A Review of the Office of Technology Assessment Report on Polygraph Validity

Clark McCauley  
*Bryn Mawr College*

Robert F. Forman  
*Haverford Community Hospital*

This article reviews the assessment of polygraphic lie detection produced by the Office of Technology Assessment (OTA) in November 1983. The review argues that use of the lambda statistic to summarize polygraph accuracy in the OTA report was inappropriate because the studies examined differ widely in base rates of guilt and innocence. Using Lykken's (1981) average accuracy statistic and avoiding overlapping data sets found in the OTA report, the review finds that field studies, analog studies, and guilty knowledge studies produce very similar average accuracy (82% to 88%). Guilty knowledge tests tend to be biased toward false negatives, whereas control question tests tend to be biased toward false positives. The striking similarity of results for field and analog studies using control question tests contradicts the common belief that results of analog testing are weak and not generalizable to field practice. It is argued that analog studies are in fact the only hope for improving the validity of polygraph testing.

In November 1983, the Office of Technology Assessment (OTA) of the U.S. Congress released a 132-page technical memorandum entitled *Scientific Validity of Polygraph Testing: A Research Review and Evaluation*. The memorandum was requested by the Committee of Government Operations, U.S. House of Representatives, in response to President Reagan's National Security Decision Directive 84 (NSDD--84), issued March 11, 1983. This directive authorized executive agencies and departments to require em-

---

Requests for reprints should be sent to Clark McCauley, Department of Psychology, Bryn Mawr College, Bryn Mawr, PA 19010.
ployees to take a polygraph examination in the course of investigations of unauthorized disclosures of classified information. As an immediate consequence, approximately 2.5 million government employees and 1.5 million employees at companies that do business with the government became subject to polygraph testing. Although this directive was later rescinded (February 1984), the OTA report has been considered—by other government agencies, American industry, and many polygraph researchers—to be an authoritative evaluation of polygraph testing. In light of the significance of the OTA report, especially in Congressional hearings concerning regulation of polygraph testing, a review of the review is in order.

This article (a) challenges the report's use of the lambda statistic to summarize and compare studies of polygraph accuracy and (b) retables studies reviewed in the OTA report using a more appropriate statistic, Lykken's (1981) "average accuracy." When accuracy is calculated for only the conclusive judgments and when only independent data sets are averaged, an assessment of polygraph accuracy is obtained that differs substantially from the OTA summary.

**ACCURACY MEASURES**

Various statistics have been offered as vehicles for describing the accuracy of polygraph testing. In one common practice, the percentage of examinees accurately identified from the total group of examinees is given. For example, in the first issue of Polygraph, the journal of the American Polygraph Association, Barland and Raskin (1972) summarized the results of their study by stating that "53% of the [subjects] were correctly categorized, 12% were incorrect, and 35% were inconclusive" (p. 24). Unfortunately, such statistics give no information concerning detection of innocence versus detection of guilt. That is, it is not possible to discern the rates of false positives and false negatives from these statistics.

In the OTA report, the shortcomings associated with using an overall accuracy percentage were partly avoided by separately presenting results in terms of detection of guilt, detection of innocence, and inconclusives. But comparisons between studies are awkward and are of limited value when accuracy can be presented only in separate percentages for detection of guilt and detection of innocence, because more of the guilty can usually be identified at the expense of misidentifying more of the innocent. What is needed is a single statistic that describes the accuracy obtained in a particular study so that comparisons between studies can be facilitated.

In the OTA report, lambda was the summary statistic chosen. Lambda is a measure of predictive association that shows the proportional reduction in the probability of error in predicting one category (in this case, deception or
innocence) when a second category (in this case, polygraph examination results) is known. Essentially, lambda provides an index of the percentage improvement in prediction when polygraph examinations are compared to prediction from base-rate information alone (OTA, 1983, p. 51).

Unfortunately, lambda is an inappropriate statistic for comparing and summarizing the studies reviewed in the OTA report. In their article introducing lambda, Goodman and Kruskal (1954) pointed out: "In some cases, particularly when comparisons between different populations are important, the measures lambda a, lambda b, or lambda may not be suitable, since they depend essentially on the marginal frequencies" (p. 745). The studies cited in the OTA report have base rates of guilt and innocence (criterion marginals) that differ widely (e.g., 43% guilty for Bersh's, 1969, zone of comparison test vs. 75% guilty for Raskin, 1976); consequently, lambda is not appropriate for the purposes of comparing or summarizing the studies considered by OTA. Fortunately, an alternative summary statistic is available that avoids the limitations associated with lambda.

Lykken (1981) pointed out that simply averaging the rates of truth detection and lie detection together gives a fair summary of the accuracy obtained in a particular study. For example, consider the following hypothetical situation. A classified document is stolen from a government agency. Only 3 of the agency's 100 employees are involved in the theft. After testing all the employees, the polygraph examiners fail to identify anyone as guilty. In this instance, it could be said that the polygraph investigation was 97% accurate because all innocent employees were accurately diagnosed. It would be specious to present 97% as the accuracy rate because the 3 guilty employees went undetected. Using Lykken's suggestion, we would obtain \((100\% \text{ truth detection } + 0\% \text{ lie detection}) / 2 = 50\%\) as the overall accuracy rate. Scores obtained using Lykken's average accuracy will range between 0% to 100%, with 50% representing chance accuracy. Thus, Lykken's statistic is unaffected by base rates and gives a fair estimate of polygraph accuracy. In our article, Lykken's average accuracy is used to report and compare accuracies obtained in the studies considered in the OTA report.

REEXAMINATION OF POLYGRAPH ACCURACY IN STUDIES REVIEWED BY OTA

Our Table 1 follows OTA Tables 5, 6 and 7 in presenting results separately for field studies (those studies that examine charts obtained from real criminal investigations using some form of control question test), analog studies (those studies conducted with mock or small crimes in an experimental setting and using some form of control question test), and guilty
knowledge studies (those studies conducted in an experimental setting and using some form of the guilty knowledge test).

Table 1 summarizes the results of our reexamination of the studies on which the OTA report was based. The statistics in Table 1 differ from corresponding statistics in the OTA report (OTA Tables 5, 6 and 7) in three respects.

First, three studies were deleted from our consideration because they did not provide independent estimates of detection of both guilty and innocent examinees (Heckel, Brokaw, Salzberg, & Wiggins, 1962; Timm, 1982; Widacki & Horvath, 1978).

Second, our accuracy figures are calculated without inconclusives. In the OTA report, inconclusive judgments were presented separately in Tables 5, 6 and 7, but ultimately inconclusives were treated as errors by OTA (p. 97). Because inconclusives usually lead to retests or suspension of judgment in actual polygraph testing, it is misleading to consider inconclusives as errors. That is, if polygraph tests were 100% accurate except when inconclusive, the question of polygraph validity would be considerably less pressing. A fairer assessment of field practice requires calculation of accuracy rates on

The control question test is the most common question format used in polygraph testing. Essentially, the examinee's response to control questions (e.g., "Before the age of 25, did you ever steal anything from a place you worked?") are compared with responses to relevant questions (e.g., "Did you steal the $200 from the cash register at work?"). The control question is one which is unrelated to the matter under investigation but of similar, though less serious, nature, and one to which the subject will, in all probability, lie; or at least his answer will give him some concern with respect to either its truthfulness or its accuracy. (Buckley, 1980, p. 1191)

If the examinee's response to a specific control question is greater than the examinee's response to a paired relevant question, then a judgment of truthfulness is indicated; an opposite result would indicate deceptiveness.

The control question test is often used as a generic term for a variety of polygraphic techniques that utilize the same kind of control and relevant questions. Other versions of this general format are the lie control test, the zone of comparison test, the quadri-zone test, and the modified general question test.

The guilty knowledge test, originally described by Lykken (1959), has been studied experimentally but has not had wide application in the field practice of polygraphy. In this approach, the examinee is asked in a multiple-choice format questions about which only the investigator and the guilty examinee have knowledge. For example, in an investigation in which stolen money was recovered, a typical guilty knowledge test question might be "Was the money hidden in (a) the back of the store, (b) a car trunk, (c) a wall safe, (d) a filing cabinet, (e) a desk drawer?" The guilty examinee reveals guilty knowledge by producing a greater response to the correct location of the hidden money. As Lykken (1981, p. 298) pointed out, the incorrect alternatives provide a real control for the critical alternative in the guilty knowledge test, whereas the "control question" in the control question test provides only a weak approximation of a true control for the critical question.
TABLE 1
Accuracy Results Without Inconclusives

<table>
<thead>
<tr>
<th>Guilty</th>
<th>Innocent</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOE</td>
<td>NOJ</td>
</tr>
</tbody>
</table>

Field studies
Bersh (1969)
GQT unanimous
32 32 97 36 36 89 93
ZO C unanimous
38 38 89 51 51 94 92
GQT and ZOC unanimous
70 70 93 87 87 92 92
GQT and ZOC majority
34 34 71 25 25 80 75
Total
104 104 86 112 112 89 88
Horvath and Reid (1971)b
Horvath and Hunter (1973)c
Slowick and Buckley (1975)d
Wicklander and Hunter (1975)'
Horvath (1977)b
W. A. Davidson (1979)d
Raskin (1976)
Numericalad
12 77 100 4 21 95 98
Nonnumericalad
12 198 93 4 54 69 81
Barland and Raskin (1976)
Panel
47 40 98 17 11 45 72
Judicial
33 30 100 8 7 43 72
Kleinmuntz and Szucko (1982)b
M (SD) of 9 independent studies
86 (7)
87 (18)
77 (10)
Analog studies
Barland and Raskin (1975)
36 26 88 36 21 71 80
Podlesny and Raskin (1978)
20 17 81 20 19 96 89
Raskin and Hare (1978)
24 21 100 24 19 95 97
Rovner, Raskin, and Kircher (1978)
36 31 90 36 34 85 88
Kircher (1983)
50 32 94 50 39 97 96
Dawson (1980)
12 11 100 12 10 70 85
Bradley and Janisse (1981)
Electrodermal response
96 71 82 96 65 86 84
Heart ratead
96 54 63 96 51 63 63
Szuc k and Kleinmuntz (1981)g
15 90 71 15 90 49 61
Ginton, Daie, Elaad, and Ben-Shakhar (1982)
2 2 100 13 13 85 92
Honts and Hodes (1982a)
No countermeasures
9 6 100 12 6 67 83
Honts and Hodes (1982b)
No countermeasures
19 16 100 19 9 67 83
Hammond (1980)
32 24 96 30 18 67 81
M (SD) of 12 independent studies
92 (9) 78 (14) 85 (9)
Guilty knowledge studies
Lykken (1959)
50 50 88 48 48 100 94
P. O. Davidson (1968)
12 12 92 36 36 100 96
Podlesney and Raskin (1978)
10 10 80 10 8 100 90

(Continued)
the basis of the conclusive judgments made. Consequently, in our Table 1, the percentage correct columns reflect the number of correct judgments divided by the total number of conclusive judgments made. The number of inconclusive judgments in a study is the difference between number of examinees and number of conclusive judgments in Table 1.

Finally, in calculating average percentage truth and lie detection for the studies in Table 1, results from the same examinees entered our average only once, whereas in the OTA report it appears that multiple analyses or combinations of the same data were included in calculation of OTA average percentages.\(^2\) In our Table 1, four versions of Bersh results were not included in the field study average because they were subsets of the "Bersh total" results that were included. Results for Raskin numerical and nonnumerical were not averaged in because both sets of results were based on 16 examinees selected from a larger study by Barland and Raskin (1976; see also Raskin, Barland, & Podlesny, 1978). Barland and Raskin (1976) panel results were included in the average, but results from Barland and Raskin (1976) judicial were not because these two sets of results had 35 examinees in common and the former had the larger number of total examinees. (Nonindependent results are retained in our Table 1—see Table Footnote a—to facilitate comparison with OTA tables.) The upshot is that

\(^1\)Although the OTA (1983) summary indicated that "ten individual field studies" contributed to the averages reported on p. 97 of the OTA report, it is not clear which of the 18 percentages associated with these studies in OTA Table 5 were averaged.
the average percentages for field studies in our Table 1 are based on 9 independent data sets rather than on the "ten individual field studies" contributing to the OTA (1983) averages (p. 97).

Similar considerations arise for the analog studies. Results from Bradley and Janisse (1981) based on electrodermal response (EDR) were included in the average, but heart rate (HR) results for the same examinees were not because EDR appears to be generally the most accurate of the physiological measures (Waid & Orne, 1981) and was more accurate than HR in this study. Results for the guilty examinees using countermeasures in the two studies by Honts and Hodes (1982a, 1982b) are not presented in Table 1 because the other analog studies were conducted with no countermeasures. Thus average percentages for analog studies in our Table 1 are based on 12 independent studies rather than on "fourteen individual analog studies" contributing to the OTA (1983) summary (p. 97).

In the guilty knowledge studies, EDR results from Bradley and Janisse (1981) again entered the average, although HR results from the same examinees did not. And results from the first test conducted on examinees in Balloun and Holmes (1979), rather than results from the second test, were included in the average because the first test obtained considerably higher accuracy. The average percentages for guilty knowledge studies in our Table 1 are thus based on six independent studies.

The mean percentages for the three kinds of study can now be compared. The 9 independent field studies averaged 86% in the detection of guilt, 77% in the detection of innocence, and 82% for Lykken's average accuracy. The 12 independent analog studies averaged 92% in the detection of guilt, 78% in the detection of innocence, and 85% for Lykken's average accuracy. The 6 independent guilty knowledge studies averaged 79% in the detection of guilt, 96% in the detection of innocence, and 88% for Lykken's average accuracy.

These average percentages are noteworthy in several respects. First, the average true positive and true negative rates for field studies are little different in Table 1 (86% and 77%) from those in the OTA (1983) summary (86% and 76%), but for analog studies they are 20% to 30% higher in Table 1 (92% and 78%) than in the OTA summary (64% and 58%). The differences for the analog studies is chiefly attributable to the substantial numbers of inconclusive judgments in these studies, which lowered the OTA percentages as a result of the OTA summary's treating inconclusives as errors.

The second noteworthy aspect of the average percentages in Table 1 is that field studies and analog studies tend to be biased against the innocent, whereas guilty knowledge studies tend to be biased in favor of the innocent. That is, field and analog studies average better in detecting guilt (86% and 92%, respectively) than in detecting innocence (77% and 78%, respective-
ly), whereas guilty knowledge studies average better in detecting innocence
(96%) than in detecting guilt (79%). This pattern has been noted before
(e.g., Waid & Orne, 1981) and is consistent with Lykken's a priori analysis
of the weakness of control question tests such as those used in both the field
and analog studies in Table 1. Briefly, Lykken (1981) argued that the
difficulty of finding a control question ("Did you ever steal anything before
the age of 15?") as arousing to an innocent examinee as the relevant
question ("Did you rob the Friendly Loan Company?") is likely to produce
a relatively high rate of false positives. Indeed, both field and analog studies
in Table 2 show false positive rates averaging about 20% (compared to false
negative rates averaging only about 10%).

The third and most important aspect of Table 1 is the consistency of
Lykken's average accuracy statistic across field, analog, and guilty knowl-
edge studies (82%, 85%, and 88%, respectively). This convergence indi-
cates that the difference between field and analog studies and between
control question and guilty knowledge testing may be less than has been
assumed (Lykken, 1981), although the guilty knowledge test may still be
worth special attention for its bias toward false negatives rather than false
positives. The similarity of average results for field and analog studies is
particularly striking. Field and analog averages differ by only a few
percentage points for true positive rate, true negative rate, and Lykken's
average accuracy, and the standard deviations of the average percentages
are also very close for the two kinds of study. These results lead us to
question the common pessimism about the generalizability of analog studies
to field polygraph practice.

FIELD VERSUS ANALOG TESTING

Belief in the superiority of field research is based on the premise that,
because the "laboratory subject has little to lose" (Abrams, 1972, p. 145) by
being detected, the subject will be less detectible. This belief appears to be
held by many polygraph investigators (Abrams, 1972, p. 145; Berrien, 1939,
p. 542; OTA, 1983, p. 62; Orne, 1972, p. 173), and, as already noted, the
OTA summary (p. 97) offered support for this belief by indicating that field
results averaged 20% to 30% more accurate than analog results. But our
Table 1 indicates that the only difference between field and analog results is
a tendency toward a higher proportion of inconclusives in analog studies.
Average accuracy in field and analog studies could hardly be more similar
once inconclusives are set aside, and we believe these results put the burden
of proof on anyone who would claim superiority of detection in field
polygraph testing. Results aside, it cannot even be maintained that field
studies are methodologically superior because both field and analog studies have obvious methodological problems.

A major problem with any field study is that the criterion can rarely be verified with certainty. Because all field studies involve actual criminal cases, the guilt or innocence of the examinee must come from judicial outcome, confession, or panel reviews of evidentiary files. Unfortunately, judges, juries, and panel review boards are liable to make erroneous judgments. Even confessions can be misleading because they are sometimes the result of plea bargaining ("Plead guilty to this burglary, and we won't prosecute you for these other two") and may have little to do with the actual guilt of the defendant. Consequently, all field studies suffer to some extent from dubious criterion determination—a problem that will tend to reduce observed accuracy.

Another problem common to all field studies is the lack of independence of polygraph judgments and determination of the criterion. When ground truth is established by a judicial panel, the problem is that the panel makes its judgment on the basis of evidentiary files that were also to some extent available to the polygraph examiner. When ground truth is established by judicial outcome or confession, the problem is that polygraph results may have influenced the judicial outcome or may have stimulated the confession. In any field study, therefore, there is at least the potential of criterion contamination that will tend to inflate the accuracy observed (thus opposing the effect on observed accuracy of errors in criterion determination).

Although analog studies solve the criterion problem with certainty, they are usually criticized for being poor models of field polygraph testing. The OTA (1983) report identified six problems concerning the generalizability of analog results to field testing:

The subject population differs, and, apparently most important, the consequences for 'suspects' differ dramatically between the field and laboratory. In addition, in analog studies, the questions and question techniques most often are not tailored to individual examinees. In actual criminal field investigations, case information about the crime and the subject usually provide a basis for tailoring questions. (p. 62)

It is important to note, however, that none of these problems is insurmountable. As Ginton, Daie, Elaad, and Ben-Shakhar (1982) and P. O. Davidson (1968) have shown, it is possible to design crime situations that have real-life consequences for the examinees' actions and subsequent dishonesty. In principle, experienced polygraph examiners may be employed in analog studies, although only the study by Ginton et al. (1982) reported using professional examiners in an analog study. Similarly, testing situations may be created in the laboratory that more faithfully model field
testing situations (e.g., Ginton et al., 1982; P.O. Davidson. 1968). And the concern that analog subjects differ from those involved in actual polygraph testing can be addressed by enlisting analog subjects from appropriate populations. Finally, there is no reason that analog studies could not tailor their questions to background information about the subject being investigated, as analog questions are already tailored to the one mock crime at issue for all examinees.

Given that analog studies have at least the potential of strong external validity, their indisputable advantage is that they enable the researcher to test innovations in polygraph testing. Professional examiners are understandably reluctant to introduce new questioning or scoring techniques in field practice, but new techniques must be compared and evaluated if polygraph accuracy is to be improved. In analog studies, it is possible to examine the relative accuracy of different scoring techniques (Ginton et al., 1982; Szucko & Kleinmuntz, 1981), physiological parameters (Balloun & Holmes, 1979; Bradley & Janisse, 1981; Podlesny & Raskin, 1978), and questioning procedures (Bradley & Janisse, 1981; Podlesny & Raskin, 1978). It appears that analog studies, especially analog studies that more closely model field polygraph practice, have a great potential for improving field technique. Indeed, it seems unlikely that field accuracy can be significantly improved without innovation and development made possible by analog research.

REFERENCES


