Numerical Evaluation and Wise Decisions

Donald J. Krapohl, Brett A. Stern, & Yazmín Bronkema

In the first part of the 20th century, rules for the interpretation of physiological data in lie detection were appearing in scientific journals as scientists began to explore the use of bodily responses for this purpose. The level of scientific attention in the succeeding decades was uneven, but never strong. As a result, many or most of the polygraph chart interpretation rules that found their way into current polygraph practice were developed by the practitioners and sages of the profession, rather than through rigorous scientific methods. These practitioner rules now largely dominate the field, and have been repeated in various books, school handouts, seminar materials, and polygraph publications. A keyword search of the articles in the technical journal *Polygraph* over the last 30 years finds 123 articles have appeared on the topic in that one publication alone, with the majority of these reiterating beliefs of the authors rather than producing convincing data.

When one examines the larger psychophysiological literature, much seems to be known about physiological responding. There is far less agreement among polygraphers, however, as to the best way to evaluate charts. Disagreements and public debates over scoring and decision rules have persisted for decades: which numerical threshold or cutting scores give the best accuracy; against which comparison question should the relevant question be scored, is hyperventilation (or some other particular feature) diagnostic, etc? Despite a lack of consensus among polygraph professionals, these and other scoring questions actually do have correct and defensible answers. They are founded on empirical research and rational decision-making, rather than the historical approach of personal preferences and school traditions. In this paper we will review scientific principles and findings, with an eye toward identifying those that lead to the best results.

We warn that some of the conclusions in this paper will contend with the different schools of thought regarding chart interpretation. This is not necessarily by design, but rather the result of the approach to the problem we have taken. Instead of a recitation of a particular system's scoring rules, or criticism of someone's method, or the introduction of a new means of evaluation, we looked at scoring from a fresh perspective: how can scientific principles be brought to bear on the issue of polygraph accuracy? We base our approach on well-established principles that relate to diagnostic techniques of all types, from medical tests to interpretation of satellite images, from latent fingerprint analysis to, of course, polygraphy.

Our emphasis on the science should not be taken that polygraph data analysis is exclusively a scientific undertaking. In polygraphy or other fields, some small degree of art is helpful. However, exploiting the science before the art permits us to minimize idiosyncratic preferences, abandon unproductive rules, facilitate the training of new examiners, and help move the field toward more general acceptance. Our goal is to urge polygraphers to examine their assumptions about polygraphy, to consider what our sister sciences have to offer us, and to invite the profession to take advantage of those principles and scientific evidence that can improve our practices individually and collectively.

In that vein, let us first discuss some important concepts.

This article originally appeared in Polygraph, 2003, 32(1)

Concepts

Validity

In order to appreciate how scoring practices affect polygraph accuracy, we must understand what this term means, and how it relates to other principles. There are several forms of validity but the one most applicable to polygraph decisions is criterion related validity. This concept refers to how well polygraph decisions match ground truth. It is generally recognized that the polygraph does not detect lies, the criterion of interest. The polygraph merely monitors and records selected physiological functioning within the human body, permitting inferences about a person's veracity. The validity of polygraphy is gauged by how well the underlying phenomenon (differential physiological arousal) predicts deception. If decisions that are based on the phenomenon afford a high degree of agreement with ground truth, it is said to have a high criterion related validity. Conversely, if the phenomenon does not correspond with ground truth, the validity is said to be low. Consequently, the best way to have high validity is to use only phenomena that reliably occur during deception, employ a method of interpretation that identifies and weights them according to their diagnostic value, and use optimized decision rules. Inclusion of unreliable phenomena can erode accuracy, as can sub-optimal scoring and decision rules.

Reliability

A companion concept to validity is reliability. Reliability is the measure of repeatability or reproducibility. There are three forms of reliability: test-retest, intra-rater, and inter-rater reliability. Test-retest reliability relates to testing the same person with the same test on multiple occasions to assess whether the same outcome is reached on all occasions. Intra-rater reliability is how often the same person agrees with himself or herself on multiple occasions with the same data. For example, are the charts interpreted the same on Monday as they were on the previous Friday? The most frequently researched form of polygraph reliability is inter-rater, that is, the degree of agreement among different evaluators who are looking at the same data. Inter-rater agreement is extremely important, and it will set the limit for how valid a technique can be. For example, if there is only 70% agreement among scorers in the field, the accuracy of the technique cannot exceed 70%. This is because at least 30% of the scorers did not get the right answer (assuming it was the 70%, and not the 30% who did get the right answer!) High agreement doesn't mean high accuracy, though. There can be high agreement and low validity. This appears to be the case with the current voice stress technologies used in deception detection. Through practice and exercises, voice stress technicians can achieve high rates of agreement, however their criterion related validity is still poor because the underlying phenomenon is weak or not valid. In the extreme, there can be 100% agreement, and yet the evaluators can be wrong every time. It is important to recognize that high agreement is necessary for, but does not guarantee, high validity.

Variance

Most everything varies. Polygraph scores certainly do. So do IQs, body sizes, pulse rates, amount of sunshine on any given day, number of family members, reaction time, number of sweat pores on the palm, and an uncountable number of other things. Variance must be taken into account when research is done, because samples also have variance between them. The extent to which one can rely on research findings depends upon the source from which samples were taken and how large they were. We often read articles how one scoring system had 90% accuracy while another had 80% or 100% or some other number. Sometimes authors will claim that 90% is better than 80% or worse than 100%, usually to the advantage of the author's argument. As a particularly good (or bad) example of this, in 2001 an author in *Polygraph* characterized a reduction of a mere three inconclusive decisions (from 8 to 5) as a 37.5% change, and touted this percentage as proof of the author's point of view. Had the author conducted a test of proportions, that difference would have been found to be meaningless ($z=0.87$, ns). This example of improper statistical methods, and others like it, prompted the present paper.

Is a finding of 90% really better or worse than other percentages? Is a reduction of 37.5% something we can rely on? If the sample sizes aren't adequate, the differences

are probably not very stable. That's because a different sample might have come up with entirely different percentages, possibly upending previous conclusions. If one had large samples, the finding of a difference becomes more reliable, but even then there should be a replication. Unless the research has at least been replicated with another sample, one can never be sure that the findings extend beyond the original sample. Other factors separating good science from bad science include the sampling method, criterion selection, exclusion criteria, and selection of statistical methods.

Like most everything else, the physiological responses of examinees, and the scores that result from them, also vary. Though deceptive examinees tend to have scores more in the negative direction, and nondeceptive examinees in the positive direction, there is a range and distribution of scores for these groups. Some liars produce extremely negative scores, while other liars tested on the same issue might have scores significantly more in the positive direction. The frequency of scores tends to graph into the shape of the familiar bell curve (See Figure 1). Those scores most often obtained will cluster near the middle of the distribution, and their frequency tapers off as the scores become more extreme. The most negative score obtainable in the 7-position manual scoring system with three charts and three relevant questions is -81. It is exceptionally unusual for anyone to have a score this low. Similarly, a nondeceptive person with the same number of questions and charts can receive a total of +81 points. Again, it is exceedingly rare to see scores this high. It is this relationship between scores and their frequency that produce the bell curve.

Figure 1. Bell curve.

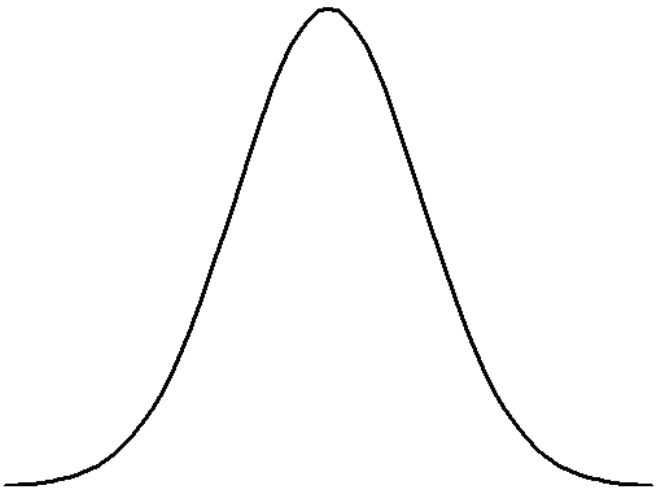


Figure 1 is the idealized bell curve, though there has been scant attention paid to whether polygraph scores produce such a perfect shape or one that is more skewed. Nevertheless, the following principles are fairly robust, and will apply to all except markedly non-normal distributions.

The evidence suggests that the scores from deceptive examinees fall mostly in one bell curve, and the scores from nondeceptive examinees in another. When a technique has poor diagnosticity, the two bell curves will be heavily overlapped. Using Figure 2 as a hypothetical example, the distribution of scores for the liars covers an area shared largely with those of truthtellers. Because of this overlap, the diagnosticity of the technique is poor, regardless of where one placed the cutting score.

Figure 2. Two hypothetical distributions, suggesting poor diagnosticity for a test.

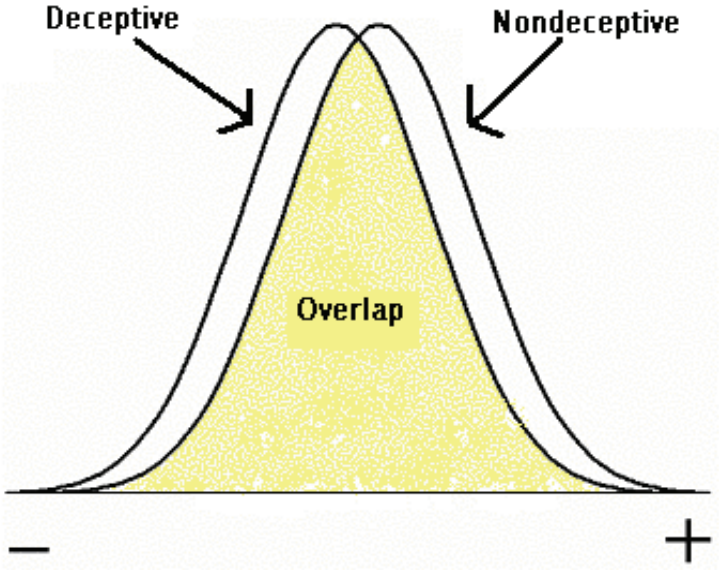
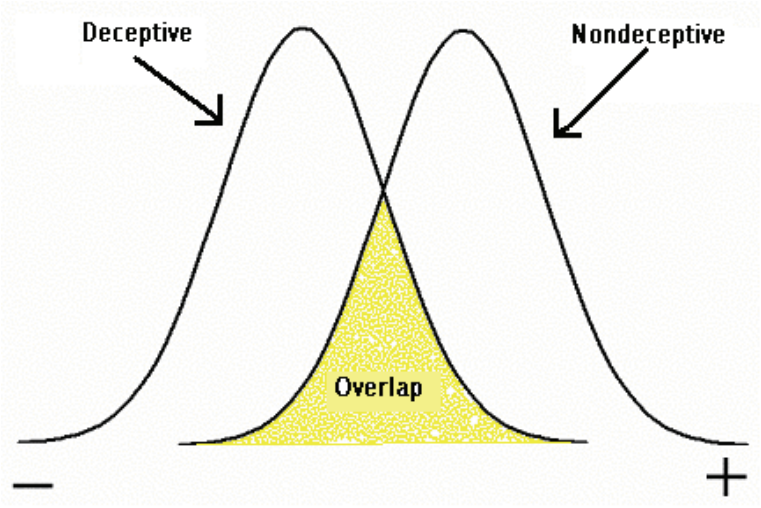


Figure 3 on the following page shows two distributions that have less overlapping area. If the method used valid features and rules to create the distributions, any cutting score used with Figure 3 would always outperform any cutting score with Figure 2 in terms of proportions of correct decisions.

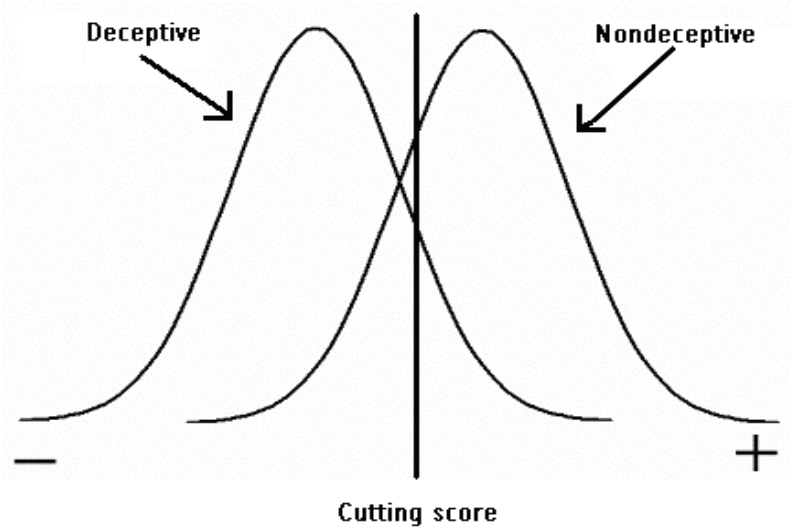
Figure 3. Two hypothetical distributions indicating relatively good diagnosticity.



A careful look at these graphs will show that cutting scores do not really affect accuracy as much as they affect the type of errors that are made. For example, moving a cutting score more toward the positive direction permits the scorer to capture a larger portion of the deceptive examinees (See Figure 4). However, there is a cost of

misclassifying nondeceptive examinees. Similarly, shifting the cutting score the other way helps identify more truthtellers, but causes a loss of detecting liars. Catching more of one type means losing some of the other type. Most examiners grasp this relationship intuitively. Nevertheless, there are a number of research articles from examiners reporting efforts to find "perfect" cutting scores that result in the least number of errors. As Figure 4 reveals, such efforts have been, and will always be unsuccessful because the underlying distributions always overlap in deception testing.

Figure 4. Graph depicting how adjusting a cutting score to increase detection of deception affects the detection of truthfulness.



Though shifting the cutting scores may be a poor method of improving accuracy, there are other methods that are effective. One such method is to include only the most diagnostic information in the scoring system. As more predictive physiological features are added to the polygraph scoring system, the distributions of scores of the truthful and deceptive examinees are pulled in opposite directions. As they shift away from one another, there is less overlap, and consequently less error across cutting scores. Compare Figures 2 and 3. Figure 3 shows distributions further apart, indicating that the scores were based on more predictive features than those in Figure 2, thereby improving accuracy in a way that adjusted cutting scores cannot.

A note of caution is warranted here. Merely adding more scoring features, scoring rules, or increasingly complex processes to the mix will not move the two bell curves apart. They may even make the process of manual chart interpretation more difficult and less objective. Only valid features have a positive effect. When the number of features is much larger than perhaps a dozen or so, marginally effective or even ineffective features are creeping into the model. This is especially the case when working with physiological data, which tends to have high inter-subject variability.

If "more" meant "better" for polygraph features, one could hypothetically develop a scoring system with a thousand features, and claim that the larger number makes that system superior to those with fewer than a thousand features. The fallacy of this line of reasoning should be obvious, and it is the rarest of diagnostic fields that can claim to have more than 20 individual features that are found reliable enough for use by human evaluators. As a case in point, in a technical report to the US government, Harris, Horner, and McQuarrie (2001) found that 22 of the manual scoring criteria reported by Swinford (1999) could be reduced to four and deliver the same information, the remaining features

being either redundant or ineffective. Similarly, the simple three-feature Objective Scoring System enjoyed better accuracy than human scorers of the same data sets that had 22 features in their scoring system (Krapohl & McManus, 1999; Krapohl & Norris, 2000).

There is also a human factor to consider. Unlike computer algorithms, which can accommodate extremely complex calculations with perfect reliability, the reliability of human decision-making correlates with simplicity. Increasing complexity erodes reliability among human scorers, and as discussed earlier, this reliability is essential for validity. Adding rules, features, or decision rules can, at some point, diminish accuracy. It is an example of when more is less.

Next, let us consider inconclusive outcomes. How one comes to inconclusive decisions is another factor that can affect accuracy. The wider the inconclusive band, the fewer errors are made. Looking again at Figure 3, one can see that it is possible to have a scoring system that produces no errors, though the inconclusive rate might be as high as 70% or more. Having a very narrow inconclusive zone will increase the number of correct decisions, and also incorrect decisions because, numerically, more definitive decisions are being rendered.

This relationship between inconclusives and error reveals that the question "How accurate is the polygraph test?" is overly simplistic. The most accurate answer to this question is, it depends. It can be as high as 100% accurate, or as low as perhaps 80%. Wide inconclusive zones decrease error, potentially approaching zero when the inconclusive area is very large. Or, with no inconclusives, accuracy is in the area of about 80%. Compare polygraphy to the field of latent fingerprints. The common wisdom is that fingerprinting produces virtually perfect decisions. The rarely discussed other side of this coin is that the technique also only produces decisions in a small minority of cases. That is because latent prints are usually of insufficient quality to be helpful. This is one reason law enforcement will only make an effort to search for them in the most serious of crimes. If polygraphy were permitted an inconclusive rate similar to that of latent fingerprinting, accuracies might well be similar.

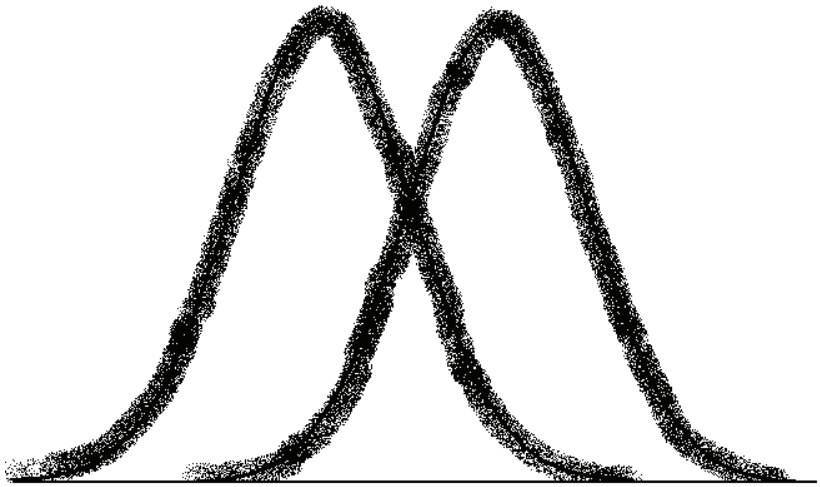
With this understanding of the relationship among cutting scores, accuracy, and inconclusives, we are now ready to contemplate the best approach to determining numerical thresholds, or cutting scores. There are three core issues for deciding where cutting scores should be placed. They are the proportion of false positive errors, false negative errors, and inconclusive results the consumer of the polygraph results can tolerate. Examiners can control only two of these three factors with their decision rules. It is not possible to simultaneously have no errors and no inconclusives when a test has imperfect validity. One can choose to have low false negatives and inconclusives, but it must be paid for in false positives. Or, it is possible to have low inconclusives, but an increase in errors is inevitable. Or, a zero error rate is achievable, but the reader will know by now that the inconclusive rate will be unacceptably high.

Complicating matters more, variability also occurs among scorers. For those who have not seen it themselves, a scoring exercise at the next polygraph association meeting could be an eye-opening experience: scores can vary in some cases to produce opposite outcomes. Variability in scores is not a trivial problem. If every scorer arrived at the exact same score, at least reliability would have been achieved, and possibly validity. Perfect reliability is now only possible with systems that rely exclusively on measurements, such as with Lykken scoring (Lykken, 1959), the Objective Scoring System (Krapohl & McManus, 1999), the Rank Order Scoring System (Honts & Driscoll, 1988), and any of the automated computer algorithms. None of the semi-objective scoring systems have demonstrated the potential of achieving this reliability.

The inter-scorer variability that accompanies the semi-objective field scoring methods in common practice makes setting fixed cutting scores problematic. Scorers who come from the same training, or work in the same agency tend to have better agreement than those scorers who do not.

However, even among those trained in the same methods, there are almost always differences in the composite or total scores when analyzing the same charts. This makes for a somewhat fuzzy bell curve of scores, and highlights the challenge of using universal fixed cutting scores for scorings that vary from scorer to scorer. Figure 5 characterizes the problem of cutting scores and scorer variability. As one can see, this effect reduces the reliability of any estimate of polygraph accuracy with manual scores and any set of cutting scores.

Figure 5. Two hypothetical distributions of polygraph scores showing the blurred curves produced by scorers whose scores vary for the same cases.



Levels of Rules

To maximize the efficacy of polygraph decisions, it is useful to look at the problem hierarchically. The problem begins at its base with the selection of polygraph features for scoring, followed by determining how numbers should be assigned to those features. Then, decision rules using those numbers must be formulated so that the best decisions result. It is easy to recognize that working in any other order is less effective: testing cutting scores before deciding on scoring features results in little useful information. The following sections are organized in this fashion.

Features

Scientists looking at the physiological data have found 10 polygraph tracing features reliable for manual scoring. Below are listed these individual features, along with the supporting citations from the peer reviewed literature or official government sponsored research:

Respiration

1. Suppression (including apnea)

   Barland & Raskin (1975)
   Cutrow, Parks, Lucas, & Thomas (1972)
   Harris, J.C., Horner, A., McQuarrie,D.R. (2000)
   Nakayama (1984)
   Patrick & Iaconno (1991).
   Wakamatsu & Yoshizumi (1968)

2. Increase in cycle time (decrease in the cyclic rate/slowing)

   Barland & Raskin (1975).
   Cutrow, Parks, Lucas, & Thomas (1972)
   Patrick & Iaconno (1991).

3. Change in inhalation/exhalation ratio

   Benussi, V. (1914)
   Burtt, H.E. (1921a)
   Burtt, H.E. (1921b)
   Landis & Gullette (1925)

4. Baseline rise

   Harris, Horner, & McQuarrie (2000)
   Kircher & Raskin (1988)

Electrodermal

5. Amplitude of phasic response[1]

   Harris, Horner, & McQuarrie (2000)
   Kircher & Raskin (1988)
   Kugelmass, et al (1968)
   Patrick & Iaconno (1991)
   Podlesny & Truslow (1993)

6. Duration of response

   Kircher & Raskin (1988)
   Podlesny & Truslow (1993)

7. Complexity of response

   Harris, Horner, & McQuarrie (2000)
   Kircher & Raskin (1988)

Cardiovascular

8. Baseline amplitude increase

   Barland & Raskin (1975).
   Harris, Horner, & McQuarrie (2000)
   Kircher & Raskin (1988)
   Podlesny & Truslow (1993)

9. Duration of response

   Harris, Horner, & McQuarrie (2000)
   Kircher & Raskin (1988)

Vasomotor

10. Reduction of pulse wave amplitude

    Kircher & Raskin (1988)
    Patrick & Iaconno (1991)

Some polygraph schools teach that there are more, sometimes dozens more, diagnostic polygraph features for manual scoring. We often hear of a polygraph examiner reporting that he or she has seen other physiological response patterns beyond these 10 on a given set of charts, and when the examiner confronted the examinee, a confession of guilt was elicited. Anecdotes and selective recollections fall far short of proof of a relationship between that particular response pattern and ground truth for most examinees, however. Though perhaps providing the makings of an interesting case, experiences like this tell little about what is generally true for all examinees. With manual numerical evaluation and current instruments, it is unlikely that more than a few important and reliable new diagnostic features will be identified by scientists any time soon.

While not amenable to human interpretation, there are data contained within the traditional polygraph channels that have shown promise as additional criteria. Foremost is respiration line length (RLL; Timm, 1982). RLL is the measure of the length of the respiration waveform over a specified period of time. RLL is a summary measure that captures respiratory suppression, change in inhalation-exhalation ratio, and increases in cycle time in a single value. So diagnostic is RLL that Harris, Horner, and McQuarrie (2000) determined that it could replace all other respiratory features currently taught. However, RLL is difficult to measure manually, and a channel that displays RLL in a meaningful way is not available on any commercially available polygraphs. For this reason, RLL is taught in only a few polygraph schools.

A second feature that has value is pulse deceleration (Patrick & Iaconno, 1991). For a brief period after stimulus onset, a deceptive response is often associated with a slowing of the pulse. Because most polygraphs are not currently configured to separate pulse rates from the more complex cardiograph waveform, human evaluators are less able to recognize these decelerations

[1]Note: This is but a partial list. Virtually no study has failed to find EDR amplitudes to be diagnostic.

when they occur. With the advent of computer polygraphs, pulse deceleration and RLL could easily become additional data channels for examiners, and we are hopeful that manufacturers will see the potential in adding them.

As suggested earlier, the intractable problem that prohibits adding tracing features for manual scoring beyond the 10 above is that these additional features are usually suitable only for a very small number of examinees, are but irrelevant or counterindicative for most others. Take for example hyperventilation. For the rare examinee, deception may be accompanied by a noticeable increase in breathing rate for a few cycles. For virtually all other examinees, the increase in breathing rate is the type of random variation that is characteristic of physiological data in general, or perhaps a deliberate manipulation of respiration. The increased rate hardly ever means anything in terms of a person's veracity, but can be merely a normal fluctuation in breathing behavior unrelated to deception. Hyperventilation would certainly be diagnostic for some small subset of examinees, but distinguishing for whom it is diagnostic from the much larger group for which it is not is a problem no one has yet solved. Returning to the bell curves earlier in this article for clarification, because hyperventilation is not reliably diagnostic, it does not move the two bell curves apart -- it does not increase accuracy nor decrease error. It can be characterized as noise, and there are literally dozens of similar examples of individual features (e.g. premature ventricular contractions, etc.) taught in scoring systems across the profession. One day there will be additional data sensors added to the polygraph. Those sensors will be selected by how much the information they provide moves the two bell curves apart. Currently, the 10 listed above have been shown by the scientific method to be the most reliable features for manual scoring. No others taught in the field meet this high standard.

Number Assignment

There are two main approaches to manual numerical evaluation: rank order, and the 7-position numerical scale. Rank order scoring is not as widely practiced, and we will not expend much space discussing it, except to note that it does afford potentially outstanding inter-rater agreement because of its simplicity. Those interested in further reading on rank order scoring approaches are invited to read articles by Gordon and Cochetti (1987), Honts and Driscoll (1987), Krapohl, Dutton and Ryan (2001), and Miritello (1999).

Most polygraphers are familiar with the 7-position numerical scale. Traditional 7-position scoring has a notable similarity to the Likert scale (Likert, Roslow & Murphy, 1934), the tool used in psychology for over 65 years to measure attitudes. In the Likert scale, the choices are typically: strongly agree, agree, neutral, disagree, and strongly disagree. The respondent's choices are converted to scores on a 5-position scale, and attitudes thereby quantified. Analogous to the Likert scale is the polygraphic 7-position scoring: –3 (strongly deceptive), -2 (deceptive), -1 (somewhat deceptive), 0 (neutral), +1 (somewhat truthful), +2 (truthful), and +3 (strongly truthful). Though the computation rules are different for Likert and polygraph scoring scales, the longevity of the Likert number-assignment strategy is cause for reassurance that the basis of the 7-position polygraph scoring system is sound.

Though there are subtle differences in all scoring systems, a point of significant divergence among polygraph practitioners is the question of which comparison question should be used when scoring a given relevant question. The approach promoted in the Utah scoring system is to score each relevant question against the comparison question that immediately precedes it on the test. The Utah method also systematically rotate questions, to ensure each relevant question can be scored against each comparison question over the course of the testing (Raskin & Honts, 2001). In the Federal method, a relevant question is scored against the stronger of two adjacent comparison questions, if the relevant question is bracketed by them (DoDPI, 2001). Otherwise, the relevant question is scored against the nearest comparison question. In the Backster system, the scoring decision relies on whether the examinee responded significantly to the relevant question (Matte, 1996). If so, it is scored against the least

reactive of the adjacent comparison questions. When the relevant question does not evoke a significant physiological response, it is scored against the stronger of the two adjacent comparison questions.

These are dissimilar approaches, and one might expect that they would have different effects on the scoring data. Absent conclusive data, it would be premature to posit which of the three performs best. It may prove to be the case that none is the best, but to achieve similar accuracies for both truthful and deceptive examinees, cutting scores will be different among the methods. This possible outcome would be consistent with the overlapping bell curves premise. A research project currently underway by the present authors will attempt to find an answer.

There is at least one important difference among these approaches that is worthy of comment. For two of the methods, Backster and Federal, examiners must choose which comparison question to use based on subjective assessments of response significance and intensity. The Utah method avoids this potential source of scorer variability by using only the comparison question immediately preceding the relevant question. From a psychometric point of view, the Utah method is more scientifically defensible because it reduces individual subjective judgments. This is an attractive advantage, and one worth serious consideration.

There may also be benefit in taking the Utah question-rotation strategy an additional step, and change the positions of both the relevant and comparison questions with each chart, but in opposite directions. With each new chart, the relevant questions could be rotated forward, and the comparison questions backward, or vice versa. In addition to permitting each relevant question to be scored against a different comparison question, this method should moderate an effect on scores that may arise from within-chart habituation (Olsen & Harris, 1998). The suggested double-rotation method is more unwieldy than static question sequences, but it mitigates two possible sources of noise variance in polygraph scoring.

Some polygraph examiners score charts with the 3-position scoring system, an abbreviated version of the 7-position system. Instead of assigning scores between –3 and +3, the range of scores in the 3-position system are between –1 and +1. Because studies have shown that field polygraph examiners assign values between –1 and +1 about 80% - 90% of the time when using the 7-position method, the 3-position system is the de facto method of scoring for the majority of cases (Capps & Ansley, 1992; Krapohl, 1998). One advantage of the 3-position system is that it may reduce some of the subjectivity associated with scoring. Another is that it can be performed much faster because there are fewer judgments to make. However, because there is a more restricted range of scores with the 3-position scoring system, there is a higher proportion of inconclusive calls when the cutting scores are not adjusted. Opposite decisions between the 3- and 7-position scoring systems are exceptionally rare.

Decision Rules

Once numbers have been effectively assigned to the individual response comparisons, it is essential to use decision rules that optimize the outcomes. Readers may have noticed that we have couched our language so to leave open the possibility of something few in the profession have considered: using the same cutting scores for all circumstances will not yield maximum benefit. While examiners reading this idea have a moment to contemplate this heresy, we should like to articulate our rationale for proposing it.

Decision errors are inevitable in any diagnostic technique, including polygraphy. Use of a fixed set of cutting scores for all cases implies that the user understands and accepts the error rates these cutting scores incur, and that these cutting scores minimize the types of errors the user wants to avoid. It seems patently unlikely that the costs of errors are equal in all settings, and under certain conditions, the best decision could be to suspend judgment (i.e., make an inconclusive or No Opinion call), even when numerical scores call for a definitive decision of deception or nondeception.

There is a long-standing axiom in the polygraph profession that states, "Believe in your charts." Yet, we advocate here that even when a numerical threshold is met it might be more prudent to suspend rendering a definitive call—at least for the moment. Consider the following examples:

Example 1.

The examiner scores the test charts, and would be prepared to make a decision of No Deception Indicated (NDI) based on very positive scores. However, the examiner notes a couple of anomalies in the charts. One is that the pneumograph scores are highly positive, while the other two channels are moderately negative, though the net effect is positive composite totals when all channels are summed. Also, the respiration responses to the comparison questions appear nearly identical to one another. Should the call be NDI, or should the examiner suspend judgment?

Example 2.

The examiner tests four individuals, one of whom most certainly must have committed the crime in question. Three of the four examinees are clearly NDI by a wide margin, a call that is supported with Concealed Information Tests (CITs). The fourth examinee just barely meets the NDI numerical threshold. The CITs indicate that the fourth examinee knows more about the crime than he should. Should the call be NDI for the fourth examinee, or should the examiner suspend judgment?

Example 3.

In a murder case, the examiner very scrupulously scores the charts, which would lead to a call of Deception Indicated (DI), but by a single point. The final polygraph results will shape the prosecutor's decision as to whether to seek the death penalty. Should the call be DI, or should the examiner suspend judgment?

Example 4.

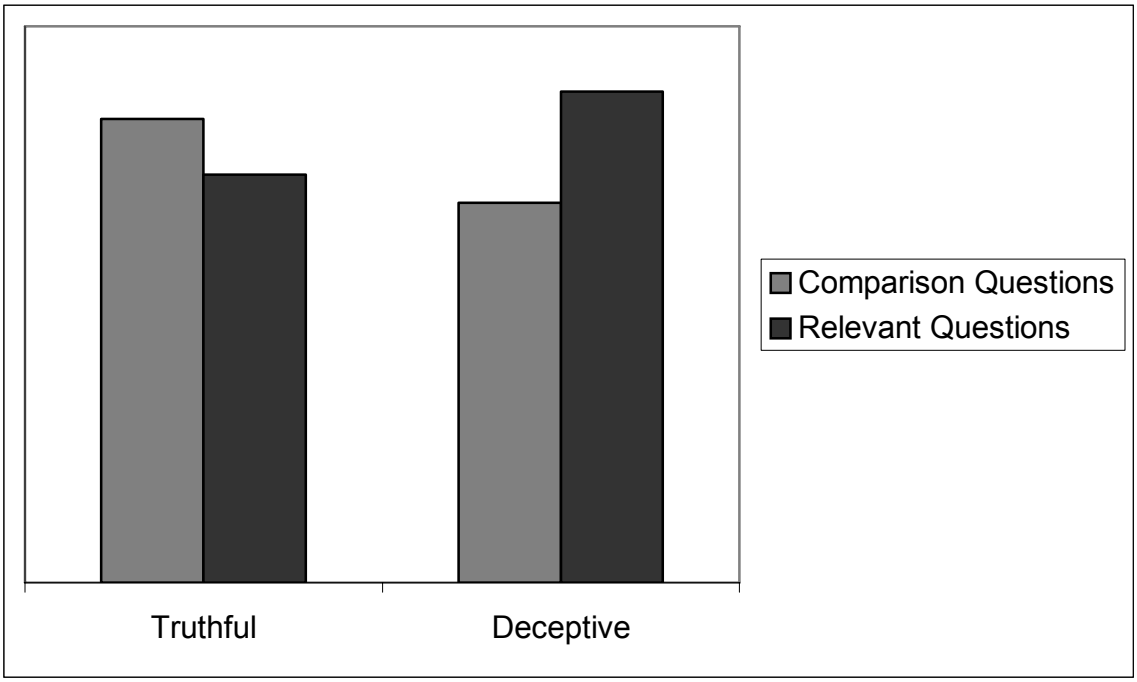
During the pretest interview of a pre-employment screening test the examinee reported extensive drug use. Subsequent test results warrant a call of NDI, however, the weakest analysis spot score was on the question dealing with involvement with illegal drugs. While printing the examinee's last chart, the examiner queries him for his impression of his test results, and the examinee volunteers that he believes he didn't pass the issue regarding drug involvement. Would such a disclosure warrant further exploration—perhaps through a specialized test focusing specifically on drug involvement?

Please observe from these examples that we do not suggest making opposite calls from what scoring may indicate, but that attention paid to non-scoring factors might cause the prudent examiner to withhold a decision until a retest is completed. Though more demanding of resources, waiting for the results of a retest under some circumstances could increase an examiner's accuracy.

There is another factor to consider when establishing cutting scores. Research over the last 25 years has repeatedly found that the responses of liars are not the reverse of those of truthtellers (Franz, 1989; Krapohl, 1998; Raskin, Kircher, Honts, & Horowitz, 1988). Liars, on average, give stronger reactions to relevant questions than truthtellers, on average, give to comparison questions (See Figure 4). Examiners know intuitively from experience that it is much easier to identify liars than truthtellers. Few, however, are aware that the reason is because the underlying phenomenon they are scoring, differential responding, is not symmetrical.

Some scoring systems have cutting scores symmetrical around zero, usually +/- 6. Imposing symmetrical cutting scores on an asymmetrical phenomenon means that accuracy rates will not be equal for both liars and truthtellers. One group will be detected better than the other. Because scores from field polygraph cases are normally shifted in the negative direction, symmetrical cutting scores result in better detection of liars than of truthtellers. To balance the accuracies, it would be necessary to move the cutting score for truthful decisions more toward zero, or the cutting score for DI calls more in the negative direction. However, an important lesson that bell curves teach us is that moving the cutting scores does not improve overall accuracy. It only changes the kinds of errors that result from the cutting scores.

Figure 4. Average relative response intensity to relevant and comparison questions for truthful and deceptive examinees by individual question comparisons, in arbitrary units. Data from Krapohl, Dutton, & Ryan (2001)



The unbalanced accuracy rates are also affected by the Spot Score Rule (SSR). The SSR is a decision rule that depends not only on the total score of the case, but the total score of each individual question (Light, 1999). If any one question indicates deception, though it is essentially the same question as the other relevant questions, the call is DI, even if the other questions score highly positive. The SSR makes the test even more sensitive to deception, while proportionately reducing its sensitivity to truthfulness (as again predicted by the bell curves.)

So, what are the best decision rules? Remember, the three factors for setting cutting scores are: the proportion of false positive errors, false negative errors, and inconclusive outcomes that the consumer can tolerate. Only two of these three can be controlled at one time. As a general rule, for fewer false negatives, the Backster system and the Federal cutting scores along with the SSR are probably the best. For more equal proportions, the Utah method is a step in the right direction. There is no scoring system for single-issue examinations in common practice with decision rules that favor the detection of truthfulness over the detection of deception. As a final thought, one might consider the available automated algorithms, some of which have been validated. Each has perfect reliability, and a validity that equals or exceeds the performance of experienced polygraphers in blind scoring polygraph charts.

Conclusion

It may come as a surprise to some that the most important player in the development of "optimal" decision rules is the consumer of the polygraph results, not polygraphers. It is the consumer who must weigh the intrinsic costs of errors and inconclusives, and it is possible to align polygraph cutting scores to correspond with the consumer's acceptance of risk. Judicious flexibility in decision rules has advantages in error containment. However, fixed decision rules have the benefit of increasing decision agreement across examiners. The fixed-threshold approach may be desirable, even necessary within agencies, though not necessarily between agencies or across the profession.

We hope we have dispelled the notion that optimization can be found simply by adding unvetted tracing features or scoring rules to the scoring system. The evidence here should make clear that byzantine manual scoring systems are to be avoided in favor of more simple and elegant methods. Diagnostic decision-making is always a difficult task, and is made more problematic by unscientific approaches sometimes seen in the field.

Polygraphy is in the proud company of medicine, psychiatry, engineering, forensic sciences, and the other fields that attempt to make critical diagnostic decisions with very complex and noisy data. With 30 years of good polygraph research to draw upon, combined with rational decision principles borrowed from other fields, the profession may be poised for the adoption of a universal scoring system, one that will be empirically tested, and will consider essential factors such as base rates, and costs and probabilities of errors. In addition, it may selectively add more automation, to increase reliability and precision. The goal is to make the process more simple, reliable, accurate, and useful than our current state of practice.

References

Barland, G.H., & Raskin, D.C. (1975). An evaluation of field techniques in detection of deception. *Psychophysiology*, 12, 321-330.

Benussi, V. (1914). Die atmungssymptome der lüge [The respiratory symptoms of lying]. *Archiv fuer die Gesamte Psychologie*, 31, 244-273. English translation printed in 1975 in *Polygraph*, 4 (1), 52-76.

Burtt, H.E. (1921a). The inspiration/expiration ratio during truth and falsehood. *Journal of Experimental Psychology,* 4(1), 1-23.

Burtt, H.E. (1921b). Further technique for inspiration/expiration ratios. *Journal of Experimental Psychology,* 4, 106-111.

Capps, M.H., & Ansley, N. (1992). Comparison of two scoring scales. *Polygraph*, 21(1), 39-43.

Cutrow, R.J., Parks, A. Lucas, N. & Thomas, K. (1972). The objective use of multiple physiological indices in the detection of deception. *Psychophysiology*, 9, 578-588.

DoDPI (2001, January). *Test Data Analysis (Manual Scoring Process): 7-Position Numerical Analysis Scale*.

Franz, M.L. (1989). *Technical report: Relative contributions of physiological recordings to detect deception*. Contract Number MDA 904-88-M-6612.

Gordon, N.J., & Cochetti, P.M. (1987). The horizontal scoring system. *Polygraph*, 16(2), 116-125.

Harris, J.C., Horner, A., McQuarrie,D.R. (2000). *An Evaluation of the Criteria Taught by the Department of Defense Polygraph Institute for Interpreting Polygraph Examinations*. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-00-7272.

Honts, C.R., & Driscoll, L.N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16(4), 241-257.

Kircher, J.C., & Raskin, D.C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73, 291-302.

Krapohl, D.J. (1998). Short report: Proposed method for scoring electrodermal responses. *Polygraph*, 27(1), 82-84

Krapohl, D.J. (1998). A comparison of 3- and 7- position scoring scales with laboratory data. *Polygraph*, 27(3), 210-218

Krapohl, D.J., Dutton, D.W., & Ryan, A.H. (2001). The Rank Order Scoring System: Replication and extension with field data. *Polygraph,* 30(3). 172-181.

Krapohl, D.J., & McManus B. (1999). An objective method for manually scoring polygraph data. *Polygraph*, 28(3), 209-222.

Kugelmass, Sl, Lieblich, Il, Beh-Ishai, A., Opatowski, A, & Kaplan, M. (1968). Experimental evaluation of Galvanic Skin Response and blood pressure change indices during criminal interrogation. *Journal of Criminal Law, Criminology, and Police Science*, 59(4), 632-635.

Landis, C. & Gullette, R. (1925). Studies of emotional reactions: III. Systolic blood pressure and inspiration-expiration ratios. *Journal of Comparative Psychology*, 5, 221-253.

Light, G.D. (1999). Numerical evaluation of the Army zone comparison test. Polygraph, 28(1), 37-45.

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology*, 5, 228-238.

Lykken, D.T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43(6), 385-388. Reprinted in *Polygraph*, 7(2), (1978), 123-128.

Matte, J.A. (1996). *Forensic Psychophysiology Using the Polygraph: Scientific Truth Verification – Lie Detection*. Williamsville, NY: J.A.M. Publications.

Miritello, K. (1999). Rank order analysis. *Polygraph*, 28(1), 74-76.

Nakayama, M. (1984). Suppression of respiration on the critical item and the rebound component. *Reports of the National Institute of Police Science*, 40, 32-37. Abstract in English.

Olsen, D., & Harris, J. (1998). *Study of habituation and other systematic changes in responses during field polygraph tests*. Johns Hopkins University Applied Physics Laboratory. SSD/PL-98-0368.

Patrick, C.J., & Iaconno, W.G. (1991). A comparison of field and laboratory polygraphs in the detection of deception. *Psychophysiology*, 28(6), 632-638.

Podlesny, J.A., Truslow, C.M. (1993). Validity of an expanded-issue (Modified General Question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78(5). 788-797.

Raskin, D.C., & Honts, C.R. (2001). The comparison question test. In M. Kleiner (ed.) *Handbook of Polygraph Testing*. Academic Press: London.

Raskin, D.C., Kircher, J.C., Honts, C.R., & Horowitz, S.W. (1988*). A study on the validity of polygraph examinations in criminal investigations*. Final report to the National Institute of Justice. Grant No. 85-IJ-CX-0040.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven position numerical analysis scale at the DoD Polygraph Institute. *Polygraph,* 28(1), 10-27.

Timm, H.W. (1979). The effect of placebos and feedback on the detection of deception. *Dissertation Abstracts International*, 40(12-B). University Microfilm No. AAD80-13808.

Wakamatsu, T., & Yoshizumi, K. (1968). Study on respiratory wave during polygraph examination. *Reports of the National Research Institute of Police Science*, 21(3), 158-164. Abstract in English.