1987

The Validity of Tests: Caveant Omnes

David H. Kaye
Penn State Law

THE VALIDITY OF TESTS:
CAVEANT OMNES*

D. H. Kaye†

It is a tale full of sound and fury, told by an idiot, signifying nothing. So spake Macbeth, lamenting the loss of his beloved but perfidious wife. The tale that I wish to write about is also full of sound and fury; however, it is not told by idiots. On the contrary, the protagonists of my tale are *David T. Lykken,* Professor of Psychiatry and Psychology at the University of Minnesota; *David C. Raskin,* Professor of Psychology at the University of Utah; and *John C. Kircher,* Assistant Professor of Educational Psychology at the University of Utah. In the last issue of this Journal, Dr. Lykken, a well-known (and well-qualified) critic of the polygraph test, argued that polygraphs could be invalid when applied to a group in which the base rate for lying is low.[1] He made the same observation about drug tests and other screening tests. Dr. Raskin, one of the few prominent and scientifically respectable defenders of the use of polygraphs in criminal cases, together with Dr. Kircher, accused Lykken of supplying ''incomplete, incorrect, and misleading'' information.[2] Lykken, in rebuttal, accused Raskin and Kircher of mischievous rhetoric and misstatement.[3]

The exchange among these experts is as timely as it is lively. Proposals and projects involving widespread testing of government employees for drugs and

[1]David Lykken, *The Validity of Tests: Caveat Emptor,* 27 JURIMETRICS J. 263 (1987).

[2]David C. Raskin & John C. Kircher, *The Validity of Lykken's Criticisms: Fact or Fancy?* 27 JURIMETRICS J. 271 (1987).

[3]David T. Lykken, *Reply to Raskin & Kircher,* 27 JURIMETRICS J. 278 (1987).

deception threaten individual privacy and freedom.[4] Required diagnostic testing for certain diseases—most notoriously, for AIDS—raises similar concerns. Incorrect conclusions about who has taken illicit drugs, who has AIDS, and who is lying can be devastating. Yet, perfect knowledge is unattainable. Errors are inevitable. How can we measure the tendency of such tests to err? Which measures are appropriate for deciding whether to use a screening test? And what do the measures of error have to do with the admissibility of evidence in court? These are the kinds of questions that bubble about in the debate between Lykken and Raskin-Kircher.

I would like to stir a few more ingredients into the cauldron. At the risk of placing myself amidst the slings and arrows of outraged psychologists, I shall elaborate on Lykken's description of the terminology commonly applied to medical screening tests,[5] and I shall try to show how the biostatistical and psychometric quantities relate to issues of public policy and forensic proof.

I. MEASURING ACCURACY WITH THE PVP

A medical or psychological test detects certain symptoms of a disease or condition. Unfortunately, the test may not always register the symptoms when they are present, or it may register them when they are absent. Therefore we need two numbers to describe the accuracy of the test. As Lykken reports, these are known as the sensitivity and the specificity. The sensitivity η is the probability that a person with the disease is correctly diagnosed. The specificity $\ominus$ is the probability that a disease-free individual is correctly diagnosed. Letting D be the event that a person has the disease and S stand for the event that the test indicates that the person has the disease, we may write $\eta = \Pr(S|D)$ and $\ominus = \Pr(\bar{S}|\bar{D})$. In other words, the sensitivity is the conditional probability that the test detects the symptoms given that the person has the disease, and the specificity is the conditional probability that the test does not detect the symptoms given that the person does not have the disease.

These quantities can be estimated on the basis of experience with other people who have been tested and subsequently shown with apparent certainty to have (or not to have) the disease. Similarly, a more accurate confirmatory test

---

[4]Lower federal and state courts have been striking down random drug tests of government employees as unreasonable searches and seizures. *See* William J. Curran, *Compulsory Drug Testing: The Legal Barriers,* 316 N. ENG. J. MED. 318 (1987). For a discussion of the application of the Fourth Amendment to polygraph testing of federal employees, *see* 69 CORNELL L. REV. 896 (1985).

[5]The notation as well as much of the mathematical analysis that I present is shamelessly taken from a paper by Joseph L. Gastwirth, of the statistics department of George Washington University, entitled *"The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data,"* and scheduled for publication in the journal STATISTICAL SCIENCE. I also parrot several of Gastwirth's observations about the implications of the mathematical relationships. An intriguing aspect of Gastwirth's analysis that I do not discuss is his estimation of the variance in the statistics relating to the accuracy or efficacy of screening tests.

can be used as the criterion to determine the probabilities. For instance, it has been reported that the ELISA test for AIDS, which is used to screen donated blood, correctly classified blood known to be contaminated 86 out of 88 times, and that the ELISA test correctly classified uncontaminated blood 275 out of 297 times.[6] These figures yield an estimated sensitivity η = 86/88 = .977 and an estimated specificity $\ominus$ = 275/297 = .926.

Both Lykken and Raskin-Kircher focus on Pr(D|S), the conditional probability that a person has the disease given that the test detects the symptoms. This probability is related to η and $\ominus$, but, as Lykken emphasizes, it also depends on the base rate—the prevalence π of the disease in the population from which the subjects for testing presumably are picked at random. Lykken shows the impact of π with the following hypothetical example of airline crew members required to submit to urinalysis before each flight:

> Let us assume that the urine test is 95 percent accurate in both of its jobs, detecting drug users and detecting drug-free persons. Let us also assume that as many as 5.556 percent of airline pilots smoke pot or sniff coke from time to time. . . . Of every 100,000 tests administered, 95 percent or 5,278 of the 5,556 guilty drug-users should be detected. . . . But 5 percent of the 94,444 drug-free pilots, 4,722 of them, will also fail the urine test! Of the 10,000 tests that are failed, nearly half (47 percent) will be false-positive errors. The accuracy of the failed tests will not be 95 percent, but rather, little better than the accuracy of a coin toss.[7]

In the notation that I have introduced, the sensitivity η is .95, the specificity $\ominus$ is .95, and the prevalence π = .05556. Lykken computes (a) the expected rate at which D and S occurs (Pr(DS) = .5278), and (b) the expected rate at which $\bar{D}$ and $\bar{S}$ occurs (Pr($\bar{D}\bar{S}$) = .4722). Since there are 5,278 drug-users correctly detected for every 10,000 persons classified by urinalysis as drug-users, the probability of a true positive classification is Pr(D|S) = .5278. This conditional probability is also called the Predictive Value Positive, or PVP. As Lykken observes, in this instance it is not very different from the probability of a fair coin coming up heads.

There is, however, a problem with the implicit (and perhaps unintended) suggestion that the urine test on this population is comparable to detecting drug usage by pitching pennies at the crews. Fair coins have a sensitivity and specificity of .5 rather than .95. Applying these values to the hypothetical drug-testing scenario yields a PVP of Pr(D|S) = .05556.

The careful reader should be able to verify this numerical result with the ad

---

[6]S. H. Weiss, J. J. Goedert, & M. G. Sarngadharan, *The AIDS Seroepidemiology Working Group;* R. C. Gallo & A. Blattner, *Screening Test for HLTV-III (AIDS Agent) Antibodies,* 253 J. AM. MED. Ass'n 221 (1985).

[7]Lykken, *supra* note 1, at 265-66.

hoc analysis that Lykken (and I) have used so far, but it will be helpful to exhibit the general formula for the PVP. The formula, which is easily derived,[8] is

$$Pr(D \mid S) = \frac{\pi\eta}{\pi(\eta + \ominus - 1) + (1 - \ominus)} \qquad (1)$$

The computationally inclined reader may wish to verify that substituting Lykken's hypothetical values for the sensitivity, specificity and prevalence in (1) yields a PVP of .5278, as I have claimed. Likewise, the PVP for a coin flipping test on this population is $(.05556)(.5)/[.05556(.5 + .5 - 1) + (1 - .5)]$ = .05556. The PVPs are the same because the terms involving η and $\ominus$ for the coin cancel themselves out! But this should come as no surprise. Since the coin toss gives us no information about drug use, our best prediction of the probability that a randomly selected person used drugs remains the prevalence π of drug usage in the population.

In short, it would be wrong to equate the drug test that has a sensitivity and specificity of .95—and is fairly informative—to a coin toss that has a sensitivity and specificity of .5—and is utterly uninformative. Yet, the observation that the informative drug test has an unimpressive PVP with a population characterized by the low prevalence π = .05556 is unimpeachable.

II. IMPLICATIONS OF PVP FOR SCREENING TESTS

What, then, are we to make of the PVP? First, it seems clear that in the context of *screening* tests, PVP *is* a helpful figure. When it is low because we are screening a population with a low incidence of the disease or condition, we know that the probability of a correct conclusion that the tested individual has the disease is small. Further testing of the selected group is therefore desirable. At this point use of a more expensive test with a higher sensitivity and specific-

---

[8]By definition, the conditional probability of the disease D given the detection of the symptoms is found by looking at all instances of the disease and determining the fraction of these cases in which the test detects the symptoms. In symbols, $Pr(D \mid S) = Pr(DS)/Pr(S)$. (The "DS" may be read as "D and S." The equation states that the conditional probability of D given S is the relative frequency with which both D and S occur when D occurs.)

The numerator can be rewritten as $Pr(DS) = Pr(SD) = Pr(D)Pr(S \mid D)$. But $Pr(D)$ is the prevalence π of the disease in the population being sampled, and the $Pr(S \mid D)$ is our old friend, the sensitivity η. Hence, the numerator is just $\pi\eta$.

The denominator $Pr(S)$ can be expanded as follows:
$$Pr(S) = Pr(SD) + Pr(S\bar{D}) = Pr(D)Pr(S \mid D) + Pr(\bar{D})Pr(S \mid \bar{D})$$
$$= \pi\eta + [1 - Pr(D)][1 - Pr(\bar{S} \mid \bar{D})] = \pi\eta + (1 - \pi)(1 - \ominus).$$
Substituting these expressions for the numerator and denominator and rearranging a few terms gives Equation (1).

ity is justified, or, depending on the relative seriousness of false positives and false negatives as well as the cost of testing, multiple screening tests may be applied to confirm the preliminary finding of the disease. Thus, in AIDS testing, it makes sense to use a series of inexpensive ELISA tests on a population with a low prevalence of AIDS in the first instance. However, since the more definitive Western blot test is available for confirmatory testing, and because it is important to avoid incorrectly identifying a patient as an AIDS victim, it would not be appropriate to stop with a single ELISA test.[9]

So too, *if* the polygraph were to be used strictly as a screening test for identifying secret foreign agents in sensitive positions, and *if* the sensitivity and specificity were high in the case of such mendacious subjects with every motive to induce false positives, then a low PVP should not ipso facto preclude the use of the test. Whether the experience with the Air Force Seven Screens Program establishes that these conditions are fulfilled, as Raskin-Kircher seem to imply, or whether the polygraph aspect of the program was black magic, as Lykken contends, is an issue I leave to others to debate.

A second point to note about a low PVP for a test applied to members of a low base rate population is that the inverse proposition is also true. The PVP can be high when the test is used on people from a population characterized by a moderate or high base rate. Consequently, prescreening that raised the prevalence in the population subject to testing would lead to a more impressive PVP. For instance, if probable cause were required *before* resorting to drug tests, if magistrates and judges were serious and scrupulous about insisting on probable cause, and if they were able to know probable cause when they saw it, then rather than testing airline crews for which $\pi = .05556$, the prevalence in the group being tested might be .5 or greater. Substituting a figure like $\pi = .55$ in (1) along with the previous hypothetical value of .95 for η and $\ominus$ gives a PVP of .959. In other words, a low PVP as applied to a low prevalence population does not mean that the test is worthless when used on a different population.

There is, of course, much more that could be said about the PVP.[10] Since this is an essay rather than an encyclopedia, however, I shall turn to another issue that divides Lykken and Raskin-Kircher. It is the idea of admitting polygraph results only when they support a defendant's claim of innocence. Their anticlinal analyses of this proposal will lead us to something called the PVN, a sibling of the PVP.

---

[9] If the costs of false positives and false negatives can be quantified, then decision theory can be used to arrive at an optimal testing strategy. *See* D. J. Fink and R. S. Galen, *Probabilistic Approaches to Clinical Decision Support,* in 2 COMPUTER AIDS TO CLINICAL DECISIONS 1 (B. T. Williams, ed., 1982).

[10] Mikel Aickin, for instance, has suggested to me that a statistic known as the odds ratio would be superior to PVP as a measure of efficacy. Taking a cue from yet another paper by Gastwirth, I discussed the odds ratio as a measure of discrimination in *Statistical Evidence of Discrimination in Jury Selection,* in STATISTICAL METHODS IN DISCRIMINATION LITIGATION 13, 20-21 (D. H. Kaye & M. Aickin eds. 1986). The odds ratio conceivably could be used to measure the probative value of a positive test result. This approach would not change the conclusions developed *infra* in Part IV about the distinction between PVP and probative value.

III. MEASURING ACCURACY WITH THE PVN

There is precedent for a rule that would permit polygraph tests to be introduced to establish innocence but not guilt. For many years, the rule in paternity actions was that the results of blood group tests on the defendant, mother and child were not admissible—except to establish that the defendant was not the father. The theory behind this apparently lopsided rule was that the probative value of an exculpatory finding was immense, while that of an inculpatory result was slight.[11]

In arguing against similar proposals to admit polygraph test results solely to exculpate defendants, Lykken observes that "[w]hen we are testing a large number of persons, of whom the majority are in fact offenders, the validity of negative or exculpatory polygraph tests will be much lower than the overall validity of the technique."[12] He offers the following illustration involving "a naive defendant" and "a court-appointed polygraph examiner":

> The only test results that will be presented in court under the proposed rule will be of those tests that are passed. Since he has little to lose if he fails, and yet quite a bit to gain if he passes, every sensible defendant will agree to be tested. Let us assume, against the best evidence, that non-adversarial polygraph tests are 80 percent accurate in general [and] that, among all criminal defendants actually brought to trial, 80 percent are in fact guilty. Out of every 1,000 defendants, on these assumptions, 80 percent or 160 of the 200 innocent defendants will pass and those 160 test results will be offered in evidence. But, at the same time, 20 percent, or 160 of the 800 guilty defendants also will pass. With respect to the subset of 320 polygraph tests considered at trial, the actual validity will be only 50 percent (chance) rather than 80 percent.[13]

Raskin-Kircher take issue with this example. They claim that they have "empirically derived estimates from actual cases" that base rates for guilt among criminal suspects who volunteer for tests are .5 or .6, and that "[u]sing conditional probability analysis and the laboratory and field accuracy estimates, that translates into . . . confidence in truthful . . . test outcomes" ranging from .83 to .97. Such high values for the conditional probability, they write, "are acceptable for evidentiary purposes."[14]

To see who has the better of this exchange, we need to understand what the disputants are calculating and what it has to do with the admissibility of evidence. I shall elaborate on Lykken's illustration, indicate where Raskin-Kircher's "empirically derived estimates" come from, and explain how they

---

[11]When the number of identifiable blood groups was small and the groups were fairly common, the failure to exclude a man did little to narrow the class of possible fathers. The introduction in recent decades of more probative genetic testing has undermined this traditional rule. *See, e.g.,* McCormick on Evidence § 205 (E. Cleary, 3d ed. 1984).

[12]Lykken, *supra* note 1, at 268.

[13]Id.

[14]Raskin & Kircher, *supra* note 2, at 275.

lead to conditional probabilities (what Raskin-Kircher call "confidence"). Then I shall consider what these probabilities tell us about the admissibility of evidence. My answer to the last question, for those who cannot stand the suspense, is "Not much."

A. Computing the Predictive Value Negative

The numbers that Lykken and Raskin-Kircher bandy about in their discussions of the merits of exculpatory polygraph testing are not values of the PVP. The Predictive Value Positive, it will be recalled, is the probability $Pr(D\,|\,S)$ of the disease (here, deception) given that the test classifies the patient (here, defendant) as symptomatic (here, as deceptive). Since Lykken and Raskin-Kircher are interested in the accuracy of a negative diagnosis of deception, they correctly focus on the conditional probability $Pr(\bar{D}\,|\,\bar{S})$ that a randomly drawn defendant is *not* deceptive given that the polygraph analyst so certifies. This conditional probability is called the Predictive Value Negative, or PVN. The general formula is a simple variation of the expression already given for the PVP:[15]

$$Pr(\bar{D}\,|\,\bar{S}) \; = \; \frac{Pr(\bar{D}\bar{S})}{Pr(\bar{S})} \; = \; \frac{\Theta\,(1-\pi)}{\Theta\,(1-\pi)\,+\,\pi(1-\eta)} \tag{2}$$

In Lykken's example, the specificity η, the sensitivity Θ, and the prevalence of deception π all have the value .8. Inserting this value into (2) yields PVN $=$ .5.

Raskin-Kircher do not challenge this bit of mathematics. How could they? Rather, they complain that "Lykken made a fatal error by assuming that because 80 percent of defendants are guilty, 80 percent of those who take such tests are guilty. His analysis is incorrect because innocent defendants actually volunteer to take the tests with greater frequency than do guilty defendants."[16]

Yet, the PVN of .5 does not result from the choice of .8 by itself, or from the assumption that all defendants would submit to polygraph examinations if the results could not be used against them. The culprit, if there is one, is the equality among the three variables on the right hand side of (2). As long as $\pi = \eta = \Theta$, the value of PVN must be .5, regardless of the value of π, η and Θ.[17] The selection of .8 is mere window-dressing.

Thus, Raskin-Kircher must be arguing that Lykken's hypothetical values

---

[15]The proof is perfectly analogous to the derivation provided *supra* in note 9 for PVP.

[16]Raskin & Kircher, *supra* note 2, at 275.

[17]Just let $\pi = \eta = \Theta = x$, where x is any real number between zero and one. Then from (2) it follows that:

$$PVN = \frac{x(1-x)}{x(1-x)\,+\,x(1-x)} = \frac{1}{2}$$

for η and $\ominus$ as well as π are inappropriate and that more realistic choices establish that PVN is well above the value of .5. Let us consider how they propose we should estimate PVN. There are two intertwined components to the analysis: the "empirically derived estimates" of π, and "laboratory and field accuracy estimates" of the sensitivity η and the specificity $\ominus$. Both merit scrutiny.

B. Estimating Prevalence in a Subpopulation

The estimation of the prevalence in a subpopulation like the criminal defendants who "volunteer" for polygraph testing is slightly tricky. I glossed over the point in Part I. The method comes from Poland, where Steinhaus proposed its use in connection with blood group testing in paternity litigation.[18] Steinhaus's method rests on the fact that the probability $Pr(S) = p$ that a person randomly selected for testing will be diagnosed as having a disease or condition (a) can be estimated as the proportion $\hat{p}$ of people tested who are so diagnosed, and (b) can be expressed in terms of the prevalence π in the population being sampled for testing, the sensitivity η of the test in this group, and the specificity $\ominus$ of the test. It turns out that the prevalence π is given by

$$\pi = \frac{p - (1 - \ominus)}{\eta + \ominus - 1} \tag{3}$$

While I shall not bother to derive this result,[19] I might point out that it has the desirable property that as the sensitivity and specificity approach one, the estimate of the prevalence is simply the proportion of the entire population that would have positive test results. This is a plausible property because, for a perfectly accurate test, everyone classified as diseased or deceptive is in fact diseased or deceptive.[20] In this limiting case, $\pi = p$ and we can estimate the population proportion p with the sample proportion $\hat{p}$.

Implicitly relying on equation (3), Raskin-Kircher assert that "[u]sing empirically derived estimates from actual cases, base rates of guilt among suspects who volunteer for tests are approximately 43-48 percent in law enforcement settings and 60 percent in so-called 'friendly tests' performed confidentially for defense attorneys."[21] Knowing that Raskin-Kircher are using (3) to generate

---

[18]I have not seen Steinhaus's discussion, but am relying on the description of his method in Gastwirth, *supra* note 7, and in Finkelstein & Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 490 (1970). The latter authors cite H. STEINHAUS, THE ESTABLISHMENT OF PATERNITY, Prace Wroclawskiego Towarzystwa Naukowego ser. A, No. 32, at 5 (1954). I would be grateful to anyone who can tell me more about Steinhaus.

[19]Compulsive or unusually curious readers should consult Gastwirth's paper, which develops this equation in a few lines.

[20]The only possible error in the estimate arises from the fact that p is a proportion for a limited sample of a larger population.

[21]Raskin & Kircher, *supra* note 2, at 275.

these estimates, we can see that these base rates are no better than the underlying values for p, η and $\ominus$. In this way, we are led to ask how well these quantities are known.

An article by Raskin published a year ago in the *Utah Law Review* reveals the sources of the .43–.48 and .60 figures.[22] The only law enforcement setting mentioned there is a series of criminal investigations by the U.S. Secret Service in the three years from 1980 to 1982 for which the overall $\hat{p}$ was said to be .5. The only confidential tests for defense attorneys were those that Raskin himself performed from 1973 to 1985. Raskin determined that the fraction $\hat{p} = .66$ of the defendants who came to him over these twelve years were deceptive.

To the extent that the polygraph test is not completely standardized, to the degree that other factors affecting the defendant may vary, and to the extent that estimates of p, η and $\ominus$ are each subject to sampling error, the figures for π that Raskin-Kircher pull out of the hat should not be taken as firm. Raskin observes, for instance, that the Secret Service has "the highest quality polygraph program among law enforcement agencies,"[23] and it is a safe bet that he regards the examinations he conducts as well above average. Tests performed by other agencies and examiners will have lower sensitivities and specificities, and these other examinations may give rise to different sample proportions. Consequently, it probably would be a mistake to interpret Raskin-Kircher as insisting that the base rate for deception among volunteers in law enforcement settings is rigidly confined to the 43-48 percent range or that the rate with "friendly polygraphers" is exactly 60 percent. No doubt, these are estimates, and the plausible range for the prevalences is somewhat broader. Just how fuzzy the estimates are is an appropriate topic for further empirical study.

C. From Estimates of Prevalence to Estimates of PVN

Having seen where the estimates of the prevalence π of deception in the "volunteer" population come from, we are in a position to consider the claim that the "confidence" in negative polygraph findings is no lower than 83 percent and as high as 97 percent. The procedure for moving from the estimated prevalence to the "confidence" is straightforward. Raskin-Kircher use the estimates of π from (3) in the expression (2) for the PVN. The 97 percent "confidence," for instance, results from setting π equal to .6 (deduced from the Secret Service experience) in (2). Of course, estimates of the specificity η and the sensitivity $\ominus$ also enter into (2).[24] To obtain these numbers, Raskin-Kircher

---

[22]Raskin, *The Polygraph in 1986: Scientific, Professional, and Legal Issues Surrounding Applications and Acceptance of Polygraph Evidence*, 1986 UTAH L. REV. 29.

[23]*Id.* at 59 n. 91.

[24]These quantities also are required to estimate π. With the kind of numbers we are talking about, however, any fairly high values for η and $\ominus$ will give estimates of π close to p. For this reason, I did not dwell on the uncertainty in π introduced by not knowing the precise values of η and $\ominus$.

apparently rely on Raskin's selection in his *Utah* article of five particular laboratory studies of polygraph accuracy.[25] Raskin pooled these studies to obtain an estimated sensitivity and specificity of $\eta = .97$ and $\ominus = .92$. Putting these values into (2) with the estimated prevalence of .6 yields an estimated PVN of .97, as Raskin-Kircher report.

Naturally, the .97 figure can be no more precise than the .6 figure whose fuzziness I have already mentioned. In addition, problems of sampling error and generalizability of the values of η and $\ominus$ taken from the pooled laboratory studies contaminate the computed PVN of .97. How substantial these problems are is open to debate. At bottom, the disagreement between Lykken and Raskin-Kircher emanates more from their polar opinions about the accuracy (sensitivity and specificity) of polygraph testing than from their dispute over the likely base rate for deception in exculpatory polygraph testing. Lykken proceeds on the premise that values for η and $\ominus$ of as high as .8 are overly generous. Raskin-Kircher, on the other hand, are prepared to start with substantially higher values of the sensitivity and specificity. Interestingly enough, if we use Raskin-Kircher's estimates of .97 and .92 with Lykken's hypothetical base rate of .8, we find from (2) that PVN is .88, a figure that Raskin-Kircher think should make the polygraph "acceptable for evidentiary purposes." As I said before, the argument over whether the base rate is .8 or something less turns out to be a huge red herring. The real quarrel concerns the sensitivity and specificity of the polygraph "diagnosis" of deception. The Department of Defense, the staff of the Office of Technology Assessment, the American Polygraph Association, and various academic psychologists have expressed divergent views about the magnitude and variability of measures of the accuracy of the polygraph for lie-detection. Having clarified (I hope) some of the statistical aspects of the exchange between Lykken and Raskin-Kircher, I make no effort to revisit this much mooted question.

IV. WHY PVP OR PVN IS NOT PV

Raskin-Kircher's talk of "evidentiary purposes" takes us to the final matter that I want to raise. Does a high value of a conditional probability like PVP or PVN demonstrate such high probative value (PV) that the polygraph evidence should be admissible in evidence? Of course, one might insist that even if polygraph operators could detect deception or truthfulness with near certainty, their testimony nevertheless should be excluded. However, I want to put such arguments to the side and to focus entirely on the preliminary question of the adequacy of PVP or PVN as a measure of probative value.

My own view is that these quantities cannot generally be taken to quantify PV, and I think I can show this to almost anybody's satisfaction with a simple

---

[25]Raskin, *supra* note 22, at 43 (Table 1).

example. I have a fantastic test for deception, the Magic Penny Toss. Since it is *de rigueur* to employ an acronym for psychological tests (as in TAT, WISC, ITBS and LSAT, but not as in Rorschach or Meyers-Briggs), I call this test the MPT. My Magic Penny is a weighted coin that will come up heads 75 percent of the time. When it does come up heads, I conclude that the defendant is deceptive. When it comes up tails, I announce that he is truthful. Since the MPT is sheer balderdash, the sensitivity and specificity of the MPT are .5. Anyone who has the foggiest idea of the true nature of the MPT will hold that the MPT has no, and I mean no, probative value.

Assume further that every defendant must submit to an MPT—there is no prohibition on unreasonable searches and seizures and no privilege against self-incrimination in my jurisdiction. Finally, imagine that we have acquired powerful empirical evidence to the effect that 85 percent of all defendants tested in the past twenty years are guilty and that this prevalence varies not a whit from year to year—life is very stable in my jurisdiction. On these purposefully silly assumptions, I can say with some assurance that $\pi = .85$.

Now, as we saw in Part II, when the sensitivity and specificity are each .5, the PVP is the same as the base rate for deception: PVP $= \pi$. The MPT provides no valid information that would warrant modifying the prior probability π that a defendant is deceptive. Hence, we can be 85 percent confident that a score of deceptive on the MPT is correct: PVP $= .85$. Yet, we have introduced no evidence to show that the defendant is guilty or innocent. Our ''confidence'' in deception arises entirely from the empirical findings about defendants as an undifferentiated class.

We may draw two conclusions from the peculiar case of the MPT. The probative value of a test can be zero, while the PVP can be close to one. In these circumstances, the PVP does not express or quantify evidentiary value. Beyond this, it is possible that any measure that involves π or an estimator of π 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. A court might exclude estimates of the prevalence 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, but it does pose a problem for those who might seek to use PVP to convince a court to admit or exclude a polygraph test. Specifically, when Raskin-Kircher characterize a high value of PVN as acceptable for evidentiary purposes, at least part of that value is attributable to the prevalence for truthfulness among many people and across many years. Even if the sensitivity and specificity of the polygraph were as high as they maintain, this much of their analysis is problematic from the forensic standpoint.

More generally, the case of Lykken against Raskin-Kircher underscores the need to be clear about the meaning of probative value. Since ''probative

value'' is a lawyer's phrase, it is no criticism of these authors to say that they did not focus on this issue. Let me close this essay, therefore, with a few remarks about the meaning of probative value. After all, if you agree that PVP is not PV, you should be asking what is.

Federal Rule of Evidence 401 defines "relevant evidence" as "evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence." At the same time, the rules do not define "probative value," although Rule 403 requires the trial judge to exclude evidence whose "probative value is substantially outweighed" by factors such as prejudicial impact and undue consumption of time.

Since 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 a prior probability.[26] If D stands for the event that a person lied during the polygraph test, S represents the classification of deception, and X represents all the other evidence introduced prior to the polygraph testimony, then $PV = Pr(D|SX) - Pr(D|X)$.

There are two major points to note here. First, if the polygraph evidence is the only evidence in the case, then there is no other evidence X to worry about, and $PV = Pr(D|S) - Pr(D) = PVP - \pi$. Under the change-in-the-probability interpretation of probative value and the dubious assumption of no other evidence against the accused, PVP overstates the probative value. Second, in any real litigation there will be other, probative evidence X, so $Pr(D|X)$ will 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, $Pr(D|SX)$ is not the PVP that Lykken and Raskin-Kircher discuss, because $Pr(D|S)$ ignores X.

Another definition of PV, pursued by a goodly number of statisticians and philosophers, is the likelihood ratio (or its logarithm). In the current context, this measure of probative value would be $LR = Pr(S|D)/Pr(\bar{S}|D) = \eta/(1 - \ominus)$. It tells us how many times more probable a diagnosis of deception is under the assumption that the subject is dissembling than under the hypothesis that he or she is speaking truthfully. 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.

The likelihood ratio is an appealing measure of PV, largely because it can be understood as the factor that transforms the prior probability into the posterior one. Suppose that after hearing some evidence X tending to prove that a

---

[26]For a development and defense of this interpretation of PV, *see* Richard D. Friedman, *A Close Look at Probative Value*, 66 B.U. L. REV. 733 (1986).

defendant embezzled money from his employer, we would judge the probability that she is deceptive in pleading not guilty to be about $Pr(D|X) = 4/5$. Now we learn that a polygraph analyst found defendant to be lying when she denied her guilt, and, like Lykken, we treat (for the sake of argument) the polygraph test as having a sensitivity and specificity of .8. The probability of deception given this further evidence will be higher than the prior probability of 4/5.

To calculate the change, it is convenient to talk in terms of odds rather than probabilities. If the probability of an event is p, the odds in favor of the event are $p/(1-p)$. For instance, since the prior probability of deception (the probability estimated in light of the evidence prior to the polygraph results) is 4/5, the prior odds are $(4/5)/(1/5) = 4$, or 4-to-1. According to an elementary formula of probability theory known as Bayes's rule (which lay at the heart of our explication of the meaning of PVP and PVN), the posterior odds are just the prior odds times the likelihood ratio LR:

$$\frac{Pr(D|SX)}{1-Pr(D|SX)} = LR \frac{Pr(D|X)}{1-Pr(D|X)} \qquad (4)$$

Since $LR = \eta/(1-\ominus) = .8/.2 = 4$, the odds in favor of deception have gone from 4 to 16—a fourfold increase.

Had the polygraph evidence been introduced at a different point, say when the prior odds were only 2, the posterior odds would have jumped to 8—a fourfold increase. In short, the likelihood ratio indicates how much the odds grow with the introduction of the polygraph results. In our example, they grow by a factor of four. On the other hand, if we were to use Raskin's estimates of $\eta = .97$ and $\ominus = .92$, we would conclude that the odds grow by a much larger factor, namely, $LR = .97/.08 = 12$.

The crucial point, of course, is that under the likelihood interpretation of probative value, PV is not PVP or PVN. It is LR. Whether we use the likelihood interpretation or the difference-in-the-probabilities interpretation of probative value, in judging whether polygraph results have substantial probative value it is a mistake to fuss about PVP or PVN. Raskin-Kircher may (or may not) be right in believing that exculpatory polygraph results are "acceptable for evidentiary purposes," but any implication in their paper that this admissibility flows from a high PVN is mistaken. Probative value is not the same as PVP or PVN, and these conditional probabilities cannot capture the confidence that one should have in a defendant's guilt or innocence.