.
10. Swets JA, Birdsall TG. Deferred decision in human signal detection: a preliminary experiment. Perception Psychophysics 1967;2:15–28.
11. Banks WP. Signal detection and human memory. Psychol Bull 1970;74:81–6.
12. Koppell S. Decision latencies in recognition memory: a signal detection theory analysis. J Exp Psychol Hum Learn Mem 1977;3:445–57.
13. Lockhart RS, Murdock BB. Memory and the theory of signal detection. Psychol Bull 1970;74:100–9.
14. Murdock BB. Recognition memory. In: Puff CR, editor. Handbook of research methods in human memory and cognition. New York: Academic Press, 1982;2–26.
15. Levi K. A signal detection framework for the evaluation of probabilistic forecasts. Organizational Behav Hum Decision Processes 1985;36: 143–66.
16. Craig A. Signal detection theory and probability matching apply to vigilance. Hum Factors 1987;29:645–52.
17. Jerison HJ. Signal detection theory in the analysis of human vigilance. Hum Factors 1967;9:285–8.
18. Thompson SC. Detection of social cues: a signal detection theory analysis. Personality Soc Psychol Bull 1978;4:452–5.
19. Berbaum KS, Franken EA, Dorfman DD, Barloon T, Ell SR, Lu CH, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986;21:532–9.
20. Berbaum KS, Dorfman DD, Franken EA. Measuring observer performance by ROC analysis: indications and complications. Invest Radiol 1989;24:228–33.
21. Mackinnon WB, Barry PA, Malycha PL, Gillett DJ, Russell P, Lean CL, et al. Fine-needle biopsy specimens of benign breast lesions distinguished from invasive cancer ex vivo with proton MR spectroscopy. Radiology 1997;204:661–6.
22. Swets JA. ROC analysis applies to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109–21.
23. Swets JA, Pickett RM, Whitehead SF, Getty DJ, Schnur JA, Swets JB, et al. Assessment of diagnostic technologies. Science 1979;205: 753–9.
24. Fombonne E. The use of questionnaires in child psychiatry research: measuring their performance and choosing an optimal cut-off. J Child Psychol Psychiatry 1991;32:677–93.
25. Loke WH. Diagnostic evaluations using signal detection analysis. Indian J Psychol Med 1989;12:87–91.
26. Mossman D, Somoza E. ROC curves, test accuracy, and the description of diagnostic tests. J Neuropsychiatry Clin Neurosci 1991;3:330–3.
27. Somoza E, Mossman D. "Biological markers" and psychiatric diagnosis: risk benefit balancing using ROC analysis. Biol Psychiatry 1991;29:811–26.
28. Wyshak G, Barsky AJ, Klerman GL. Comparison of psychiatric screening tests in a general medical setting using ROC analysis. Med Care 1991;29:775–85.
29. Kullendorff B, Nilsson M, Rohlin M. Diagnostic accuracy of direct digital dental radiography for the detection of periapical bone lesions: overall comparison between conventional and direct digital radiography. Oral Surg Med Oral Pathol Oral Radiol Endod 1996;82: 344–50.
30. Versteeg CH, Sanderink GCH, Lobach SR, van der Stelt PF. Reduction in size of digital images: does it lead to less detectability or loss of diagnostic information? Dentomaxillofacial Radiol 1998;27:93–6.
31. Abraham SC, Furth EE. Receiver operating characteristic analysis of serum chemical parameters as tests of liver transplant rejection and correlation with histology. Transplantation 1995;59:740–6.
32. Yong WH, Southern JF, Pins MR, Warshaw AL, Compton CC, Lewandrowski KB. Cyst fluid NB/70K concentration and leukocyte esterase: two new markers for differentiating pancreatic serous tumors from pseudocysts. Pancreas 1995;10:342–6.
33. Helstrom CW. Quantum detection and estimation theory. New York: Academic Press, 1976.
34. Kassam SA. Signal detection in non-Gaussian noise. New York: Springer-Verlag, 1987.
35. Poor HV. An introduction to signal detection and estimation. 2nd ed. New York: Springer Verlag, 1994.
36. Whalen AD. Detection of signals in noise. New York: Academic Press, 1971.

37. Ben-Shakhar G, Lieblich I, Bar-Hillel M. An evaluation of polygrapher's judgements: a review from a decision theoretic perspective. J Appl Psychol 1982;67:710–3.
38. Szucko JJ, Kleinmuntz B. Statistical versus clinical lie detection. Am Psychol 1981;36:488–96.
39. Bolt R, et al. On the theory and practice of voice identification. Washington: National Academy Press, 1979.
40. Friedman M. Introduction to pattern recognition: statistical, structural, neural, and fuzzy logic approaches. River Edge, NJ: World Scientific, 1999.
41. Whittaker DK, Brickley MR, Evans L. A comparison of the ability of experts and non-experts to differentiate between adults and child human bite marks using receiver operating characteristic (ROC) analysis. Forensic Sci Int 1998;92:11–20.
42. Andersen L, Wenezel A. Individual identification by means of conventional bitewing film and subtraction radiography. Forensic Sci Int 1995;72:55–64.
43. Swets JA. The science of choosing the right decision threshold in high-stakes diagnostics. Am Psychol 1992;47:522–32.
44. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press, 1982.
45. DeForest PR, Gaensslen RE, Lee HC. Forensic science: an introduction to criminalistics. New York: McGraw Hill, 1983.
46. Saferstein R. Criminalistics: an introduction to forensic science. 6th ed. Upper Saddle River, NJ: Prentice-Hall, 1998.
47. Stoney DA. Fingerprint identification. In: Faigman DL, Kaye DH, Saks MJ, Sanders J, editors. Modern scientific evidence: the law and science of expert testimony, Vol 2. St. Paul, MN: West, 1997; 50–78.
48. Rowe WF. Firearms identification. In: Saferstein R, editor. Forensic science handbook, Vol II. Englewood Cliffs, NJ: Prentice-Hall, 1988; 393–461.
49. Giannelli PC, Imwinkelried EJ. Scientific evidence. 2nd ed., Vol 1. Charlottesville, VA: The Michie Co., 1993.
50. Thornton JI, Rios FG. Firearms evidence sourcebook. Washington: National Institute of Justice (NCJ 133282, Grant #89IJCX0009, unpublished computer diskette).
51. Biasotti A. The principles of evidence evaluation as applied to firearms and toolmark identification. J Forensic Sci 1964;9:428–55.
52. Biasotti A, Murdock J. Firearms and toolmark identification. In: Faigman DL, Kaye DH, Saks MJ, Sanders J, editors. Modern scientific evidence: the law and science of expert testimony, Vol 2. St. Paul, MN: West, 1997;124–55.
53. Ellen D. The scientific examination of documents: methods and techniques. New York: Halsted Press, 1989.
54. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42:778–86.
55. Peterson JL, Mihajlovic S, Gilliland M. Forensic evidence and the police. Washington: National Institute of Justice, 1984.
56. Moenssens A. Novel scientific evidence in civil and criminal cases: some words of caution. J Criminal Law & Criminology 1993;84: 1–21.
57. Bromwich MR. The FBI Laboratory: an investigation into laboratory practices and alleged misconduct in explosives-related and other cases. Washington, DC: U.S. Department of Justice, Office of the Inspector General, 1997.
58. Starrs J. Mountebanks among forensic scientists. In: Saferstein R, editor. Forensic science handbook. Vol II. Englewood Cliffs, NJ: Prentice-Hall, 1988;2–37.
59. Kelly JA. Questioned document examination. In: Imwinkelried EJ, editor. Scientific and expert evidence. New York: Practicing Law Institute, 1981;695–707.
60. Giannelli PC, Imwinkelried EJ. Scientific evidence. 2nd ed., Vol 2. Charlottesville, VA: The Michie Co. 1993.
61. Peterson JL, Fabricant EL, Field K. Crime laboratory proficiency testing research program—final report. Washington: U.S. Department of Justice, 1978.
62. Peterson JL, Markham PN. Crime laboratory proficiency testing results, 1978–91, I: Identification and classification of physical evidence. J Forensic Sci 1995;40:994–1008.
63. Peterson JL, Markham PN. Crime laboratory proficiency testing results, 1978–1991, II: resolving questions of common origin. J Forensic Sci 1995;40:1009–29.
64. Forensic Sciences Foundation/Collaborative Testing Services, Crime laboratory proficiency testing program—firearms analysis report No. 85-3, 1985.
65. Forensic Sciences Foundation/Collaborative Testing Services, Crime laboratory proficiency testing program—questioned documents analysis Report No. 87-5, 1987.
66. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. J Math Psychol 1969;6:487–96.
67. Gescheider GA. Psychophysics: method, theory, and application. 2nd ed., Hillsdale, NJ: Lawrence Erlbaum, 1985.
68. Dorfman DD, Berbaum KS. RSCORE-J: pooled rating-method data: a computer program for analyzing pooled ROC curves. Behav Res Methods Instruments and Computers 1986;18(5):452–62.
69. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992;27:723–31.
70. Quenoille M. Approximate tests of correlation in time series. J R Stat Soc (Series B) 1949;11:68–84.
71. Tukey, JW. Bias and confidence in not quite large samples. Annals of Mathematical Statistics 1958;29:614.
72. *Daubert* v. *Merrell Dow Pharmaceuticals,* 509 U.S. 579 (1993).
73. Department of Justice, Office of Law Enforcement Standards, National Institute of Standards and Technology, Forensic sciences: review of status and needs. Gaithersburg, MD, 1999 (20899–8102, February 1999, NCJ 173412).
74. Fienberg SE, Martin ME, Straf ML. Sharing research data. Washington, DC: National Academy of Science, 1985.

Additional information and reprint requests:
Michael J. Saks
College of Law
Arizona State University
Box 877906
Tempe, AZ 85287-7906