Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

VOLUME 53 2024 NUMBER 1

Page Contents

Published Semi-annually

P.O. Box 8037, Chattanooga, Tennesse 37414-0037

[1]Bayesian Decision-Making Mitigates Effects of Base Rates on Outcome Confidence: A Monte Carlo Simulation

Russ Warner

Abigail Anderson

John C. Kircher

ABSTRACT

The accuracy of a test for detection is the proportion of individuals the test classifies correctly as truthful or deceptive, whereas outcome confidence is the probability that the test outcome is correct. Many consider test accuracy paramount, but accuracy is of no value if there is little chance the outcome is correct. Outcome confidence is affected by the prevalence of deception in the tested population, known as the base rate of deception. A low base rate of deception reduces confidence in deceptive test outcomes and may render them useless.

We conducted a Monte Carlo simulation that randomly sampled test scores from two hypothetical distributions of scores, one for guilty and one for innocent subjects. We used Bayes' Theorem to combine information from the test scores and the base rates to classify subjects as truthful and deceptive. The computer recorded decision accuracy and outcome confidence at base rates of deception that ranged from 1% to 99%. Although accuracy decreased significantly at extremely low base rates, outcome confidence remained above 66%. When base rates of deception were low, the Bayesian approach extended the range of useful credibility assessments.

[1] The APA editorial staff thanks Dr. Raymond Nelson for accepting the role of Action Editor for this manuscript

Introduction

The present study explored the effects of base rates of deception (BRD) on outcome confidence when the BRD served as the prior probability in a Bayesian analysis of test scores. The prior probability that a person will be deceptive in a criminal investigation could be 20% or more than 70%, depending on whether they are one of several suspects in the early stages of an investigation or a defendant in court. When screening potential employees for drug use or serious crimes, the prior depends on the issues covered by the test. It might be 40% for drug use but only 5% for serious crimes. The prior probability might be well below 0.1% of the tested population when testing for espionage or sabotage.

Base rates and test accuracy affect outcome confidence (Grubin et al., 2016; Nelson, Handler & Thiel, 2021; Raskin, 1984; 1986; 1987). Accuracy for deceptive individuals is the test's sensitivity; it is the proportion of deceptive individuals the test classifies as deceptive. Accuracy for truthful individuals is the test's specificity. Specificity is the proportion of truthful subjects correctly classified as such.

In contrast to test accuracy, outcome confidence is the likelihood that a particular test outcome is correct. Given a job applicant, employee, or suspect failed a polygraph test, what is the probability that the test outcome is correct? The probability that a deception-indicated (DI outcome is correct is the test's Positive Predictive Power (PPV), whereas the probability a no-deception-indicated (NDI) or credible decision is correct is its Negative Predictive Power (NPV). For those who requested the credibility assessments or determine the disposition of tested individuals, outcome confidence (PPV and NPV) is more important than the test accuracy.

To calculate accuracy on truthful and deceptive people and outcome confidence, we cross-classify test subjects into one of four cells of a 2 X 2 table depending on whether they are guilty or innocent (ground truth) and whether we classify them as deceptive or truthful on the test. Table 1a represents the frequencies of these events as the letters A through D, and Table 1b shows the formula to calculate sensitivity, specificity, PPV, and NPV.

Table 1a. Frequencies of outcomes in credibility assessments

		Decision		
		"Deceptive"	**"Truthful"**	Sum
True State	Guilty	A	B	(A + B)
	Innocent	C	D	(C + D)
	Sum	(A + C)	(B +D)	N

Table 1b. Accuracies are proportions of row and column sums.

Statistic	Formula	Meaning
Sensitivity	A / (A + B)	Proportion of deceptive people classified as deceptive
Specificity	D / (C + D)	Proportion of truthful people classified as truthful
Positive Predictive Value	A / (A + C)	Proportion of deceptive outcomes that are correct
Negative Predictive Value	D / (B + D)	Proportion of truthful outcomes that are correct

For example, suppose in a group of 1,000 people to be tested, there are 150 deceptive and 850 truthful people. If the test is 85% accurate, it will correctly classify .85 X 150 = 128 of the deceptive people and will err on (1 - .85) x 150 = 22. Similarly, the test will classify .85 X 850 = 723 truthful subjects correctly and (1 - .85) x 850 = 127 incorrectly. These values appear in the Tables 2a and 2b.

Table 2a. Example frequencies of outcomes

	Deceptive	Truthful	Sum
Guilty	128	22	150
Innocent	127	723	850
Sum	255	745	1,000

Table 2b. Accuracies where 150 of 1,000 test takers (15%) lie on the test

Statistic	Formula	Meaning
Sensitivity	128 / 150 = .85	Proportion of liars classified as deceptive
Specificity	723 / 850 = .85	Proportion of truth-tellers classified as truthful
Positive Predictive Value (PPV)	128 / (128 + 127) = .50	Proportion of deceptive outcomes that are correct
Negative Predictive Value (NPV)	723 / (723 + 22) = .97	Proportion of truthful outcomes that are correct

When the BRD is 15%, the probability that a deceptive outcome is correct is only 50%, the same as the toss of a fair coin.

Raskin (1984; 1986; 1987) discussed how outcome confidence (PPV and NPV) varies with changes in base rates. Generally, as deception in the population decreases, the BRD decreases, PPV drops and NPV increases. This results because the population contains a higher proportion of truthful people, and the false positive errors on those many truthful people make up a higher proportion of deceptive outcomes. Conversely, as deception in the population increases PPV increases, and NPV decreases.

Effects of BRD on Outcome Confidence

The effects of base rates on PPV and NPV are mathematical consequences of the laws of probability. Formally,

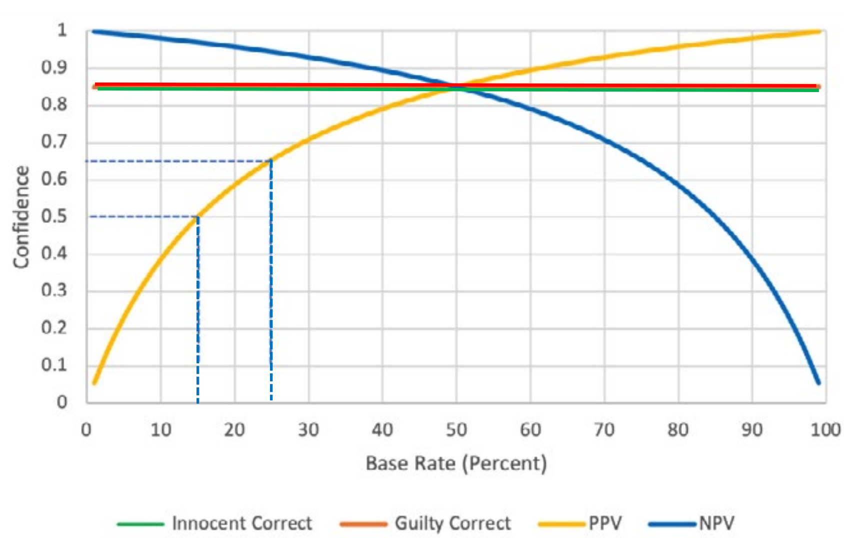
PPV = (sensitivity X BRD) / ((sensitivity X BRD) + (1-specificity) X (1 – BRD))

NPV = (specificity X (1-BRD)) / ((specificity X (1-BRD) + (1-sensitivity) X BRD))

Figure 1 shows the PPV and NPV for base rates of deception ranging from 1% to 99% for a hypothetical credibility test that achieves 85% accuracy on both truthful and deceptive individuals. The relationship between outcome confidence and base rates follows the same pattern regardless of the assumed level of test accuracy. As the BRD approaches zero (everyone in the testing population is truthful), the confidence in a deceptive outcome (PPV) drops precipitously. Conversely, as the BRD approaches 1 (everyone in the testing population is deceptive), the confidence one can have in a truthful outcome drops. There are overlapping horizontal lines at 85% that show equal sensitivity and specificity.

The leftmost dotted lines in Figure 1 indicate 50% confidence that a deceptive outcome is correct when the BRD is 15%. An outcome confidence of 50% means a deceptive outcome is just as likely to be wrong as correct. Ginton (2023) noted that outcome confidence of about 65%, although low, is still of value for a large national police force. To achieve 65% confidence with a test that is 85% accurate, the base rate would have to exceed 25% (see rightmost dotted lines). If fewer than 25% of the tested population is deceptive, confidence in a deceptive outcome would drop below 65%.

Figure 1. Effects of base rates on positive (PPV) and negative predictive value (NPV)



These observations, as important as they are, have been in the literature for decades (Raskin, 1986; 1987; National Research Council, 2003) but seem to have had little impact on the application of credibility assessments in programs that screen for rare events.

The Purpose of this Article

This project explores the effects of a Bayesian analysis of the BRD and test data on outcome confidence (PPV and NPV). We began with the base rate of deception (BRD). The BRD served as the prior probability that the subject was deceptive in a Bayesian analysis that combined information about the BRD and the test score. The analysis yielded a posterior probability of deception that the computer then used to classify the subject as truthful or deceptive. In the present study, Bayes' formula took the following form:

Pr(deceptive/test score) = { BRD X Pr(test score/deceptive) } /
{ BRD X Pr(test score/deceptive) + (1-BRD) X Pr(test score/truthful) },

where BRD was the base rate of deception; Pr(test score/deceptive) was the probability of the test score given the person was deceptive; and Pr(test data/truthful) was the probability of the test score given the person was truthful (Kircher & Raskin, 1988). The Pr(deceptive/test score) was the posterior probability of deception. It was the probability of deception in light of the test data and was the basis for classifying people as truthful or deceptive.

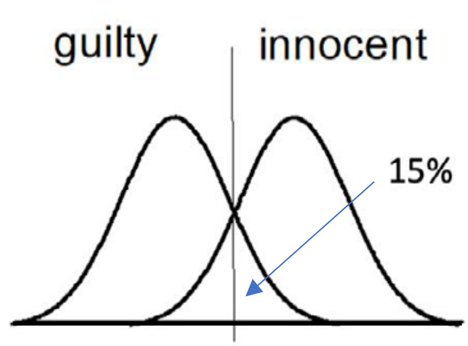
Methods

We conducted a Monte Carlo simulation to assess the effects of the Bayesian approach on outcome confidence. We programmed a computer to perform Monte Carlo simulation experiments by sampling randomly from distributions of hypothetical test scores. The test scores might represent total numerical scores by polygraph examiners, or they might be scores generated by a computer algorithm based on automated measurements of physiological reactions to questions on a polygraph or ocular-motor test.

The leftmost curve in Figure 2 shows a hypothetical distribution of test scores for people guilty of a crime and lie about it on the test. The rightmost curve shows the distribution of scores for innocent, truthful subjects. The distributions are bell-shaped and equidistant but on opposite sides of zero. Most of the scores for guilty people fall below zero, although 15% are positive. Likewise, most scores for innocent subjects exceed zero, although 15% are negative.

The vertical line at zero is the optimal cutoff for classifying subjects as guilty or innocent, assuming we are equally concerned about false positive and false negative decision errors. If we classify those with positive scores as innocent and those with negative scores as guilty, we would correctly classify 85% of innocent and 85% of guilty subjects.

Figure 2. Hypothetical distributions of test scores for guilty and innocent subjects



Moving the cutoff below zero would misclassify more guilty subjects and classify more innocent subjects correctly. Moving the cutoff above zero would detect more liars but misclassify more innocent subjects.

Monte Carlo Simulation

The computer sampled 1,000,000 scores at random from these two distributions. It did this 300 times at each base rate ranging from 1% to 99% and averaged the results. If the base rate was 50%, the computer sampled half of the scores from the guilty distribution and half from the innocent distribution. If the base rate was 10%, the computer sampled 10% of 1,000,000 or 100,000 scores from the guilty distribution, and the remaining 900,000 from the innocent distribution. The means of the distributions for guilty and innocent subjects were -1.04 and +1.04, respectively, and both distributions had standard deviations of 1. The cut value zero isolated the upper and lower 15% of the guilty and innocent distributions, respectively.

To draw a single score from one of these distributions, we used the Box-Muller algorithm to generate a normally distributed random number with a mean of zero and a standard deviation of one (Press et al., 1992). To simulate a guilty subject, we added the mean of the guilty distribution (-1.04) to that random number. To simulate an innocent subject, we added the mean of the innocent distribution (+1.04) to the random number.

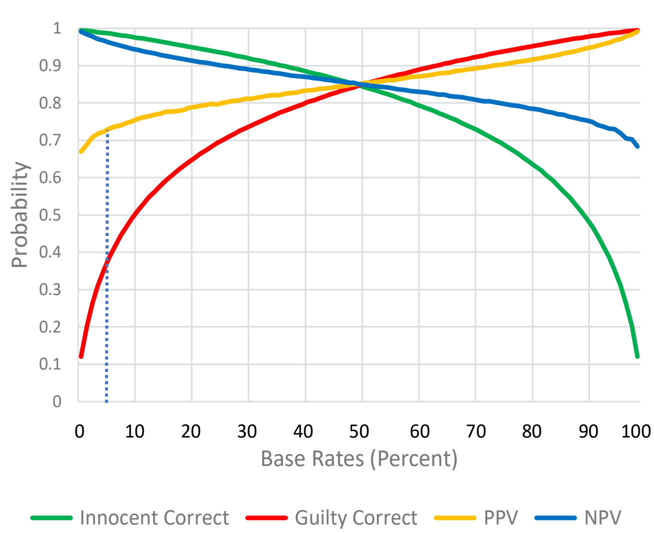
The computer used each randomly selected score to calculate two conditional probabilities, Pr(test data/deceptive) and Pr(test data/truthful) (Kircher & Raskin, 1988). Bayes' formula used those conditional probabilities and the BRD to compute the posterior probability of deception. When the posterior probability was greater than 0.5, the computer classified the person as guilty, and when it was less than or equal to 0.5, it classified the person as innocent.

For each sample of 1,000,000 cases, we computed the accuracy for guilty and innocent subjects, PPV, and NPV and saved the four values. We repeated this process 300 times at base rates ranging from 1% to 99% and averaged the results. We needed a large sample size for each iteration to stabilize the tails of the PPV and NVP curves.

Results

Figure 4 shows the means of the 300 data sets. The X-axis shows BRD ranging from 1% to 99%. The curves show the effects of different BRD on accuracy for guilty and innocent subjects, PPV, and NPV.

Figure 4. Accuracy and PPV and NPV with Bayesian Adjustments for Changes in Base Rates



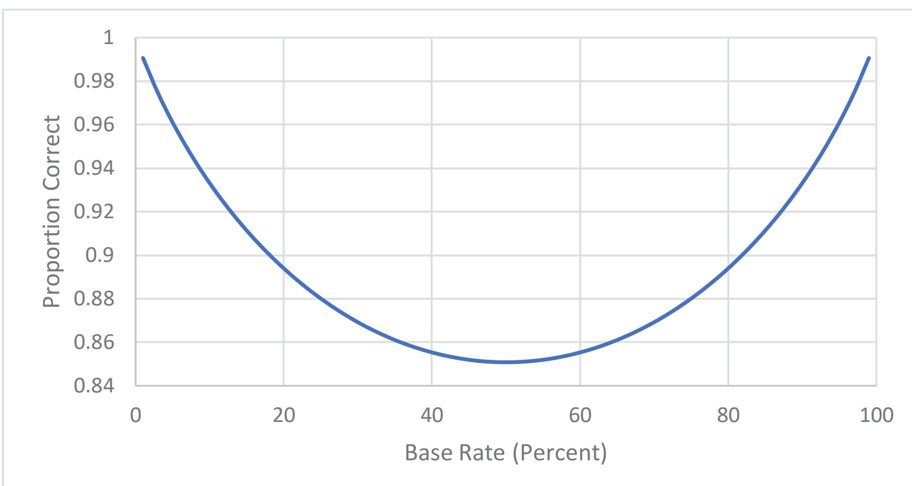
Sensitivity, Specificity, and Confidence

When the BRD was as low as 1%, accuracy on guilty subjects (red) dropped to 12%. Accuracy increased to 85% when the BRD was 50% and approached 100% when the BRD was 99%. The opposite effect occurred for innocent subjects. Accuracy for innocent subjects (green) was almost 100% when the BRD was very low and only 12% when the BRD was very high.

With the Bayesian approach, decisions on one group or the other were inaccurate at extreme base rates. However, PPV and NPV remained above 66% for BRD between 1% and 99%. For example, when the BRD was 5% (dotted line), only 35% of guilty subjects were correctly classified, but when a subject was called deceptive, there was a 72% chance the decision was correct.

Figure 5 shows the proportion of all subjects classified correctly as a function of the BRD. The total number of people classified correctly was highest at the two most extreme base rates because the base rates were themselves so diagnostic of the person's deceptive status. The fewest number of people classified correctly occurred when the BRD was 50%. In that case, the BRD canceled out of Bayes' equation, and accuracy was equal to the validity of the test. The mean accuracy of decisions across base rates from 1% to 99% was 89.4%. When the decision maker does not account for the BRD, accuracy remains constant at the lowest point along the curve (85%).

Figure 5. The proportion of correct classifications at base rates from 1% to 99%



Discussion

If base rates are not considered in credibility assessment testing, the consumer may place too little or too much confidence in test results than they deserve. Assuming a BRD of 15% and a test that is 85% accurate, a truthful outcome has a 97% chance of being correct. However, a deceptive outcome is just as likely to be wrong as it is correct. Unlike a truthful outcome, a deceptive outcome would lack utility.

The present experiment assumed a hypothetical credibility test with 85% accuracy when testing truthful and deceptive individuals. It showed that a Bayesian approach to decision-making extended the range of useful test outcomes. For example, the Utah numerical scoring system (Bell et al., 1999) uses a fixed cutoff decision policy and does not consider variability in BRD. We can expect that type of decision policy to provide at least 75% confidence when base rates are between 35% and 65% for a test with 85% sensitivity and specificity. For the same test, a Bayesian decision policy provides comparable confidence levels for base rates that range from 5% to 95%. We can use the Bayesian approach in a broader range of settings than a traditional decision policy and achieve similar confidence in the outcome.

Increasing test accuracy generally increases outcome confidence and extends the range of base rates that provide helpful test outcomes. For clarity, the present study considered a single test accuracy of 85%. If test accuracy were 90%, BRD between 3% and 97% would produce PPV and NPV that exceed 74%. A reduction in test accuracy would have the opposite effect, decreasing the range of BRD that yields useful outcomes.

Weighted Accuracy

The Bayesian approach increased both outcome confidence and the total number of correct decisions. Without Bayesian adjustments for BRD, a test with 85% accuracy and no inconclusive outcomes yields 85% correct decisions regardless of the BRD. With the Bayesian approach, where the prior probabilities represent the base rates, accuracy varies with BRD. The present study's mean accuracy across all BRD was 89.4%. Mean accuracy across all BRD exceeded test accuracy of 85% because the Bayesian approach included information about base rates in its decisions, and the base rates became more diagnostic as they approached 0 or 1. A decision policy that does not consider the BRD forfeits a source of diagnostic information when the BRD is not 50%.

Confidence Criteria

Consumers of credibility assessments, such as police investigators, human resource personnel, attorneys, and courts, should be less concerned with test accuracy than PPV. PPV indicates the likelihood that a deceptive test outcome is correct. When the PPV is less than 50%, a deceptive outcome is not helpful because it is more likely to be wrong than correct.

At what point is a credibility test useful? Is it when PPV exceeds 65% (Ginton, 2022), or should the criterion be 75% or even 85%? Whatever value is chosen, the base rate should be sufficient to achieve that criterion. If the test is used with a population that contains too few deceptive individuals, confidence in a deceptive outcome will not meet the standard.

A deceptive outcome may not be helpful because it does not meet some criterion level of confidence. However, in the same setting, a truthful outcome will probably have value because the lowest confidence for deceptive outcomes (PPV) is associated with the highest confidence for truthful ones (NPV), and vice versa.

The criterion confidence level for a particular organization probably should depend on the testing context. When missing a deceptive individual in the early stages of a criminal investigation or a screening scenario has serious consequences, a PPV of 65% might be enough to keep the individual among those requiring further evaluation. However, when PPV is as low as 65%, the organization should realize there is a 35% chance the person who failed the test was truthful.

A traditional scoring system that uses fixed cutoffs provides the same sensitivity and specificity regardless of the BRD. Since the Bayesian approach has the disadvantage of lower test sensitivity or specificity as the BRD approaches one extreme or the other, the traditional approach may appear to be superior. However, compared to the Bayesian approach, the traditional approach has lower outcome confidence when the BRD departs from 50%. A traditional, non-Bayesian decision policy will classify 85% of deceptive subjects correctly when the base rate of deception is 10%, but what good is the decision if less than half of deceptive outcomes are correct?

In contrast, the Bayesian approach will correctly detect only 50% of guilty subjects at that BRD, but over 74% of its deceptive outcomes will be correct. For this reason, we believe the Bayesian approach has more utility than traditional decision policies that do not consider base rates. The Bayesian approach has the added advantage of higher overall test accuracy because it incorporates BRD which becomes more diagnostic as it approaches 0 or 1.

Specifying the Base Rate of Deception

The proposed Bayesian approach requires the specification of the BRD. Raskin (1987) and others (Ginton, 2022; Grubin et al., 2016) describe methods for estimating base rates. Those procedures require estimates of test accuracy and observed numbers of truthful and deceptive outcomes. Because accuracy estimates are available from the American Polygraph Association for the polygraph (APA, 2011), any agency that tracks numbers of truthful and deceptive polygraph outcomes could use those formulas to estimate the BRD for its organization.

About the Authors

Russ Warner is the Chief Operations Officer at Converus and has 10 years of experience in the credibility assessment industry. He has developed over 2,500 credibility assessment tests and 9 technical training courses for EyeDetect and VerifEye. He has trained hundreds of government agencies and private organizations in dozens of countries. He received an MBA from Brigham Young University in Provo, Utah, and has 25 years of experience working for technology companies.

Abigail Anderson is a software engineer at Converus specializing in machine learning and artificial intelligence. She has created algorithms to enhance voice recognition for audio-based credibility assessment tests. She holds a master's degree in information systems from Brigham Young University.

John C. Kircher is a widely recognized expert with more than 100 scientific publications and technical reports on the psychophysiological detection of deception. He has worked on related projects with the U.S. Department of Defense, U.S. Secret Service, U.S. Department of Homeland Security, National Science Foundation, National Research Council, Royal Canadian Mounted Police, and numerous state and local police departments. He and David Raskin developed the software for the first computerized field polygraph in 1991 and the first scoring algorithm. John received a Ph.D. in psychology from the University of Utah.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40*, 194-305.

Bell, B. G., Raskin, D. C., Honts, C. R., Kircher J. C. (1999). The Utah numerical scoring system. *Polygraph, 28*, 1-9.

Cook, A. E., Hacker, D. J., Webb, A. K., Osher, D., Kristjansson, S., Woltz, D. J., & Kircher, J.C. (2012). Lyin' Eyes: Ocular-motor Measures of Reading Reveal Deception. *Journal of Experimental Psychology: Applied, 18*(3), 301-313.

Ginton, A. (2023). Calculating the base rate in polygraph populations and the posterior confidence in obtained results in the comparison question test, built upon the proportion of outcomes: The case of Israel Police. *Journal of Police and Criminal Psychology, 38*, 165-171.

Grubin, D., Handler, M. Nelson, R., & Raskin, D.C. (2016). The confidence in your polygraph test outcome may not be what you think it is: Prior probability and accuracy determine confidence. *APA Magazine, 49*(5), 39-47.

Handler, M. (2016). Low base rate screening survival analysis & successive hurdles. *Journal of the American Association of Police Polygraphists*, March 2016.

Kircher, J. C., Raskin, D. C., 1988. Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology, 7*3, 291-302.

Krapohl DJ, Shaw KP (2015). *Fundamental of polygraph practice*. Elsevier-Academic Press.

National Research Council (2003). *The Polygraph and Lie Detection. Committee to Review the Scientific Evidence on the Polygraph*. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press

Nelson R, Handler M, and Thiel E (2021). Posterior odds of deception and truth-telling for low and high prior probabilities. *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice 50*(2):96–105

Press, WH, Teukolsky, SA, Vetterling, WT, & Flannery, BP (1992). *Numerical recipes in Fortran 77 (2nd Ed)*. Cambridge University Press.

Raskin, D.C. (1984, March). *Proposed use of polygraphs in the department of defense*. Statement before the Committee on Armed Services, US Senate.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. Utah Law Review, 1986(1), 29-74.

Raskin, D.C. (1987). Methodological issues in estimating polygraph accuracy in field applications. *Canadian Journal of Behavioral Science, 19*(4), 389-404.

AMERICAN POLYGRAPH ASSOCIATION

DEDICATED TO TRUTH