Low Base Rate Screening Survival Analysis[1] & Successive Hurdles

Mark Handler[2] AAPP Research & Information Chair

Greetings my fellow AAPP members. I hope this finds you all healthy, happy and blessed. I wanted to take a minute to share with you a concept that we, as examiners, may not really think about in our day to day work – *the Base Rate Phenomenon*.

When I was a law enforcement examiner at a large sheriff's office I ran many screening polygraph exams and never gave the base rate phenomenon much thought, if at all. I had read the research on the accuracy of the techniques I used, either the L.E.P.E.T. early on and then later the DLST with ESS scoring. I used the standard target questions most other agencies used, without giving any consideration for this thing we call the base rate for my test questions. Since then I have thought considerably about this and wanted to share some of those thoughts. President Wardwell asked me to write up this paper for the association.

To begin our discussion, we need to define 'base rate'. Base rate will simply be how much of something we are looking for exists in our testing population. If we are conducting specific-issue test for our police department or for an attorney, it would be the prior probability the subject is actually Guilty[3] of the crime. If we are screening (i.e. PCSOT or public safety pre-employment) it is still the prior probability of Guilt, but can also be regarded as the proportion of the examinees who are lying to one or more of the test questions.

Depending on the situation, the prior probability can be higher or lower than chance (where chance is 50%). In a criminal-specific setting we hope the referring detectives selected a subject who is more likely than not to be the culprit. This would make the prior probability greater than chance. But in a screening setting, depending on the targets, the base rate may be significantly lower than chance. For example, testing targets like felony crimes, recent illegal drug possession, or physical domestic violence in a law enforcement pre-employment screening setting may incur low base rates. In a PCSOT monitoring test the base rate of reoffending is probably very low. In a security screening setting, we hope the base rate of espionage, terrorism, or sabotage is very, very low.

Ultimately credibility assessment testing should help inform a decision-maker, consumer, or end-user. That test result should help him or her make a better decision. If we believe that credibility assessment testing has diagnostic value, *then the test results should matter*. But how do base rates of Guilt, sensitivity to Guilt, specificity to Innocence, false-positive and false-negative rates affect the confidence of those test results? What does it mean when someone "passes" or "fails" a credibility assessment test? How much confidence can we or our consumer place in our test result?

1. The original idea for "screening survival analysis' came from my good friend Dr. Charles Honts.
2. Mark Handler sits on the Technical Advisory Group for the Converus Corporation who manufactures the EyeDetect Technology, an unpaid position that offers some stock options should the company go public.
3. The terms "guilty" and "innocent" refer to the actual state of the subject. I ask your kind indulgence of my use of these terms here to denote ground truth states.

Some of these concepts can be tough to wrap our heads around. I will do my best to help you understand how credibility assessment tools can improve something called the *Outcome Confidence* of a test result. I will limit the examples to screening situations, though these phenomena exist in all testing. These are not new ideas. Dr. David Raskin testified before the United States Senate Committee on Armed Services on these very concerns (Raskin, 1984). Drs. Kircher and Raskin have been discussing this underappreciation of the base rate phenomenon for many years (Kircher & Raskin, 1987; Raskin, 1987).

I want to reinvigorate discussion on this topic because it is often overlooked in basic polygraph school and even in advanced training seminars. My hope is to give new and experienced examiners another look at these concepts so they may appreciate how they affect the confidence in their work product. I also want to open the door to dialogue about accuracy limits and successive hurdles. Polygraph accuracy estimates have not changed much over the years and are likely to remain fairly stable. Having alternative technologies with reportedly good accuracies, can improve the Outcome Confidence in polygraph test results by changing the base rates of the testing population. Let's take a look.

We'll focus on using a screening test for a group of job applicants with a low base rate of guilt. I hope to convince you that the test results do matter and can better inform your end-user. Credibility assessment results are far better than doing nothing. In fact, research shows that a typical person can detect deception in another person during the course of an interview about 54% of the time- this is near chance (Bond & DePaulo, 2006). Wise use of evidence-based technology can significantly increase the credibility assessment process when compared to the standard human resources interview. Using multiple *different* technologies can further improve this process.

In order to discuss these ideas, we have to use technologies that have published scientific support. For this discussion, I will reference EyeDetect® and polygraph because these tools have peer-reviewed published accuracy and error estimates. The accuracy values for EyeDetect are taken from studies Dr. David Raskin presented at the American Polygraph Association (APA) meeting in 2015 (Raskin, 2015). The values for polygraph are from Table 2 in the APA meta-analytic review (APA, 2012).

To begin, we need estimates for sensitivity (TP), specificity (TN), false positive (FP) and false negative (FN) for EyeDetect and polygraph. By definition, TN means True Negative or correct results with Innocent cases. FN means False Negative or incorrect results with Guilty cases. TP means True Positive or correct results with Guilty cases. FP means False Positive or incorrect results with Innocent cases.

I list the estimates we will use in Table 1. The Ground Truth column is the actual state of the subject and to be consistent with the language of published scientific studies, these are referred to as Innocent and Guilty. I realize in a criminal justice setting, these states are determined by the trier of fact. The columns Passed Test and Failed Test state the test result and show the accuracy and error estimates for when someone passes or fails an EyeDetect or polygraph test.

Table 1. Accuracy Rates for EyeDetect (Raskin, 2015) and PDD (APA 2012 table 2)		
Ground Truth	Pass Test	Fail Test
EyeDetect		
Innocent	0.88 (TN)	0.12 (FP)
Guilt	0.17 (FN)	0.83 (TP)
PDD		
Innocent	0.72 (TN)	0.14 (FP)
Guilt	0.08 (FN)	0.81 (TP)

Case 1: Equal Base Rates and Equal Accuracies

To introduce the concepts of base rates and Outcome Confidence, a simple example is used in Table 2. Here the accuracy values are not specific to EyeDetect or polygraph. We assume a value of 90% accuracy for Innocent and Guilty subjects and there will be no consideration of an inconclusive outcome.

These conditions are applied to a group of 1000 applicants - 500 Innocent and 500 Guilty. In other words, the base rate of Guilt is 50% or .50. These values also apply in a setting where the prior probability of Guilt for a single test subject is .50.

Table 2. Contingency Table with equal accuracy (90%) and equal base rates (50%)			
Ground Truth	Pass Test	Fail Test	Totals
Innocent	450 (TN)	50 (FP)	500
Guilty	50 (FN)	450 (TP)	500
Totals	500	500	1000
Outcome Confidence	**0.9 (NPV)**	**0.9 (PPV)**	

The numbers in the bottom row labelled “Outcome Confidence” indicate the confidence in the accuracy of the various test results and these proportions have statistical names. The proportion of correct truthful outcomes to total truthful outcomes (450/500) is known as the Negative Predictive Value (NPV). The proportion of correct deceptive outcomes to the total number of deceptive outcomes (450/500) is known as the Positive Predictive Value (PPV). Notably in this example with equal base rates, the Outcome Confidence directly mirrors the accuracy of the test (.9 or 90% for both Pass Test and Fail Test results).

Case 2: Equal Base Rates of Guilty and Innocent using EyeDetect Accuracy

Table 3 illustrates a second case where base rates are equal, but the screening test is conducted using EyeDetect and we will use its reported accuracy and error rates (Raskin, 2015). EyeDetect has a published accuracy of 83% with Guilty subjects (TP) and 88% with Innocent subjects (TN). The FP rate is 12% and FN rate is 17%. There are no inconclusive rates for EyeDetect.

This case, and the next (Case 3), are used to illustrate how slightly different screening tool accuracies and errors affect the Outcome Confidence where base rates are equal. You can see the EyeDetect and polygraph Outcome Confidence is similar for both a passed and failed test.

Table 3. Contingency Table with equal base rates (50%) with EyeDetect (Guilty = 83%, Innocent = 88%, FP = 12%, FN = 17%)

Ground Truth	Pass Test	Fail Test	Totals
Innocent	440 (TN)	60 (FP)	500
Guilty	85 (FN)	415 (TP)	500
Totals	525	475	1000
Outcome Confidence	**0.84 (NPV)**	**0.87 (PPV)**	

The Outcome Confidence is fairly balanced for passing or failing because the base rates of Guilt and Innocent are equal. One notable thing from Table 3 is that although the test is more accurate with Innocent subjects, there is greater confidence in Failed Test outcomes (PPV = .87) than with Passed Test outcomes (NPV = .84). This is because the number of Guilty and Innocent subjects who pass or fail the test changes disproportionally due to the imbalanced accuracy and error rates. This imbalance is reflected in the slight trade-off in Outcome Confidence.

Case 3: Equal Base Rates of Guilty and Innocent using Polygraph Accuracy

Table 4 illustrates a third case with equal base rates, but where the screening test was conducted with polygraph. Polygraph accuracy and error rates are 81% with Guilty subjects (TP) and 72% accurate with Innocent subjects (TN). The FP rate is 14% and FN rate is 8% (APA, 2012).

Note: There are inconclusive rates for polygraph and inconclusive results are shown in parentheses. Inconclusive results do not affect the Outcome Confidence, though they do have a potential to affect the utility of the testing technique. If inconclusive results were counted as errors, they would affect the Outcome Confidence. Agencies that take action based on inconclusive results would have to factor that into these calculations and recalculate the Outcome Confidence. I chose to leave them out in my calculations.

Table 4. Contingency Table with equal base rates (50%) with Polygraph (Guilty = 81%, Innocent = 72%, FP = 14%, FN = 8%)

Ground Truth	Pass Test	Fail Test	Totals
Innocent	360 (TN)	70 (FP)	430 (70 inconclusive)
Guilty	40 (FN)	405 (TP)	445 (55 inconclusive)
Totals	400	475	875* (125 inconclusive)
Outcome Confidence	**0.90 (NPV)**	**0.85 (PPV)**	

*Does not add to 1000 because polygraph has inconclusive results.

In Table 4, it is notable that although the test is more accurate with Guilty subjects, there is a greater degree of confidence in Pass Test outcomes (NPV = .90) than with Fail Test outcomes (PPV = .85) due to the imbalanced accuracies as previously mentioned. This imbalance is reflected in a trade-off in Outcome Confidence. Also one can appreciate there is only a slight difference in Outcome Confidence between EyeDetect and polygraph testing.

Case 4: Low Base Rate (20%) with EyeDetect and Polygraph accuracies

Tables 5 and 6 illustrate the example where the target behavior occurs in only 20% of test subjects. While this target behavior base rate may seem low, it may be quite representative of many credibility assessment testing situations. Agencies should give serious consideration to the potential base rate of their testing targets to better estimate the confidence in their test results.

Table 5 shows the example of a low (20%) base rate target using EyeDetect. In this example, disqualifying behavior occurs in 20% of test subjects and the Outcome Confidence in a Pass Test outcome is very high (NPV = .95). The Outcome Confidence in a Fail Test is lower (PPV = .63). This implies that 37% of the subjects that failed the test are actually Innocent. If an agency selects a low base rate target, they can expect similar results – even with a highly accurate test.

Table 5. Contingency Table with low base rate of Guilt (20%) EyeDetect accuracies (Guilty = 83%, Innocent = 88%, FP = 12%, FN = 17%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	704 (TN)	96 (FP)	800
Guilty	34 (FN)	166 (TP)	200
Totals	738	262	1000
Outcome Confidence	**0.95 (NPV)**	**0.63 (PPV)**	

In the following Table 6, EyeDetect accuracies have been replaced with polygraph accuracies using the same assumption of 20% for the target group base rate of guilt. In this case, using polygraph, the Outcome Confidence in a Pass Test outcome is slightly higher (NPV = .97) as compared to EyeDetect (NPV = .95). But the Outcome Confidence in a Fail Test is slightly lower (PPV = .59) than it was for EyeDetect (PPV = .63). About 41% of the subjects who failed the polygraph test are actually Innocent. Also, we excluded the 134 (13%) inconclusive polygraph results from these calculations.

Table 6. Contingency Table with low base rate of Guilt (20%) with Polygraph (Guilty = 81%, Innocent = 72%, False Positive = 14%, False Negative = 8%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	576 (TN)	112 (FP)	688 (112 inconclusive)
Guilty	16 (FN)	162 (TP)	178 (22 inconclusive)
Totals	592	274	866* (134 inconclusive)
Outcome Confidence	**0.97 (NPV)**	**0.59 (PPV)**	

*Does not add to 1000 because polygraph has inconclusive results.

Case 5: Successive Hurdles with Low (20%) base rate of Guilt - reducing the polygraph work load model.

Now we want to show how using multiple technologies can add information to the decision-making process using the Outcome Confidence. As mentioned, accuracies for polygraph are unlikely to change much and so in any unbalanced base rate setting our Outcome Confidence is likely to remain low for a "failed test". We really can't hope for a lot more of an increase in accuracy, but we can adjust the base rate of Guilt for those we test with polygraph to improve the Outcome Confidence.

For the next example, a group of 1000 applicants are screened using two technologies in a *successive hurdles* approach (Meehl, & Rosen, 1955). For this example, we want to keep the base rate of guilt for the disqualifying behavior at 20%. The first screening tool we use is EyeDetect; results are shown in Table 7. We will see that using the EyeDetect first will alter that base rate of Guilt upward, improving our polygraph Outcome Confidence. Recall, it is the low base rate that drives the low Outcome Confidence because many of the positive test results are false-positive results. That is because we are testing primarily Innocent subjects. For these analyses I will assume the two technologies are independent.

Table 7. First Hurdle contingency table with low base rate of Guilt (20%) with EyeDetect. (Guilty = 83%, Innocent = 88%, FP = 12%, FN = 17%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	704 (TN)	96 (FP)	800
Guilty	34 (FN)	166 (TP)	200
Totals	738	262	1000
Outcome Confidence	**0.95 (NPV)**	**0.63 (PPV)**	

In this setting, if a subject passes the EyeDetect test, the Outcome Confidence is 95%. There are very few false-negative results and most of the negative results are true-negatives. In this case where the Outcome Confidence is 95% for those passing the test, the use of one screening tool may be sufficient to move those subjects to the next phase in the hiring process.

If, however, a subject fails the EyeDetect test, there is only a 63% confidence in that outcome. There are agencies that may decide that a 63% Outcome Confidence is not sufficient to warrant disqualifying the subject that failed the EyeDetect test. In that case, those failing the EyeDetect test can be moved to polygraph as a second test or hurdle.

Of the 262 people that failed the EyeDetect test, there are 166 Guilty and 96 Innocent subjects. But now in the next test (polygraph), the base rate of Guilt is 63%. EyeDetect testing raised the base rate of guilt for disqualifying behavior in test subjects from 20% to 63%. Raising the base rate of Guilt changes the Outcome Confidence for the polygraph test results - let's take a look at table 8.

Table 8. Contingency Table with a base rate of Guilt of 63% with Polygraph (Guilty = 81%, Innocent = 72%, FP = 14%, FN = 8%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	69 (TN)	13 (FP)	82 (14 are inconclusive)
Guilty	13 (FN)	134 (TP)	147 (19 are inconclusive)
Totals	82	147	229 (33 are inconclusive)
Outcome Confidence	**0.84 (NPV)**	**0.91 (PPV)**	

If a subject failed the polygraph test after failing the EyeDetect test, the Outcome Confidence of guilt is 91%. By applying polygraph as the second hurdle, the Outcome Confidence of failing the test increased from 63% to 91%! This increase in Outcome Confidence can be helpful to the decision-maker, consumer or end-user.

Question: What if someone failed the EyeDetect but then passed the polygraph test?

Recall from Table 7 that 262 people failed the EyeDetect test—166 were Guilty and 96 were Innocent. About 82 of those 262 people (31%) will pass the polygraph test after failing the EyeDetect. This is inevitable because all psychometric testing is imperfect. Also, it would not be desirable to have perfect correlation between the EyeDetect and the polygraph. If they agreed perfectly, anyone failing the EyeDetect would fail the polygraph. Anyone passing the EyeDetect would go on to pass the polygraph test and we would not realize the benefits of the two different technologies.

But even in the case where the technologies disagree, the Outcome Confidence in the passed test is still quite high at 84%. Thoughtful consideration around what this means seems advised. Perhaps at this point the agency may wish to conduct a more thorough investigation on these select individuals.

Case 6: Successive Hurdles with Low (20%) base rate of Guilt - reducing false negative results model.

What if we wanted to take a conservative approach and minimize the false negative rate of the final applicant pool? To do this we could retest anyone who passed the first technology using a second technology. What would this look like in a successive hurdles model? What would the cost versus benefits be?

For table 9 let's take only the people who passed the EyeDetect in the first round of testing and move them on to polygraph. Remember they still have a low (20%) starting base rate of Guilt before the EyeDetect.

Table 9. First Hurdle contingency table with low base rate of Guilt (20%) with EyeDetect. (Guilty = 83%, Innocent = 88%, FP = 12%, FN = 17%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	704 (TN)	96 (FP)	800
Guilty	34 (FN)	166 (TP)	200
Totals	738	262	1000
Outcome Confidence	**0.95 (NPV)**	**0.63 (PPV)**	

738 of the original 1000 passed the EyeDetect and move on to polygraph. Of those 704 were actually Innocent and 34 were actually Guilty. The base rate of Guilt moving on to polygraph is very low at 5% (34/738 subjects). Look at what happens in Table 10.

Table 10. Contingency Table with a base rate of Guilt of 5% with Polygraph.
(Guilty = 81%, Innocent = 72%, FP = 14%, FN = 8%)

Ground Truth	Pass Test	Fail test	Totals
Innocent	507 (TN)	99 (FP)	606 (98 are inconclusive)
Guilty	3 (FN)	28 (TP)	31 (3 are inconclusive)
Totals	509	127	637 (101 are inconclusive)
Outcome Confidence	**0.99 (NPV)**	**0.22 (PPV)**	

The Outcome Confidence for a passed polygraph after a passed EyeDetect is extremely high-99%. If we decide our testing goals are to keep the false negative results to a minimum, the conservative approach is one good way to do it. But we run into the concept of diminishing returns at the cost of considerable effort. In this model we end up polygraph testing 738 of the original 1000 applicants whereas in the earlier model we only had to polygraph 262. The return on this approach for the added effort is a modest increase of 4% in the Outcome Confidence of a passed test - with the earlier model we had a 95% confidence and here we have a 99% confidence.

The question is: *Does the modest increase in Outcome Confidence justify the cost of all the extra polygraph testing?* That decision is best left to the consumer and should be assessed ahead of time after considering their testing goals and target base rates. In this model we would have conducted 738 polygraph examinations whereas in the earlier model we would have conducted 262. Would the extra 476 polygraph exams justify the 4% increase in Outcome Confidence? If a typical screening polygraph examination and report takes 3 hours, we would have spent an additional 1428 hours – the equivalent of one person working about 36 work weeks or nine months.

Consideration: What if someone failed the polygraph test after passing the EyeDetect test?

Recall from Table 9 that 738 people passed the EyeDetect test - 34 were actually Guilty and 704 were actually Innocent. About 127 of those 738 people (17%) will fail the polygraph test after passing the EyeDetect. Once again, this phenomenon is inevitable because all testing has error rates.

Of the 127 failed polygraph tests, 99 are Innocent and only 28 are actually Guilty - that is why that Outcome Confidence is so low (22%). Agency decision-makers should seriously consider what action(s) this low Outcome Confidence warrants, perhaps conduct a thorough background and disregard the second test result. Remember, the subject already passed the EyeDetect and that Outcome Confidence was 95%.

Take Home Points

1. All credibility assessment tests are imperfect, and we have evidence-based estimates of these errors. We can use these estimates to strategize our testing based on our goals.
2. Credibility assessment targets can have low base rates of guilt for disqualifying behavior.
    a. As you can see, base rates can have a profound effect on the Outcome Confidence in test results.
    b. Outcome Confidence is an under-appreciated phenomenon in credibility assessment testing.
3. When the base rate of Guilt is low, the test group consists of a majority of Innocent people. That means the Outcome Confidence for passing the test in the first hurdle (i.e., EyeDetect) is quite high at 95%. With that rate of accuracy, it may be cost-effective to move those test subjects to the next phase of the hiring process.
    a. But, it also means the Outcome Confidence in a failed test result is not as high (63%). See Case 5.
4. By using EyeDetect first, the base rate of Guilt for the polygraph test group was raised to 63%. This improved the Outcome Confidence from 63% to 91% for a failed polygraph test using the successive hurdles model. See Case 5.
5. There is always a cost versus benefit trade-off to consider. For example, although the Base Rate of Guilt increased from 20% to 63% in the second hurdle, there was a slight loss in Outcome Confidence for a passed test.
    a. The Outcome Confidence for a passed test drops from 95% to 84%. See Case 5.
6. The cost-benefit trade can best be appreciated in Case 6. Here our goal was to reduce false negative rates – and we achieved a modest increase towards that goal (4%).
    a. But it came at the considerable effort of nine months of testing.
    b. This approach may be more tenable with lower number testing populations.
7. There will always be disagreements between the two technologies. They are both imperfect.
    a. We estimated that in this 20% base rate setting, the test outcome of *Failing the EyeDetect test then Passing the polygraph test* will occur about 31% of the time.
    b. We estimated that in this 20% base rate setting, the test outcome of *Passing the EyeDetect test then Failing the polygraph test* will occur about 17% of the time.
    c. These phenomena are inevitable due primarily to the imbalance of the base rates of the target groups of test subjects.
    d. We would not want perfect correlations or it would negate the benefit of a second hurdle with a different technology.

References:

American Polygraph Association. (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, 40, 194-305.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214-234.

Kircher, J.C. & Raskin, D.C. (1987). The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data: comment: base rates and statistical precision. *Institute of Mathematical Statistics, Statistical Science*, Vol.2, No. 3, pp. 226-228.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.

Raskin, D.C. (1984, March). Proposed use of polygraphs in the department of defense. Statement before the Committee on Armed Services, U.S. Senate.

Raskin, D.C. (1987). Methodological issues in estimating polygraph accuracy in field applications. *Canadian Journal of Behavioral Science*, 19(4), 389-404.

Raskin, D. C. (2015, September). The Utah Technique, Presentation at the Presented at the American Polygraph Association 50th Annual Seminar, Chicago, IL.