Rejoinder

Joseph L. Gastwirth

First, I wish to thank all the commentators for their thoughtful discussion. The variety of potential applications of the screening paradigm they describe also demonstrates a wide variation in social and economic costs of the two types of error. Hence, a simple resolution of all the issues involved is unlikely to be achieved.

Although Professor Kaye is correct in noting that the polygraph is generally inadmissible unless both parties stipulate to it, recent cases suggest that the current reluctance of courts to admit the result of a polygraph test is due to the concern that juries may give it too much weight. Judge Lacey (1984) provides an example where a jury was not unduly swayed. Egesdahl (1986) cites *McMorris* v. *Israel*, 643 F.2d 458 (7th Cir. 1981) and *State* v. *Stanislawski*, 62 Wis. 2d 703, 216 N.W. 2d 8 (1974) for the view that polygraph tests have reached a sufficient degree of scientific standing that automatic rejection of expert testimony based on them is no longer warranted. A similar position was taken in *U.S.* v. *Oliver*, 525 F.2d 731 (8th Cir. 1975) a case that approved the admission of a polygraph exam that both sides had agreed to before its administration. In addition to discussing the conflict between the *Frye* criteria and the relevancy approach, Egesdahl (1986) provides many references to empirical studies of the effect of scientific evidence on juries as well as cases that considered the admissibility of such evidence. The newer studies show that mock juries are not overly influenced by scientific techniques especially when ranges for their accuracy, e.g. 70–90% for the polygraph, are presented to them.

I believe that presenting juries with the standard errors and confidence intervals for all pertinent parameters (θ, η, C and F) should aid their understanding of the degree of possible error inherent in any scientific device. The results in Tables 2 and 3 show that the reliability of any technique needs to be determined from a reasonably large study. Indeed, Dr. Goldberg's comments on the screening application reinforce the desirability of carefully determining the error rates *prior* to initiating a mass screening program. Professor Kaye's remarks on the relationship between the PVP and probative value supplement my view that the PVP as well as the sensitivity and specificity should aid in the assessment of the weight that should be given to scientific evidence.

As Professor Kaye notes the PVP is not the same as the legal concept of probative value nor did I equate them. Apparently others have stated that the PVP must exceed ½ for test results to be useful. The legal concept of probative value is broader than the usual mathematical models allow and it is unwise to fix a threshold value for admissibility in terms of the PVP or any one statistical measure. More research on how jurors utilize the other evidence to form their prior probability of guilt and how the scientific evidence subsequently changes this probability is needed. Fairly complex models may be required as jurors see and hear the same evidence and discuss the case so procedures for combining dependent data should play a role. Furthermore, the strength and amount of other relevant evidence may affect the admissibility of lesser-quality scientific information.

Although I have some sympathy with Professor Kaye's Bayesian view, he may have overemphasized the difference between $P(D|S \cap X)$ and $P(D|S)$ as it may be possible either to obtain accuracy rates for persons possessing X or to demonstrate that belonging to the group specified by X (the other evidence) does not affect the accuracy of the test or device. Thus, the relevant issue is whether the accuracy rates η and θ are known (sufficiently precisely) for the appropriate population. Indeed, Section 5 and the comments by Kircher and Raskin and Goldberg underscore the importance of verifying the accuracy rates of a screening test on the population for which it will be used.

Although I discussed the PVP, the PVN should also be considered in weighing the admissibility of scientific evidence. The most controversial cases involving the admissibility of polygraph, drug tests and other scientific evidence occur when other evidence is relatively sparse. In this situation courts must focus on the accuracy of the procedure. For example, the accuracy of the EMIT drug test arose in *Pella* v. *Adams*, 638 F. Supp 94 (D. Nev. 1986) when a prison inmate was disciplined for drug use after a positive test. As Dr. Wittes raised the issue of drug tests we quote from page 97 of the opinion:

> [5] "Because the evidence against Pella is scarce without the positive EMIT test, the Court finds that inquiry into its reliability and accuracy is appropriate. Several courts have examined the results of the EMIT test, with varying conclusions. See e.g., *Higgs* v. *Wilson*, 616 F.Supp. 226 (W.D.Ky.1985) (granted a preliminary injunction against the prison from disciplining an inmate on the sole basis of an unconfirmed positive EMIT test); *Wykoff* v. *Resig*, 613 F.Supp. 1504

(N.D.Ind.1985) (in order to meet due process requirements, ordered all positive EMIT results in the future should be confirmed by a second EMIT test or its equivalent); *Peranzo* v. *Coughlin*, 608 F.Supp. 1504 (S.D.N.Y.1985) (double EMIT testing held sufficient to satisfy due process); *Storms* v. *Coughlin*, 600 F.Supp. 1214 (S.D.N.Y.1984) (noted that substantial question was raised as to whether EMIT tests were reliable); *Jensen* v. *Lick*, 589 F.Supp. 35 (D.N.D.1984) (prison officials could impose sanctions on prisoners based upon the unconfirmed EMIT test); *Kane* v. *Fair*, 33 Cr.L. 2492 (Mass. Superior Court, August 5, 1983) (the state failed to show that knowledgeable scientists would accept an unsubstantiated EMIT-positive result as evidence of drug use and required the positive result be accompanied by an alternative method of testing); *Smith* v. *State*, 250 Ga. 438, 298 S.E.2d 482 (1983) (the EMIT test is sufficiently reliable to stand as the only evidence in a parole revocation hearing). This Court finds that a substantial issue of material fact exists which precludes granting summary judgment as to Pella's claim that his due process rights were violated."

I should mention that the court ultimately decided that the administration of the test would have been proper even if a requirement of *probable cause* had to be satisfied as seeds and leafy green substances were found in Pella's living area. As the court allowed the disciplinary action, it appears to be thinking along the lines I described in my discussion of *Capua*. Hence, a high value of $\hat{C}$ along with a low standard error may assist a judge in determining the probative value of scientific evidence. Further support for also considering the PVN, especially when other evidence is limited, e.g., when there is only one witness who identifies the defendant, appears in Wicker (1953). He reports that Mr. Frye, who could not introduce the negative polygraph result at trial, was later released from prison after someone else admitted guilt.

Professors Kircher and Raskin provide an informative survey of the potential uses of the polygraph as well as its limitations. I believe their emphasis on the importance of verifying that the estimates of η and θ obtained from mock crime experiments are appropriate for other contexts reinforces the point I made in Section 5. Due to their emphasis on the polygraph they may have overlooked the analysis of the ELISA test data in Table 4 and the 1985 study of drug tests by the Centers for Disease Control, which provide other examples of the need for such studies. As they note, the lack of empirical research on the validity of the polygraph in screening situations rendered it impossible for me to obtain precise values of η and θ. Because we all seriously question the routine use of lie detectors on the general work force, it seems preferable to illustrate the inherent problems with accuracy rates that are overestimates rather than be criticized for "biasing" our analysis. In any case, the one field study I found, by Kleinmutz and Szucko (1984), reported accuracy rates of only 73% that support the general thesis that accuracy rates of any screening test are likely to be lower in the field than the laboratory or certification study.

It is hard to trace the history of the problems of screening tests when used on populations with low base rates. Kircher and Raskin cite Meehl and Rosen (1955) as the first article in the polygraph area. The results appear in the biostatistical literature slightly earlier (Dunn and Greenhouse, 1950) and Professor Greenhouse believes it was known in the vital statistics field before then. Of more significance is the fact that the same lesson has to be taught policymakers and business leaders over and over again in all these areas of application. I appreciate their noting that the sampling error formulas are new to the polygraph literature.

As Dr. Wittes states, the PVN is of primary interest when a highly specific confirmatory test will be given prior to the final diagnosis. In order to have a high PVN, in my discussion of the ELISA test I included the "borderline" group as positive, otherwise η would have been less than .977 and θ greater than .926. Perhaps that point should have been given greater emphasis. As the cases reviewed by Judge Reed in *Pella* tell us, however, a confirmatory test is *not* always required in nonmedical uses of screening test although both Dr. Wittes and I would recommend that they should be.

Dr. Wittes' discussion of the role of the time elapsed from onset of the disease (or development of antibodies) until the test is administered is quite important as the data in Tables 2 and 3 of Simmonds, Peutherer and McClelland (1987) demonstrate. They show that there is substantial variability in the number of weeks from the date infected blood is transfused to the time antibodies are detected in the recipient. Moreover, sometimes the confirmatory test detects the antibodies earlier than the screening test although sometimes the reverse is true.

I have refrained from discussing the *Collins* (yellow Cadillac) case here as there is a substantial legal literature concerning the case, which I reference in Chapter 12 of Gastwirth (1988). Dr. Wittes' point about carefully distinguishing conditional from unconditional probabilities merits careful attention.

Dr. Goldberg's review of the different terminologies is quite important and I am grateful for her clarifying the various uses of the same expression. As Morgan (1984) notes, the more applied fields tend to call

$1 - C$ the false positive rate while the epidemiological literature refers to $1 - \theta$. I believe that our analysis of the data on Danish and Vermont blood donors illustrates her point that "the false positive rates will vary from group to group." The recent study by Kuhnl, Seidl, Ray, Kulkarni and Chandanayingyong (1987) demonstrating that the specificity of the ELISA test varies among racial groups also justifies her emphasis of this point. In a similar vein, the April 24, 1987 issue of the *Weekly Epidemiologic Record* noted that KS occurs in about 27% of AIDS cases although such patients formed 58% of the group used by NIH to determine the sensitivity of the ELISA test. Hence, further studies of the sensitivity may be needed, especially when donated blood is being tested and blood passing only one screen is available for transfusion. The importance of group differences in accuracy rates is illustrated by the recent data from Cincinnati reported by Blanton, Balakrishnan, Dumaswala, Zelinski and Greenwalt (1987). Of 211 blood donors who were *repeatedly reactive* on the ELISA test, only 9 of 102 males and *none* of the 109 females were classified as HIV carriers by the confirmatory Western blot. Apparently, the ELISA test has lower specificity for women than men that implies that the basic parameters should be estimated separately for the sexes. Blanton, Balakrishnan, Dumaswala, Zelinski and Greenwalt (1987) cite other studies, including Sayers, Beatty and Hanson (1986) indicating that women who have had a pregnancy have higher HLA cross-reactivity with the ELISA test. Unfortunately, a large scale study of the accuracy rates for each sex separately does not appear to have been published.

It should be emphasized that requiring repeated reactivity on the ELISA test increases its specificity and hence its PVP. Petricciani (1985) reports the values of η and θ for the three major manufactures as: Abbott ($\eta = .934, \theta = .998$), ENI (99.6, 99.2) and Litton (.989, .996). The sensitivity was determined on AIDS patients although the specificity was estimated from random samples of blood and plasma donors assuming zero prevalence. Hence, the values of θ should be slight underestimates. Even if these specificities apply in the field, the expected PVP is less than .5 if the prevalence is low (.001). Petricciani also describes the development of the test and the recommendations for its use in screening blood. He notes that the three kits were not tested on the same population and that some of the assumptions made are not strictly valid. Given the larger donor samples used to estimate θ, which ranged from 2,000 (Litton) to 16,000 (ENI), it is surprising that accuracy rates by sex and race apparently were not examined.

In cooperation with Professor W. O. Johnson we had initiated a Bayesian analysis similar in spirit to Professor Geisser's discussion. His remarks provide further stimulus for us to give high priority to this research. His discussion of the problem faced by a hospital administrator is new and relevant, although one may need to incorporate latency period considerations before the results will be immediately applicable. It is important to note that the predictive approach advocated by Professor Geisser is the natural one for checking the stability of the parameters over time. For example, the biweekly data in Kuritsky, Rastogi, Faich, Schoor, Menitove, Reilly and Bove (1986) indicate that the fraction of blood donors who were classified as carriers of the virus on the basis of repeated ELISA tests *rose* during the year; however, the rate of *confirmed* positives remained constant. Whether this was due to changes in the manufacturing process of the major supplier (Abbott) that led to a slightly lower specificity or to less qualified or inexperienced users of the kit in the field is unclear. By using data from the recent past one could develop a predictive distribution for the next period's $\hat{p}$. If the observed value is in the extreme right tail of this distribution, we would conclude that the process changed and try to determine what was responsible for it. Incidentally, the parameter p is meaningful as it denotes the total proportion of donated blood that will not be used and $(p - \pi)/p$ is the fraction that is unnecessarily lost by the blood banks. Professor Geisser properly focuses on the fundamentality of the exchangeability assumption. Thus, the predictive approach also requires that η and θ be determined on a population sufficiently similar to the one on which the test will be used.

Dr. Gladen points out that the usefulness of the standard error of $\hat{C}$ depends on the particular application and that its expected value C has the major role. As she notes at the end of her remarks both of us are in basic agreement; however, it is good statistical practice to obtain standard errors of estimates. After examining them we may decide that they are small enough to neglect. Furthermore, the standard errors are useful in determining the sample sizes needed to estimate the accuracy rates η and θ. Because it is relatively easy to obtain healthy individuals and the cost of an ELISA test is only about \$5 to \$10, society can afford to determine θ quite accurately for the major subgroups of the nation. In contrast, we may need to implement a screening test for a new variant of AIDS as soon as possible after a few cases occur in the country. The value of η will have to be taken from other countries or determined from a small sample of cases. This is not a hypothetical problem as cases of HTLV-1, which originally occurred only in southwestern Japan, have been found in the West Indies and the United States (Cappel and Chow, 1987). Fortunately, Kamihara, Nakasima, Oyakawa, Moriuti, Ichimaru, Okuda, Kanamura and Oota (1987) report a sharp decrease in the transmission of the virus in Japan after the

initiation of a mass screening program with a new agglutination method.

To further illustrate the practical relevance of our conclusion that larger sample sizes need to be used to estimate the specificity of screening tests, we note that only 50 controls were used in the Kleinmutz and Szucko (1984) study of the polygraph and 120 were used in data cited by Raskin (1986). In the drug testing literature, van der Slooten and van der Halen (1976) assessed the accuracy of the EMIT method on a sample of 111 drug users and 13 nonuser controls.

Dr. Gladen also mentions that Hilden (1979) noted the need for "clipping" $\hat{\pi}$, the Rogan-Gladen estimate, to $\hat{\pi}_1$. This general phenomenon for estimators of prevalence was made by Mantel (1951) and occurs whenever the parameter being estimated must lie in a restricted interval. Hilden (1979) noted that $\hat{\pi}_1$, will no longer be unbiased and will have a smaller variance than $\hat{\pi}$. Our result gives conditions for both estimators to be asymptotically equivalent. Further simulations may be useful to ascertain when the large sample theory yields a sufficiently accurate approximation. An important observation concerning the use of screening test results in prevalence estimation was made by Hand (1987) who shows that the choice of η and θ that are optimal for minimizing a weighted average of the misclassification error rates does not yield the minimum variance estimator of π.

As all of us who discussed the AIDS application agree that a confirmatory test should be required, it is important to examine the effect of dependence between the screening and confirmatory test (or a second screening test that may be used in the drug testing situation). Assuming that the test results can be scaled to follow a bivariate normal distribution, we report the expected PVP in Table 5 that was patterned on a similar table in Allen (1987) who assumed that the results of the tests are independent.

The results show that even when a near perfect ($\eta_c = 1$, $\theta_c = .998$) confirmatory test is used after a highly accurate ($\eta_s = \theta_s = .99$) screening test, a modest degree of dependence reduces the expected PVP from .98 to .80 when a population with *low prevalence* ($\pi = .001$) is screened. The reason for this is that the probability of a false positive is given by $P_\rho[Z_1 > z_1, Z_2 > z_2]$, where Z_1 and Z_2 follow the bivariate normal law with correlation ρ and the ratio

$$P_\rho[Z_1 > z_1, Z_2 > z_2]/P_0[Z_1 > z_1, Z_2 > z_2]$$

increases as z_1 or z_2 increase. Related results appear in Gastwirth (1987) and a comprehensive treatment of the effect of dependence on statistical procedures is given in Miller (1986). It should be mentioned that the choice of $\rho = .4$ to .5 is based on fitting the data in Wartick, McCarroll and Wiltbank (1987) on blood donors who were given two different screening tests (ELISAs based on different antigen sources). The fitting procedure is described in Gastwirth (1988). The high correlation, $\rho = .87$, was given in the 1985 NIH study for the Vermont donors who were retested with the same ELISA.

TABLE 5

Expected predictive value of a confirmed positive result when there is correlation ρ between the two tests and the prevalence of disease is .001

Specificity of confirmatory test	Expected PVP			
	$\rho = 0$	$\rho = .4$	$\rho = .5$	$\rho = .87$
.99	.908	.533	.435	.170
.995	.952	.660	.556	.240
.998	.980	.799	.712	.387
.999	.990	.876	.812	.533

Notes: the sensitivity (η_s) and specificity (θ_s) of the screening test are assumed to be .99. The confirmatory test is perfectly sensitive ($\eta_c = 1.0$) with specificity (θ_c) given in the first column. The cut-off points for each test are the upper points z_α of the normal curve where $\alpha = 1 - \theta$.

The results in Table 5 show that using the same screening test as a confirmatory one does *not* adequately protect the individual from a false positive diagnosis, as ρ is likely to exceed .5. Hence, some of the drug testing procedures that have received court approval might well be re-examined. The results in Table 5 imply that possible statistical dependence between the specificities of screening and confirmatory tests deserves more attention.

There are two distinct considerations in modeling the dependence between the screening and confirmatory tests. The results in Table 5 arose from considering the joint distribution of the test scores for the nondiseased population, D, and concern the specificity of the joint procedure. Of perhaps greater importance is the sensitivity of the joint procedure. I found only two articles, Sayers, Beatty and Hanson (1986) and Habibi (1987), that present dose response data enabling one to relate the ELISA score to the probability the confirmatory test is positive. Even these data sets are reported in relatively few groups (intervals of ELISA ratios). Because it is unlikely that the dependence of the two tests for members of the diseased population will follow the bivariate normal model (even after dichotomization), we will need either the original raw data (with identifying information removed) or data with more groups and the group means and standard deviations as in Spruill and Gastwirth (1982) to assess their relationship. Moreover, the data should be available separately for each sex—the degree of correlation with the confirmatory Western blot or another screening test may differ for the sexes because the specificity appears to.

Although screening tests have a long history, their use in such diverse areas as criminal law, screening for eligibility for an employment opportunity and medical diagnosis as well as the serious consequence

of a "false positive" classification on an ELISA or drug-use test has raised new questions. In addition to research on purely statistical aspects, such as the design of the experiments used to ascertain the accuracy rates of single or multiple test methods, attention should be given to the development of decision theoretic models incorporating the costs of the misclassification errors to the individuals involved and to society as a whole (including the risk to the public of not detecting a true positive). These factors are playing an increasing role in recent legal decisions as illustrated by the following cases:

McDonell v. *Hunter*, 809 F.2d 1302 (8th Cir. 1987) allowed drug testing of prison employees who have regular contact with prisoners provided that it was done by a uniform or systematic selection process, i.e., employees could be randomly chosen or systematically scheduled but they could not be selected subjectively by the supervisory staff. Otherwise, testing was permitted only upon reasonable suspicion. The testing procedure was required to be accurate and reliable.

Rushton v. *Nebraska Public Power*, 653 F. Supp 1510 (D. Neb. 1987) upheld drug testing for employees who have unescorted access to protected areas of a nuclear plant. Clearly, safety issues played a role in this case as all employees had a "diminished expectation of privacy" due to the level of security required.

Treasury Employees v. *Von Raab*, 2 IER Cases 15 (5th Cir. 1987) upheld drug testing of new applicants and employees applying for a promotion to "sensitive" posts in the Customs Service. The use of a confirmatory test plus the individual's awareness of the tests *prior* to their application were major considerations in this 2 to 1 decision. The testing program did not apply to current employees who were not applying for promotion.

Morgan v. *Harris Trust Co.*, 43 FEP Cases (N.D. Ill. 1987) upheld polygraph testing of employees of a bank who have access to cash.

Local 1812 AFGE v. *U. S. Dept. of State*, 43 FEP Cases 955 (D.D.C. 1987) upheld testing for AIDS antibodies in applicants for the Foreign Service on the grounds that proper medical facilities were unavailable in many posts and that a medical exam, including blood tests, was already part of the existing procedure.

ADDITIONAL REFERENCES

ALLEN, J. R. (1987). Scientific and public health rationales for screening donated blood and plasma for antibody to LAV/HTLV-III. In *AIDS: The Safety of Blood Products* (J. E. Petricciani, I. D. Gust, P. A. Hoppe and W. Krijnen, eds.). Wiley, New York.

BLANTON, M., BALAKRISHNAN, K., DUMASWALA, O., ZELINSKI, K. and GREENWALT, T. J. (1987). HLA antibodies in blood donors with reactive screening tests for antibody to the immunodeficiency virus. *Transfusion* **27** 118–119.

CAPPEL, M. S. and CHOW, J. (1987). HTLV-1-associated lymphoma involving the entire alimentary tract and presenting with an acquired immune deficiency. *Amer. J. Med.* **82** 649–654.

DUNN, J. E., JR. and GREENHOUSE, S. W. (1950). Cancer diagnostic tests: Principles and criteria for development and evaluation. *Public Health Service, Pub. No. 9.* U. S. Government Printing Office, Washington.

EGESDAHL, S. M. (1986). The *Frye* doctrine and relevancy approach controversy: An empirical evaluation. *Georgetown Law J.* **74** 1769–1790.

GASTWIRTH, J. L. (1987). The potential effect of unchecked statistical assumptions: A fault in *San Luis Obispo Mothers for Peace* v. *U. S. Nuclear Regulatory Commission. J. Energy Law and Policy.* To appear.

GASTWIRTH, J. L. (1988). *Statistical Reasoning in Law and Public Policy.* Academic, Orlando, Fla. To appear.

HABIBI, B. (1987). Summary of experience in France regarding screening and confirmatory tests. In *AIDS: The Safety of Blood Products* (J. E. Petricciani, I. D. Gust, P. A. Hoppe and W. Krijnen, eds.). Wiley, New York.

HAND, D. J. (1987). Screening vs. prevalence estimation. *Appl. Statist.* **36** 1–7.

HILDEN, J. (1979). A further comment on "Estimating prevalence from the results of a screening test." *Amer. J. Epidemiol.* **109** 721–722.

KAMIHARA, S., NAKASIMA, S., OYAKAWA, Y., MORIUTI, Y., ICHIMARU, M., OKUDA, H., KANAMURA, M. and OOTA, T. (1987). Transmission of human T cell lymphotropic virus type I by blood transfusion before and after mass screening of sera from seropositive donors. *Vox Sang.* **52** 43–44.

KLEINMUTZ, B. and SZUCKO, J. J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature* **303** 449–450.

KUHNL, P., SEIDL, S., RAY, V., KULKARNI, A. G. and CHANDANAYINGYONG, D. (1987). Human immunodeficiency virus antibody screening in blood donors from India, Nigeria and Thailand. *Vox Sang.* **52** 203–205.

MEEHL, P. and ROSEN, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.* **52** 194–214.

MILLER, R. G., JR. (1986). *Beyond ANOVA, Basics of Applied Statistics.* Wiley, New York.

MORGAN, J. P. (1984). Problems of mass urine screening for misused drugs. *J. Psychoactive Drugs* **16** 305–317.

PETRICCIANI, J. C. (1985). Licensed tests for antibody to HTLV-III: Sensitivity and specificity. *Ann. Int. Med.* **103** 726–729.

SIMMONDS, P., PEUTHERER, J. F. and MCCLELLAND, D. B. L. (1987). LAV/HTLV-III antibody testing: Confirmation methodologies and future prospects. In *AIDS: The Safety of Blood Products* (J. E. Petricciani, I. D. Gust, P. A. Hoppe and W. Krijnen, eds.). Wiley, New York.

SPRUILL, N. and GASTWIRTH, J. L. (1982). On the estimation of the correlation coefficient from grouped data. *J. Amer. Statist. Assoc.* **77** 614–620.

VAN DER SLOOTEN, E. P. J. and VAN DER HALEN, H. J. (1976). Comparison of the EMIT opiate assay and a gas-chromatographic—mass spectrometric determination of morphine and codeine in urine. *Clin. Chem.* **22** 1110–1111.

WARTICK, M. G., MCCARROLL, D. R. and WILTBANK, T. B. (1987). A second discriminator for biological false positive results in enzyme linked immunosorbent assays for antibodies to human immunodeficiency virus (HTLV-III/LAV). *Tranfusion* **27** 109–111.

WICKER, W. (1953). The polygraph truth test and the law of evidence. *Tennessee Law Rev.* **22** 711–727.