Statistical Science
2005, Vol. 20, No. 3, 249–260
DOI 10.1214/088342305000000223

In Search of the Magic Lasso: The Truth About the Polygraph

Stephen E. Fienberg and Paul C. Stern

Abstract. In the wake of controversy over allegations of espionage by Wen Ho Lee, a nuclear scientist at the Department of Energy's Los Alamos National Laboratory, the department ordered that polygraph tests be given to scientists working in similar positions. Soon thereafter, at the request of Congress, the department asked the National Research Council (NRC) to conduct a thorough study of polygraph testing's ability to distinguish accurately between lying and truth-telling across a variety of settings and examinees, even in the face of countermeasures that may be employed to defeat the test. This paper tells some of the story of the work of the Committee to Review the Scientific Evidence on the Polygraph, its report and the reception of that report by the U.S. government and Congress.

Key words and phrases: Bayes theorem, lie detector, meta-analysis, ROC curves, sensitivity, specificity, tabular reporting.

1. INTRODUCTION

Developed almost a century ago, the polygraph—known more colloquially as the *lie detector*— is the most famous in a long line of technological tools and techniques used for detecting deception and determining truth. Unlike earlier techniques, the polygraph makes plausible claims to have a basis in modern science because of its reliance on measures of physiological processes. Over many decades, the polygraph has become, for many in the law enforcement and intelligence communities (including counterintelligence officials), the most valued method for identifying criminals, spies and saboteurs when direct evidence is lacking.

In the wake of controversy over allegations of espionage by Wen Ho Lee, a nuclear scientist at the U.S. Department of Energy's Los Alamos National Laboratory, and whether or not he "failed" polygraphs administered to him in 1998, the department ordered that polygraph tests be given to scientists working in similar positions. Soon thereafter, at the request of Congress, the department asked the National Research Council (NRC) to conduct a thorough study of polygraph testing's ability to distinguish accurately between lying and truth-telling across a variety of settings and examinees, even in the face of countermeasures that may be employed to defeat the test. The Committee to Review the Scientific Evidence on the Polygraph was assembled in the fall of 2000 and held its first meeting in January, 2001. The committee consisted of psychologists, psychophysiologists, lawyers, decision theorists, and statisticians. One of us served as committee chair (SEF) and the other as the study director (PCS).

By the time the committee's study was released in October 2002, the climate of public concern in the United States had been changed drastically as a result of the events of September 11, 2001. Not since the 1950s had the country seen anything like the new level of concern with unseen threats to national security in the form of evildoers—infiltrators, spies, saboteurs, terrorists and the like—who might sneak up on the country unawares and inflict terrible damage. What may be new since then is people's willingness to be-

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science, Department of Statistics, Center for Automated Learning and Discovery and Cylab, Carnegie Mellon University, Pittsburgh, Pensylvania 15213, USA (e-mail: fienberg@stat.cmu.edu). Paul C. Stern is Principal Staff Officer, Division of Behavioral and Social Sciences and Education, National Research Council, Washington, D.C. 10010, USA (e-mail: pstern@nas.edu).

lieve that modern technology can detect these evildoers with precision and before they can do their damage. This belief is promulgated in numerous television dramas that portray polygraph tests and other detection technologies as accurately revealing hidden truths about everything from whether a suitor is lying to prospective parents-in-law to which of many possible suspects has committed some hideous crime. Unfortunately, the belief in precise lie detection lacks a scientific foundation. The best available technologies do not perform nearly as well as people would like or as television programs and movies suggest. This situation is unlikely to change anytime soon.

The NRC committee concluded that, rather than expanding the use of polygraph testing, security agencies and the Department of Energy (DOE) in particular, should seek to minimize the use of the tests and to find strategies for reducing threats to public safety and national security that rely as little as possible on the polygraph. Although the committee's focus was not on the use of the polygraph for forensic purposes, the committee's report suggests that courts that are skeptical about the validity of polygraph evidence are well justified in their skepticism.

Although the committee's report was widely hailed in the newspapers and the scientific community as the death knell to polygraph security testing, it was not exactly welcomed by the government agencies responsible for national security. And the committee's sponsor, the DOE, was slow to respond. In an announcement in the Federal Register in April 2003, the department finally publicly acknowledged the committee's report but said that, in the absence of an alternative to the polygraph, it planned no changes in its polygraph policy. This announcement was greeted with surprise by many in the U.S. Congress, including the two senators from New Mexico, the home of Los Alamos National Laboratory, both of whom serve on the DOE senate oversight committee. The senators scheduled a hearing on DOE's policy, which we describe in more detail below.

The polygraph has a long and checkered scientific and legal history. In the next section, we briefly summarize some of that history and explain the title of this paper. Then we outline the conclusions from the committee's report and, in the penultimate section, we describe the events surrounding a hearing of the Senate Subcommittee on Energy that took place on September 4, 2003.

2. WILLIAM MOULTON MARSTON, THE FRYE CASE, AND WONDER WOMAN

Although precursors for the psychophysiological detection of deception go back to the nineteenth century, we can trace the idea of using physiological recordings—in particular, systolic blood pressure to measure deception in laboratory and legal settings to William Moulton Marston, largely while he was a graduate student at Harvard University from 1915 to 1921.

In the now famous 1923 case of *Frye v. United States*, the defense unsuccessfully attempted to introduce Marston's expert testimony as to the innocence of the defendant on the basis of his systolic blood pressure test. According to Marston [8], Frye was accused of murder in the District of Columbia and, after first denying all knowledge of the event, confessed and provided police with correct details of the killing. A few days later, Frye recanted the confession, claiming that he admitted to the crime because he had been promised a share of the reward for his own conviction. Marston then gave Frye his deception test in a D.C. jail and found his claim of innocence to be truthful. Marston's testimony was excluded at trial and Frye was convicted of murder. His conviction was appealed on the ground that the judge erroneously excluded Marston's testimony. On appeal, the circuit court affirmed the trial judge's ruling on the following basis:

> Just when a scientific principle or discovery crosses the line between the experimental and the demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-organized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. We think the systolic blood pressure deception test has not yet gained such standing. . . .

The Frye "general acceptance" test became the dominant rule governing the admissibility of scientific expert testimony over the next 70 years (see [3]). Most courts refused to admit testimony about polygraph evidence, often with reference to Frye, and this continues to the present day.

Marston went on to prominence as an advocate of the polygraph and later as the creator in 1940 of the

first female comic book action hero—Wonder Woman. Along with her Amazonian training, Wonder Woman was known for the special powers of her equipment, including a magic lasso that "was unbreakable, infinitely stretchable, and could make all who are encircled in it tell the truth." Wonder Woman's magic lasso is an icon of the modern myth of the invincibility of the polygraph. Thus the title of this paper. The truth about the polygraph is that it is far from perfect, even when focused on specific events and when the base rate of deception is likely to be high. Wonder Woman's magic lasso is indeed fiction.

In 1993, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the U.S. Supreme Court outlined the current test for the admissibility of scientific evidence in the federal courts. The Daubert test, codified in the *Federal Rules of Evidence* in 2000, requires trial court judges to be gatekeepers and to evaluate whether the basis for proffered scientific, technical or other specialized knowledge is reliable and valid. Daubert replaced the general acceptance test of Frye, a test still used in many states, including several of the largest states in the country (e.g., New York, California, Illinois and Florida). However, more and more, courts in Frye jurisdictions apply a hybrid test that incorporates a large measure of the sensibilities of Daubert into it. These sensibilities are consistent with most scientists' predilections—that hypotheses gain strength from having survived rigorous testing.

Despite the general consistency in basic outlook about how proffered evidence such as the polygraph must be evaluated on the basis of its scientific merit, actual court decisions regarding the use of polygraphs vary widely. Some of this variability comes from the great range of legal contexts in which polygraph evidence arises. In general, courts divide into three camps regarding the admissibility of polygraph test results: (1) many apply a per se rule of exclusion to such evidence, (2) a significant number of jurisdictions permit the parties to stipulate, prior to the test's administration, to the admissibility of the examiner's opinion concerning the test results, (3) other jurisdictions mandate that trial courts determine the admissibility of polygraphs on a case by case basis and can allow polygraph evidence, subject to the discretion of the trial court.

3. THE NRC STUDY

As we noted at the outset, DOE asked the NRC to conduct a thorough study of the validity of polygraph testing. Although the NRC was asked to focus on uses of the polygraph for personnel security screening, the committee examined all available evidence on polygraph test validity, almost all of which comes from studies of specific-event investigations.

The validity of polygraph testing depends in part on the purpose for which it is used. When it is used for a specific-event investigation (e.g., after a crime), it is possible to ask questions that have little ambiguity (such as, "Did you see the victim on Monday"), so it is clear what counts as a truthful answer. When used for screening, such as to detect spies or members of a terrorist cell, there is no known specific event being investigated, so the questions must be generic (e.g., "Did you ever reveal classified information to an unauthorized person?"). It may not be clear to the examinee or the examiner whether a particular activity justifies a "yes" answer, so that examinees may believe they are lying when providing factually truthful responses, or vice versa. Such ambiguity necessarily decreases the accuracy of the test. Validity is further compromised when tests are used for what might be called prospective screening, for example with people believed to be risks for future illegal activity, because such uses involve making inferences about future behavior on the basis of information about past behaviors that may be quite different (e.g., does visiting a pornographic website, or lying about such activity on a polygraph test, predict future sex offending?).

Our study [1] examined the basic science underlying the physiological measures used in polygraph testing and the available evidence on polygraph accuracy in actual and simulated investigations. With respect to the basic science, the study concluded that, although psychological states associated with deception (e.g., fear of being correctly judged deceptive) do tend to affect the physiological responses that the polygraph measures, many other factors (e.g., fear of being incorrectly judged deceptive) also affect those responses. Such phenomena make polygraph testing intrinsically susceptible to producing erroneous results.

To assess test accuracy, the committee sought all available published and unpublished studies that could provide relevant evidence. The quality of the studies was low, with few exceptions. Moreover, there are inherent limitations to the research methods. Laboratory studies suffer from lack of realism. In particular, the consequences associated with lying or being judged deceptive almost never mirrored the seriousness of these actions in real-world settings in which the polygraph is used. Field studies are limited by the difficulty of identifying the truth against which test results

should be judged and lack of control of extraneous factors. Most of the research, in both the laboratory and in the field, does not fully address key potential threats to validity.

The study found that, with examinees untrained in "countermeasures" designed to "beat" the test, specific-incident polygraph tests "can discriminate lying from truth telling at rates well above chance, though well below perfection" (NRC [1], page 4). For several reasons, however, estimates of accuracy from these studies are almost certainly higher than actual polygraph accuracy of specific-incident testing in the field. Laboratory studies tend to overestimate accuracy because laboratory conditions involve much less variation in test implementation, in the characteristics of examinees, and in the nature and context of investigations than arise in typical field applications. Observational studies of polygraph testing in the field are plagued by selection and measurement biases, such as the inclusion of tests carried out by examiners with knowledge of the evidence and of cases whose outcomes are affected by the examination. In addition, they frequently lack a clear and independent determination of truth and are unlikely to include many of the cases with ambiguous polygraph readings that would contribute to a measure of inaccuracy. Due to these inherent biases, observational field studies are also highly likely to overestimate real-world polygraph accuracy.

How then could the committee summarize 0 all of this evidence in a simple form without appearing to deliver a message consonant with Huff's 1954 book [7]? This was our dilemma. The committee struggled with what simple representation would best capture the features that we extracted from our empirical investigation of polygraph studies and their accuracy. One was Figure 1, a "scatterplot" showing the sensitivity and specificity figures derived from each of the 52 laboratory studies that met the committee's quality criteria (the results of each study are connected by lines), overlaid with curves representing the accuracy scores encompassed by something like the interquartile range of the experimental results. (Since there does not exist a natural definition for quartiles for multivariate data of this nature, the committee first computed the area under the curve for each study, A, rank-ordered the values of A, and then used symmetric curves corresponding to the quartiles for these values. These curves also happen to enclose approximately 50% of the data points as well.) The committee chose this scatterplot, which includes essentially all of the relevant data on accuracy from the studies, rather than a single summary statistic or one with added standard error bounds because we judged it important to make the variability in results visible and because our internal analyses of the characteristics of the studies left us suspicious that the variability was nonrandom. Polygraph accuracy likely depends on unknown specifics of the test situation, and we did not want to create the impression that there is a single number that can appropriately be used to describe polygraph accuracy across situations. The committee was concerned that, if it provided any such number, some would too easily misconstrue that number as the committee's consensus finding about polygraph accuracy. This reluctance to use a summary number was enhanced by our recognition that claims about polygraph accuracy are most commonly expressed in terms of percent of correct results. The usual accuracy index gives a value of 50% for a test that performs at a chance level, and the committee did not want to leave the serious misimpression that the polygraph, which performs better than chance, could therefore catch most of the spies who take it.

The second way of summarizing the data presented calculations of the results of hypothetical series of polygraph examinations in hypothetical large populations with known proportions of liars and truth-tellers. The committee assumed a beyond-the-best-case accuracy for the polygraph—a greater level of accuracy than scientific theory or validation research suggested could be achieved by field polygraph tests, even in specific-incident investigations. A beyond-the-best-case assumption was considered appropriate because it had the same implications for policy as an assumption that might have been closer to accurate. The practical implications of any accuracy level depend on the application for which the test is used. The committee presented two tables, one for a screening application with a very low proportion of deceivers; the other for a criminal investigation setting in which only suspects are tested, so that the proportion of deceivers is much higher. The tables were presented as frequencies in hypothetical populations rather than as probabilities because of research in cognitive psychology that shows that representations as frequencies are best for avoiding cognitive errors among recipients (e.g., see the work of Gigerenzer and his collaborators [5, 6]).

The committee included Table 1 in the executive summary of the report (it is Table S-1 there) to represent beyond-the-best-case polygraph accuracy in two modes, in a screening application to a hypothetical population of 10,000 examinees of whom ten are guilty

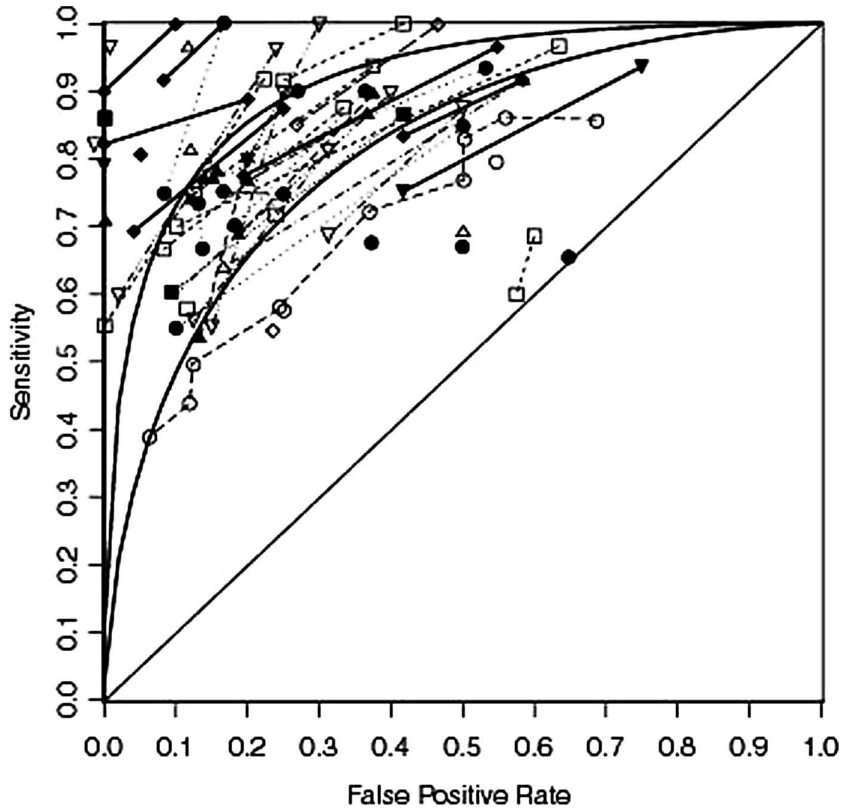

FIG. 1. *Sensitivity and false positive rates in* 52 *laboratory datasets on polygraph validity. Notes*: *Points connected by lines come from the same dataset. The two curves are symmetrical receiver operating characteristic curves with accuracy index* (*A*) *values of* 0.81 *and* 0.91. *Source*: *Figure* 5-1 *from NRC* [1].

of a target offense, such as espionage. In the "suspicious mode," the test is strict enough to correctly identify 80% of deceptive examinees. Accordingly, it identifies eight of the ten spies, but also falsely implicates 1,598 innocent employees. Someone who "failed" this test would have a 99.5% chance of being innocent. Further investigation of all 1,606 people would be needed to find the eight spies. In the "friendly mode," the test protects the innocents. Only about 39 innocent employees would fail the test, but eight of the ten spies would "pass" and be allowed to continue doing damage. This reading of the numbers in the table is

TABLE 1
Expected results of a polygraph test procedure with better than state-of-the-art accuracy in a hypothetical population of 10,000 *National Laboratory employees that includes ten spies*. *Source*: *Adapted from NRC* [1], *Table* S-2

	Examinee's true condition		
Test result	**Spy**	**Nonspy**	**Total**
(A) Suspicious mode: Test is set to detect the great majority (80%) of spies.			
"Fail"	8	1,598	1,606
"Pass"	2	8,392	8,394
Total	10	9,990	10,000
(B) Friendly mode: Test is set to protect the innocent.			
"Fail"	2	39	41
"Pass"	8	9,951	9,959
Total	10	9,990	10,000

a simple illustration of Bayes theorem, but we did not mention this in the executive summary! The committee concluded that for practical security screening applications, polygraph testing is not accurate enough to rely on for detecting deception. The table was later to play a prominent role in the public discussion of the report.

Reasonable people may disagree about whether a test with these properties is accurate enough to use in a particular law enforcement or national security application. We cannot overemphasize, though, that the scientific evidence is clear that polygraph testing is less accurate than these hypothetical results indicate, even for examinees who are untrained in countermeasures. Additionally, it is impossible to tell from the research how much less accurate: accuracy in any particular application depends on factors that remain unknown.

The committee did not discuss how best to represent the findings on polygraph accuracy for other applications. It is possible that for some applications, the difference between a beyond-the-best-case assumption and some other plausible assumption would have been consequential from a policy perspective, and the committee might have chosen a different strategy for representing its findings.

Advocates for the polygraph typically offer two justifications for using polygraph testing as an investigative tool. One is based on validity: the idea that test results accurately indicate whether an examinee is telling the truth in responding to particular questions. The other is based on utility: the idea that examinees, because they believe that deception may be revealed by the test, will be deterred from undesired actions that might later be investigated with the polygraph or induced to admit those actions during a polygraph examination. The two justifications are sometimes confused, as when success at eliciting admissions is used to support the claim that the polygraph is a valid scientific technique.

On the basis of field reports and indirect scientific evidence, we believe that polygraph testing is likely to have some utility for deterring security violations, increasing the frequency of admissions of such violations, deterring employment applications from potentially poor security risks and increasing public confidence in national security organizations. Such utility derives from beliefs about the procedure's validity, which are distinct from actual validity or accuracy. Polygraph screening programs that yield only a small percentage of positive test results might be useful for deterrence, eliciting admissions and related purposes. This does not mean that the test results can be relied on to discriminate between lying and truth-telling among people who do not admit to crimes.

Overconfidence in the polygraph—a belief in its accuracy that goes beyond what is justified by the evidence–presents a significant danger to achieving the objectives for which the polygraph is used. In national security applications, overconfidence in polygraph screening can create a false sense of security among policy makers, employees in sensitive positions and the general public that may in turn lead to inappropriate relaxation of other methods of ensuring security such as periodic security re-investigation and vigilance about potential security violations in facilities that use the polygraph for screening. It can waste public resources by devoting to the polygraph funds and energy that would be better spent on alternative procedures. It can lead to unnecessary loss of competent or highly skilled individuals in security organizations because of suspicions cast on them by false positive polygraph exams or because of their fear of such prospects. And it can lead to credible claims that agencies that use polygraphs are infringing civil liberties for insufficient benefits to the national security.

It may be harmless if television fails to discriminate between science and science fiction, but it is dangerous when government does not know the difference. In our work conducting the National Academies study, we came across many officials in intelligence, counterintelligence, and law enforcement agencies who believe that if there are spies, saboteurs, or terrorists working in sensitive positions in the federal government, the polygraph tests currently used for counterintelligence purposes will find most of them. Many such officials also believe that experienced examiners can easily identify people who are using countermeasures to try to beat the test. Scientific evidence does not support any of these beliefs; in fact, it goes contrary to all of them.

It can also be dangerous if courts or juries are overconfident about polygraph accuracy. If jurors share the misunderstandings that are common among counterintelligence experts and television script writers, they are likely to give undue credence to any polygraph evidence that may be admitted as evidence. The dangers are even greater as polygraph testing expands into forensic applications that are not subject to strong challenge in adversarial processes.

4. THE GOVERNMENT RESPONSE

On April 14, 2003, the DOE issued a notice in the Federal Register to retain the use of polygraph screening in the facilities it operates, saying that the decision

was made in light of the current national security environment, the ongoing military operations in Iraq and the war on terrorism. The notice and the accompanying statement made explicit reference to the committee's report as follows:

> DOE is undertaking this action, among other reasons, to satisfy the requirement of section 3152 of the National Defense Authorization Act for Fiscal Year 2002 to prescribe regulations consistent with the Congressionally-specified purpose of minimizing the potential for release or disclosure of classified data, materials, or information. Section 3152 further requires DOE, in formulating regulations, to take into account the National Academy of Sciences' Polygraph Review.
>
> The NAS Polygraph Review found that scientific evidence regarding polygraph validity is sparse and of low quality. The NAS's main conclusion is that polygraph testing is accurate enough for event specific investigations but that its costs outweigh its benefits when used for employee screening.
>
> However, in the alternative, the NAS report also concluded that if polygraph screening is used as a trigger for detailed follow-up investigation, and is not the basis for adverse personnel action, "[t]here are good theoretical reasons to think appropriate procedures of this sort would improve detection of deception."
>
> The notice of proposed rulemaking proposes to retain DOE's existing regulations because they are consistent with the statutory purpose of minimizing the risk of disclosure of classified data and because they are compatible with the NAS's alternative conclusion that polygraph screening, if applied as a trigger for follow-on tools versus as a basis for personnel action, can improve detection of deception. (DOE Press Release, April 14, 2003).

The Senate Committee on Energy and Natural Resources announced plans for a hearing on September 4, 2003, regarding the DOE Polygraph Program. There were to be two witnesses: Deputy Secretary of Energy Kyle McSlarrow and Stephen E. Fienberg, reporting from the NRC committee. We planned the NRC testimony based on the April 14 DOE notice and basically outlined the extent to which the DOE had distorted or misrepresented the committee's conclusions and recommendations. We also noted that the DOE's decision was incompatible with our recommendations.

The evening before the hearing, as one of us (SEF) was waiting on the airplane on his way to Washington, there was a call on my cell phone from Mr. McSlarrow. He said,"I wanted you to hear directly from me personally that the DOE has changed its mind, and I will explain the details in my testimony tomorrow." A copy of the testimony was faxed to the NRC offices that evening and we had a chance to quickly peruse it in the morning just before the hearing began.

That next morning, the hearing commenced promptly at 10 a.m. Deputy Secretary McSlarrow began his testimony as follows:

> [T]he NAS study was published in October 2002.
>
> As a result of the statutory directive, we published a notice of proposed rulemaking on April 14 of this year. In that notice, the Department indicated its then current intent to continue the current polygraph program under a new rule. At the same time, the Secretary recognized that in the longer term some changes might be appropriate, and therefore we asked explicitly for public comment during a period which ended on June 13 of this year.
>
> The Secretary then directed me to conduct a review of the current policy and to make recommendations based on my review. I have worked closely with the NNSA Administrator and the three directors of the nuclear weapons labs, and I have discussed these issues with counterintelligence professionals, polygraph experts, and as part of that review, I have also had access to classified summaries prepared by other Federal agencies.
>
> I have recently completed that review process. Let me say up front that this is one of the most difficult public policy issues I have had to confront. There is something almost talismanic about polygraphs, and that is something I can personally attest to since both the Secretary and I took a polygraph exam early in our tenure at the Department.
>
> I found many of the NAS's concerns about the validity of polygraph testing to be

> well-taken. I personally discussed this issue, as I know you have, Senator, with many employees of the Department, some of whom feel quite strongly that the polygraph is a dangerous tool that either has or will deprive us of the very talent that is needed to support our national security programs. And yet, as a policymaker, I have concluded that the utility of polygraphs is strong enough to merit their use in certain situations for certain classes of individuals and with certain protections that minimize the legitimate concerns expressed by the National Academy of Sciences, employees of the Department, and other observers.
>
> I am, therefore, recommending to the Secretary that we propose substantial changes to how we use the polygraph in the context of our counterintelligence program. In doing so, I carefully weighed considerations of fairness to employees with national security objectives, and throughout I was guided by the NAS report, a study of considerable rigor and integrity, both in the sense of what it tells us about what we know and do not know about scientific evidence relating to the polygraph and in its willingness to make clear the limitations under which the study was conducted.

He then summarized the changes that he was recommending to the current polygraph program, identifying the considerations he concluded were most important taking into account the NAS report. These changes included scaling back the polygraph program from a current estimated number of 20,000 people to approximately 4,500.

In the questions and answers after this testimony, there was the following exchange with Senator Jeff Bingaman (D-New Mexico):

> Senator BINGAMAN. Thank you very much. Let me also commend you and the Department for taking the academy's study seriously at this point. That is how I interpret your testimony. Frankly, I was not persuaded, when the earlier rulemaking came out, that there had been a serious effort in the Department to review what the academy had come up with and concluded. I appreciate the fact that you have decided to reduce the number of individuals who would be subject to polygraph exams because of what you interpret out of the National Academy study.
>
> I still have a problem, and let me just state it very generally, and then I will ask you a couple of questions. It strikes me that what the academy determined was that use of the polygraph exam as a screening tool was unreliable and that therefore they did not recommend doing so. What you are now concluding is that because of the academy's study, the Department is going to continue to use it as a screening tool but not as much. That seems to me to be better than where we were, but it certainly is not where the logic would lead us.
>
> This table that is in the National Academy study, Table S-1, seemed to me, fairly important in its conclusions. As I read that table, it said that out of a population of 10,000, there would be about 1,600 who would give false positive results or, if given a polygraph test, would be shown to be lying essentially or misleading, but it would be a false indication. They also indicated that if there were ten spies in that group of 10,000, two of those would go undetected.
>
> By reducing the number of people who take the test, we are now saying that, say, 4,500 would be subject to the test; plus, a certain percentage of this pool of 6,000 would be randomly tested. So we would perhaps see 5,000 individuals that would be tested each 5 years. That is just an estimate that seemed to me to be consistent with what you are saying.
>
> So under the academy's table, 5,000 is half of 10,000. Therefore, you would have 800 false positives instead of 1,600. That still seems to me a large number of scientists or others in our employ or in the employ of these contractors who would be placed under suspicion, inappropriately placed under suspicion. And I would be interested in your thoughts as to why that can be justified in light of what the academy has found?
>
> Mr. MCSLARROW. I do not have the table in front of me, but I have stared at it so many times I think it is imprinted on my mind at this point. [Laughter.] The NAS

study moved me. I was in a different place as a matter of personal opinion than I am today. So it moved me probably not as far as the NAS would want me to go, but it did clearly have an effect on my thinking and I think the thinking of the Secretary.

To go to the point, you are quite correct. If all we were doing was reducing the number, we would just reduce the problem. We would not have eliminated the problem of false positives. And that is why I think it is important to think of this as a two-pronged approach. Regardless of what the numbers are, if the only result of a false positive—not that this is unimportant—is to take time and resources of the Department in order to see in a further investigation whether or not there is anything to the false positive, and ultimately a polygraph result, a positive one, does not result in any adverse decision to the employee, then I would argue you have taken care of the false positive problem. You have not addressed, of course, the false negative, which is an entirely different set of problems.

My oral testimony sounded very different from the document we had prepared and which had been made available for the press and others. In it, I (SEF) summarized what we (Paul Stern and SEF) had gleaned to be the main changes in policy being proposed and how they fit with the committee's conclusions and recommendations.

One, do fewer tests. That has two components. Do testing of restricted groups in highly classified settings, and with this is a new definition of what should be top secret. And secondly, random screening, something I will come back to. I want to note that although there are fewer tests, there are still a lot of tests, and with the large number of tests, we still get the attendant false positives and false negatives.

The second change is do less with the results of the tests. Although it was not in his oral testimony, in the written testimony Deputy Secretary McSlarrow talked about treating the results as more akin to anonymous tips than definitive evidence of deception. I think that if the Department got an anonymous tip about an employee, it would not lead to a full-bore investigation. . . . There is clearly a problem with a positive test result because there is not a backup test. Once somebody tests positive on a polygraph, there is nothing that science has to offer for the Department and the security program to do as a follow-up test. Doing another polygraph is simply not good enough.

There is a second component to this change of doing less, and that is that there should be no adverse decision on access based solely on the results of the polygraph and also a recommendation to rely on the polygraph less for accelerated clearance. This is important in light of the false negative problem. I said I wanted to come back to random screening. That was my third item. This is random screening for deterrence. I want to emphasize that in our review of the literature on the polygraph, we found no scientific evidence in support of the deterrent effect of the polygraph. That does not mean that it does not exist. It is that we have never done a serious investigation of it, and especially deterrence for possible spies.

The Deputy Secretary—this is my fourth point—talked about anecdotal reports of admissions. We, in the course of our deliberations, heard many anecdotal reports. We never received written documentation that would allow us to assess them carefully or to put them in context where their scientific usefulness could be assessed. Point five, do more research. I can only applaud the Deputy Secretary's support for our position on this. He suggested in the written testimony that we do not know much about the polygraph, but I want to say that we do know something about its limits and they suggest that it simply is not up to the task that we have before us. We should not expect to make the polygraph better, but we should look for better approaches.

I concluded with the following remarks:

So while there still may be a place for polygraph testing in the labs for investigations and for small numbers of individuals with access to the most highly sensitive classified information, if the test's limited

accuracy is fully acknowledged—and this is what the DOE is now proposing to do at least in part—the question is how limited. In his statement today, Deputy Secretary McSlarrow suggests he agrees with the committee that the broad use of this flawed test for screening will probably do more harm than good, and we believe that national security is too important to be left for such a blunt instrument.

Let me just conclude by reminding you that polygraph testing rests on weak scientific underpinnings despite nearly a century of study. Much of the available evidence for judging its validity lacks scientific rigor, and our committee sifted that evidence and the report makes clear the limitations of the polygraph for the present context. Searching for security risks using the polygraph is not simply like a search for a needle in a haystack. It is true that of the large group of people being checked, only a tiny percentage of the individuals examined may be guilty of the targeted offenses. Unfortunately, tests that are sensitive enough to spot most violators will also mistakenly mark large numbers of innocent test-takers as guilty, and tests that produce few of these types of errors, such as those currently used by the DOE, will not catch the most major security violators and still will incorrectly flag truthful people as deceptive, both kinds of errors. Thus, the haystack analogy fails to recognize that unacceptable tradeoff I mentioned earlier in my testimony.

Our committee concluded that the Government agencies could not justify their reliance on the polygraph for security screening. Today's testimony and the new proposals seem much more consistent with the scientific evidence. As a Nation, we should not allow ourselves to continue to be blinded by the aura of the polygraph. We can and we should do better.

After a series of clarifying questions and answers, the hearing ended with the following exchange:

Senator BINGAMAN. Let me just ask a final question about this Atomic Energy Commission decision in 1953 to terminate their polygraph program at that time. This is information I believe you have come across in your studies. Could you give us any more information on that?

Dr. FIENBERG. At the time that we were preparing our report, I confess I did not know very much about this study. Nobody had brought it to our attention. Actually, as the report was undergoing revision, a historian, Ken Alder, who had done some research on the history of the polygraph, shared with me a number of documents, including some relatively recently declassified documents about what happened. Following the Manhattan Project, the AEC actually began a polygraph screening program at Oak Ridge in the 1940's. It was initiated by Leonard Keeler, who was one of the original creators of the physical machine we call the polygraph today, and at the time the foremost polygraph tester. He started out with a couple of hundred tests, and within six months, they were screening all of the major employees at the labs. At one point—it is a little hard to get the numbers from the released documents—well over 5,000 people were undergoing regular polygraph screening, not by Keeler but by a contractor from the outside. All of this may sound sort of eerily familiar as we look backward. It involved testing managers, scientists, engineers, technical workers, and then later contractors. Initially Keeler found nine people who admitted to having stolen product material as a result of the polygraph tests. When that was subject to closer examination and extended review, it turned out it was all a hoax. In fact, the polygraph had detected nothing at all.

Senator BINGAMAN. The polygraph had detected nothing in the sense that they may have stolen the material?

Dr. FIENBERG. No.

Senator BINGAMAN. Oh, they did not steal any material?

Dr. FIENBERG. It wasn't stolen material. It was a hoax. The polygraph [examiner] believed people when they admitted having done things. By 1952, the hue and cry was so great that the AEC was forced to set in motion a scientific review, and they created a five-person panel of what

> I would label as polygraph-friendly scientists, people who actually had either done studies or supported scientific articles. They reviewed what went on at the Oak Ridge facilities and pointed out that even though the polygraph had considerable value, there were major problems with the program at the time. There was Senate action. Senator Wayne Morse actually spoke out at length against the polygraph, and a bill was introduced in Congress to do a detailed scientific study at one point.
>
> In March 1953, almost 50 years to the day prior to the DOE announcement in the Federal Register, the Atomic Energy Commission issued a statement withdrawing the program as a result of both the objections and the concerns expressed by the polygraph-friendly scientific panel.
>
> So it is an interesting episode. What for me is especially interesting is that in the intervening 50 years, we seem not to have learned much from that original lesson. We did not learn much science, except that maybe we more fully understood the limitations of the polygraph, and we did not learn about the implications of trying to impose a large-scale security screening program on a major facility in the absence of other kinds of security measures.
>
> Senator BINGAMAN. I think that is a useful history for us to have in mind. So thank you very much for relating that, and thanks again for the report and for your testimony.

Whereupon, at 11:17 a.m. the hearing was adjourned. We left the hearing room with other staff from the NRC in amazement. Our report had had a real impact on public policy, although not as big an impact as we might have liked! And clearly, the thought and effort that had gone into the presentation of the data and the results, and in particular into the two tables, had paid. This was not a case of "how to lie with statistics," but the reverse. It was a triumph for careful statistical analysis and presentation, telling the truth about what we know about attempts to detect lying.

5. CONCLUSIONS

At the outset, we explained the seemingly compelling desire for a device that can assist law enforcement and intelligence agencies to identify criminals, spies and saboteurs when direct evidence is lacking. The polygraph has long been touted as such a device. In this article and in the NRC report on which it draws, we explain the limited scientific basis for its use, the deep uncertainty about its level of accuracy and the fragility of the evidence supporting claims of accuracy in any realistic application.

How should society, and the courts in particular, react to such a situation? At a minimum they should be wary about the claimed validity of the polygraph and its alternatives for use in the myriad settings in which they are used or proposed for use. This is especially relevant to current forensic uses of the polygraph. We believe that the courts have been justified in casting a skeptical eye on the relevance and suitability of polygraph test results as legal evidence. Generalizing from the available scientific evidence to the circumstances of a particular polygraph examination is fraught with difficulty. Further, the courts should extend their reluctance to rely upon the polygraph to the many quasi-forensic uses that are emerging, such as in sex offender management programs (see the discussion in [4]). The courts and the legal system should not act as if there is a scientific basis for many, if any, of these uses. They need to hear the truth about lie detection.

As this paper was going to press in January 2005, the Department of Energy finally announced its proposed revised polygraph rules in the Federal Register [2]. They provide a detailed plan for implementing the plan outlined in Deputy Secretary McSlarrow's September 2003 testimony. But no other federal agency has stepped forward with a plan to curb the use of polygraphs. All of them have heard the truth about polygraphs as we know it, but they have failed to acknowledge it by action.

REFERENCES

[1] COMMITTEE TO REVIEW THE SCIENTIFIC EVIDENCE ON THE POLYGRAPH (2003). *The Polygraph and Lie Detection.* National Academy Press, Washington.

[2] DEPARTMENT OF ENERGY (2005). Proposed rules. *Federal Register* **70** 1383–1396. Available at www.regulations.gov/fredpdfs/05-00248.pdf.

[3] FAIGMAN, D. L., PORTER, E. and SAKS, M. J. (1994). Check your crystal ball at the courthouse door, please: Exploring the past, understanding the present, and worrying about the future of scientific evidence. *Cardozo Law Review* **15** 1799–1835.

[4] FAIGMAN, D. L., FIENBERG, S. E. and STERN, P. C. (2003). The limits of the polygaph. *Issues in Science and Technology* Fall 40–46. Available at www.issues.org/issues/20.1/faigman.html.

[5] GIGERENZER, G., TODD, P. M. and THE ABC RESEARCH GROUP (1999). *Simple Heuristics That Make Us Smart*. Oxford Univ. Press.

[6] HOFFRAGE, U., LINDSEY, S., HERTWIG, R. and GIGERENZER, G. (2000). Communicating statistical information. *Science* **290** 2261–2262.

[7] HUFF, D. (1954). *How to Lie with Statistics*. Norton, New York.

[8] MARSTON, W. M. (1938). *The Lie Detector Test*. Richard R. Smith, New York.