Comment: The Polygraph and the PVP

D. H. Kaye

Joseph Gastwirth's paper concerns an increasingly important issue affecting public policy—the estimation of error rates in tests for such things as diseases, deception, and drugs. Gastwirth explores the estimator $\hat{C}$ of the predictive value of a positive test (PVP or $P(D \mid S)$) and the variance of $\hat{C}$. He explains how $\hat{C}$ depends on the sensitivity η, the specificity θ, the base rate π, and the sample proportion p of those whom the diagnostic test classifies as having the disease. Furthermore, for large samples, he demonstrates how the variance of $\hat{C}$ depends on η, θ, π and the sample sizes used in estimating these quantities.

Not being a statistician, I shall not attempt to address the technical aspects of Gastwirth's analysis. As an attorney, I am drawn to his discussion of the admissibility of polygraph evidence. First, I shall elaborate on his description of the standards for admissibility of such evidence. Then I shall consider the extent to which his analysis of $\hat{C}$ and $\mathrm{Var}(\hat{C})$ might be brought to bear on the legal question of the admissibility of polygraph evidence.

1. STANDARDS FOR ADMISSIBILITY

As Gastwirth observes, the leading case on the admissibility of scientific evidence is *Frye* v. *United States*, 293 F.1013 (D.C. 1923). Without explanation or precedent, *Frye* created a special test for the admission of scientific evidence—the general acceptance standard. As applied to the polygraph, most scientists in the appropriate fields must agree that conscious deception can be deduced from elevated physiologic responses and that the polygraph accurately detects these responses. In other words, there must be a consensus among scientists that the psychologic theory underlying polygraph testing is valid, and there must be a consensus that the technology for implementing this theory works with reasonable accuracy.

Although many jurisdictions have adopted and adhered to the general acceptance test, a large number have not. Cleary (1983, pages 626–631) collects many of the cases and discusses the merits of the various alternatives. The most popular alternative simply applies the principles of relevance that govern all evidence and the additional constraints on expert testimony generally. Relevant evidence typically is defined as evidence having any tendency to make the existence of any pertinent fact more probable or less probable than it would be without the evidence. At the same time, it is also recognized that even relevant evidence should be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, misleading the jury or undue delay or waste of time. In addition, to testify as an expert, a witness must possess some special qualifications, and the specialized knowledge that he or she proposes to impart must be capable of assisting the judge or jury.

Although Gastwirth describes the judicial reception of polygraph evidence as "somewhat mixed," I think it fairer to say that almost all courts exclude such evidence under either of the standards for admissibility outlined above. Within the United States, in only one jurisdiction, New Mexico, is polygraph evidence admitted over the objection of a party. Almost all courts applying the general acceptance standard have concluded that scientific acceptance of the polygraph as a lie detector is lacking. Similarly, almost all courts applying relevance principles have concluded that the balance of probative value and prejudice counsels against admitting the evidence (Cleary, 1983 and 1987 supplement).

2. THE IMPLICATIONS OF ESTIMATING PVP

With this quick and crude sketch of the legal doctrine as a backdrop, I turn to Gastwirth's suggestion that for the purpose of deciding on the admissibility of polygraph tests, "[f]ocusing on C or the PVP, as well as on the sensitivity and specificity may help in the determination of whether a technique is sufficiently reliable in a particular case." For clarity, it should be noted that when the courts speak of the "reliability" of scientific evidence, they do not mean reliability in the technical sense of statistical precision. They mean that the evidence comes from both a valid and a reliable process of measurement.

In the jurisdictions that require general acceptance, informing the judge or jury of estimates of $P(D \mid S)$ or its components in a particular case should make no difference. What counts is the attitude of the pertinent scientific community, as expressed in the testimony, publications and professional presentations of these scientists. One premise of the general acceptance test is that the courts are not competent to evaluate scientific disputes for themselves, so that they must rely

D. H. Kaye is Director, Center for the Study of Law, Science and Technology, and Professor of Law, Arizona State University, Tempe, Arizona 85287.

instead on the majority opinion of the scientific community to establish validity and reliability (Giannelli and Imwinkelried, 1986).

Of course, if the statistical analysis that Gastwirth describes would lead psychologists and others to revise their opinion of the validity and reliability of polygraphic lie detection, then this shift eventually might lead the courts to conclude that the technique clears the general acceptance hurdle. In addition, a court's own perception of the validity and reliability of a scientific technique may induce it to admit evidence that would seem barred under the *Frye* test. However, given the widely varying and bitterly contested claims for the values of the sensitivity and specificity of the technique reflected in recent research and meta-studies, it seems doubtful that a more favorable climate of scientific or judicial opinion soon will evolve.

Nevertheless, if the general acceptance test were to be satisfied, or if a non-*Frye* jurisdiction allowed the trial judge to admit polygraph evidence once he or she determined that it met the normal relevance concerns, then information about PVP or related quantities could serve two functions. It might reveal a high (or low) degree of probative value, or it might diminish the danger of unfair prejudice by sensitizing the jury to the risk of error in the test. Let us consider these possibilities in turn.

2.1 PVP and Probative Value

"Probative value" is a lawyer's phrase, and it is therefore no criticism of Gastwirth's paper to say that it does not focus on precisely what set or combination of various quantities ($\hat{C}$, PVP, η, θ or π) measures probative value (PV). Some authors act as if PVP is the appropriate measure of PV. Gastwirth notes that "when the prevalence π of deceptives [is] less than .10, the standard error [is] at least half of the expected PVP," a finding that "supports the skepticism... about the reliability of routine polygraph examinations." He adds that "[w]ith prescreening, i.e., when π is at least .5, the proportion $(1 - \hat{C})$ of false positive classifications that are erroneous is reduced to about 10% and its standard error is small." I do not know whether this is meant to imply that the probative value of a diagnosis of deception is high when there is "prescreening" in the form of a judicial finding that there was probable cause to believe that the suspect committed the offense alleged. In discussing *Capua* v. *City of Plainfield*, 55 U.S.L.W. 2170 (D.N.J. Sept. 18, 1986), Gastwirth observed that the requirement of probable cause for drug testing should tend to increase $\hat{C}$ and decrease $\mathrm{Var}(\hat{C})$.

Whether or not Gastwirth believes that a high $\hat{C}$ and a low $\mathrm{Var}(\hat{C})$ establishes substantial probative value, other writers have reasoned that when the proportion of false positives is large (which is to say that the PVP is small), the test is invalid. For instance, Lykken (1987, pages 266–267) gives a hypothetical example involving urine tests:

> Let us assume that the urine test is 95% accurate in both of its jobs, detecting drug users and detecting drugfree persons. Let us also assume that as many as 5.556% of airline pilots [the group that is tested] smoke pot or sniff coke from time to time.... Of every 100,000 tests administered, 95% or 5,278 of the 5,556 guilty drug users should be detected; so far so good. But 5% of the 94,444 drugfree pilots, 4,722 of them, will also fail the urine test! Of the 10,000 tests that are failed, nearly half (47%) will be false positive errors. The accuracy of the set of failed tests will not be 95% but, rather, little better than the accuracy of a coin toss....

The general point—that for low values of π, the PVP can be much lower than the sensitivity and specificity—is well taken. But the possibly unintended implication that the test has about the same PV as a coin flip is wrong. Certainly, a test with a high η and θ is always more valid than a coin flip for which η and θ are 0.5. The outcome of a coin toss conveys no information about deception, so $P(D \mid S) = \pi$.

The potential confusion arises, I think, from an implicit assumption that PVP is the same as PV, or perhaps from a feeling that the test is not very useful unless PVP is more than one-half. Yet, to be admissible, evidence need not by itself persuade the judge or jury that the proponent's story is more probable than not. Several cardboard boxes strategically arranged can support far more weight than each box alone will bear. The probative value PV of an item of evidence has more to do with the change it effects in the odds than with what the final odds are. In considering a polygraph test in the course of judging whether an individual is lying when he denies having committed a crime, we are not using a polygraph as a screening test. Consequently, although focusing on $\hat{C}$ and $\mathrm{Var}(\hat{C})$ seems sensible in a true screening context, such as classifying possibly contaminated blood that might then be subjected to further testing, these quantities may be less useful in the context of forensic proof.

It may be helpful to develop this idea a bit more. Suppose we adopt a Bayesian perspective and think of each piece of evidence at trial as a datum for which a likelihood ratio might be computed and applied to the prior odds in favor of the disputed event to generate posterior odds. We then say that if the final, posterior odds exceed some threshold (say, 1:1 in a civil case), the finder of fact should conclude that the event occurred.

When we recall from the discussion in Section 2.1 that the usual definition of relevance expresses the idea of a change in the probability of a disputed fact, we might be tempted to measure PV by the difference between a posterior and prior probability, as Friedman (1986) proposes. If D stands for the event that a person lied during the polygraph D, S represents the classification of deceptive, and X represents all the other evidence introduced prior to the polygraph testimony, then $\text{PV} = P(D \mid S \cap X) - P(D \mid X)$.

There are two points to note here. First, if the polygraph evidence is the only evidence in the case, then there is no X on which to condition D and PV = PVP $- \pi$. Second, in any real litigation $P(D \mid X)$ may not be π, the prevalence of lying in a population from which the person is randomly drawn. Rather, it represents the probability conditioned on whatever other evidence X already has been introduced, and this will vary from case to case. Likewise, $P(D \mid S \cap X)$ is not the PVP that Gastwirth estimates, because $P(D \mid S)$ ignores X. In either situation, the existence of substantial probative value does not turn strictly on a comparison of PVP to one-half.

Furthermore, it is possible that any measure that involves π or p is inadmissible. The law tends to exclude evidence about similar occurrences involving people other than the parties to the litigation. Proof that most physicians sued for malpractice are (or are not) negligent would not be admissible to show that the physician so charged in a particular case was negligent. The court would exclude estimates of the prevalence π of malpractice on the ground that it does not show anything about this defendant. From a statistical point of view, this result may not be entirely satisfying (see Saks and Kidd, 1980–1981), but it does pose a problem for those who might seek to use $\hat{C}$ to convince a court to admit or exclude a polygraph test.

Another definition of PV, pursued by Good (1983, page 160), is the log likelihood. In the current context, PV = log LR, where $\text{LR} = P(S \mid D)/P(S \mid \bar{D}) = \eta/(1 - \theta)$. If this representation of probative value is appropriate, then information to the effect that the sensitivity is much greater than the complement of the specificity would help establish that the polygraph evidence has substantial probative value. Neither the base rate π nor the conditional probability PVP need be considered.

Analyzing probative value in this way leads me to the conclusion that Gastwirth's Tables 2 and 3 may be of limited utility with regard to forensic proof. This is so even if one indulges the assumption that the polygraph examiners who generated the sample polygraph results used to compute $\hat{C}$ and $\text{Var}(\hat{C})$ were identical in their method and skill to the one who is testifying about the individual presumed to be randomly drawn from a pool comparable to the groups involving the n_1, n_2 and n persons needed to estimate η, θ and π.

2.2 Informing the Jury

Calculations like those summarized in Tables 2 and 3 also might be employed to focus the jury's attention on the risk of error, thereby diminishing the danger that the scientific quality of the polygraph evidence will overawe the jurors. For the reasons given in Section 2.1, however, I would be concerned that $\hat{C}$ and $\text{Var}(\hat{C})$ are less helpful than the estimates of η and θ (and the standard errors of these estimates). $\hat{C}$ does not estimate the posterior probability $P(D \mid S \cap X)$ on which a decision about deception should turn. As mentioned in Kaye (1987), giving a number that is likely to be interpreted as the posterior probability for one item of evidence, computed as if this were the only evidence in the case, while not pursuing the same policy for less quantifiable evidence, may not substantially assist a jury in weighing the full body of evidence. Of course, this is a debatable judgment about the psychology of jurors. Following Shafer (1986), one might maintain that there is room for a "constructive" use of the estimated probability. I merely mean to raise the question, and not to settle it.

3. CONCLUSION

I have approached Gastwirth's paper with professional tunnel vision, looking solely at the ways that his analysis might help courts decide on the admissibility of polygraph evidence. In particular, I have suggested that attending to $\hat{C}$ and $\text{Var}(\hat{C})$ should not do much to affect the accretion of legal doctrine on polygraph evidence. I have not commented on the many merits of his paper, for I trust that these are obvious to all who look.

ACKNOWLEDGMENTS

I am grateful to Mikel Aickin, Dennis Karjala, and Ralph Spritzer for helpful discussion of some of the issues addressed in this Comment.

ADDITIONAL REFERENCES

CLEARY, E., ed. (1983). *McCormick on Evidence.* West Publishing Co., St. Paul, Minn.

FRIEDMAN, R. D. (1986). A close look at probative value. *Boston Univ. Law Rev.* **66** 733–759.

GIANELLI, P. and IMWINKELRIED, P. (1986). *Scientific Evidence.* Michie Co., Charlottesville, Va.

GOOD, I. J. (1983). *Good Thinking.* Univ. of Minnesota Press, Minneapolis.

KAYE, D. H. (1987). Hypothesis tests in the courtroom. In *Contributions to the Theory and Application of Statistics* (A. Gelfand, ed.). Academic, Orlando, Fla.

LYKKEN, D. (1987). The validity of tests: Caveat emptor. *Jurimetrics J.* **27** 263–270.

SAKS, M. and KIDD, R. (1980–1981). Human information processing and adjudication: Trial by heuristics. *Law and Society* **15** 123–160.

SHAFER, G. (1986). The construction of probability arguments. *Boston Univ. Law Rev.* **66** 799–816.

Comment: Base Rates and the Statistical Precision of Polygraph Tests in Various Applications

John C. Kircher and David C. Raskin

In his analysis of the precision of medical screening procedures, Gastwirth discussed the effects of low base rates on the accuracy and utility of test data. The problem of low base rates has been discussed for many years in the psychologic literature (Meehl and Rosen, 1955). In general, when the prevalence of a characteristic such as AIDS or deception in the population is low, it is difficult for a test to improve upon the accuracy that would be obtained if only information about the base rate were used to make diagnoses. If the base rate of a disease is only 0.1%, then diagnosing all patients as disease-free would produce a diagnosis accuracy of 99.9%. To improve upon the accuracy attainable using only base rate information, the accuracy of a test to detect the disease would have to exceed 99.9%. Unfortunately, tests with that level of diagnostic accuracy are extremely rare, and populations with extreme base rates such as those encountered in screening situations are not uncommon.

In the polygraph literature, Raskin (1984) first called attention to the problem of low base rates in hearings before the Committee on Armed Services of the United States Senate on the proposed Department of Defense counterintelligence polygraph program. The Department of Defense was considering widespread testing of federal employees and defense contractors concerning unauthorized disclosures of sensitive information. The base rate issue was particularly important in that context because the vast majority of federal employees and contractors do not make unauthorized disclosures of sensitive information. The base rate of guilt in that population is probably less than 1 in 1000. As discussed by Raskin (1984, 1986) and as Gastwirth's analyses clearly confirm, deceptive polygraph outcomes under those circumstances would be considerably less than 50% correct, even if it is assumed that the polygraph is 90–95% accurate on populations with equal base rates of truthful and deceptive individuals.

Gastwirth focused on a different but related problem. His work reveals that the sampling error of estimates of test accuracy increases as the incidence of the trait in the tested population departs from 50%. In addition to reducing confidence in test outcomes, skewed base rates increase the error in estimating test validity. This makes an already bad situation worse. To our knowledge, this important issue has not been addressed in the polygraph literature, nor has it been discussed in the broader literature on psychologic assessment.

Although we agree with the statistical conclusions drawn by Gastwirth, the implications of his work for applications of polygraph techniques merit further comment. Polygraph tests are used in many different contexts. Law enforcement and private polygraph examiners administer polygraph tests to suspects, defendants and witnesses during criminal investigations. Businesses make extensive use of polygraph tests to screen job applicants and to test employees periodically. Government agencies use polygraphs in criminal investigations and in cases involving risks to national security. The base rate of deception and the costs associated with false positive decision errors are more problematic in some contexts than in others.

Analyses of data from the United States Secret Service for a 2-yr period suggest that the base rate of guilt is about 45% in their criminal investigations (Raskin, 1986). Raskin also reported findings from 292 polygraph tests that he had conducted over a 12-yr period on a confidential basis for defense attor-

John C. Kircher is Assistant Professor of Educational Psychology and David C. Raskin is Professor of Psychology at the University of Utah, Salt Lake City, Utah 84112.