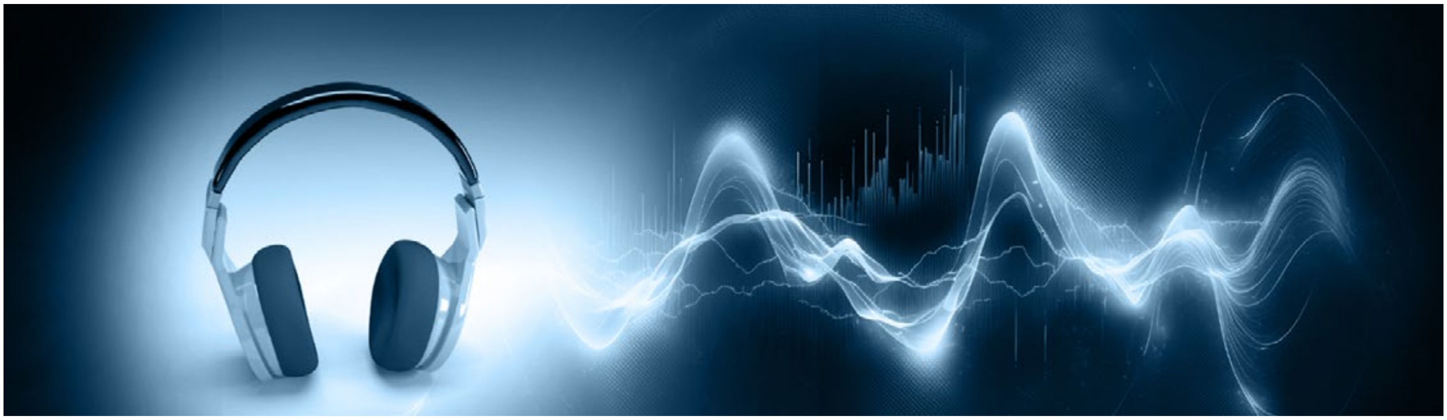

Practical Polygraph: How to Write Probability Information in Evidentiary Polygraph Reports (Standard 1.8.3)

Raymond Nelson

APA Standard 1.8.3 states that probabilistic information shall be provided along with categorical test results of evidentiary exams – defined in Section 1.1 of the APA Standards of Practice as those exams conducted with the intention that the results are to be used as a basis of information in a legal proceeding. Provision of statistical information has become a common standard in science, scientific testing, and forensics. The reason for this requirement is to educate consumers as to the inherently probabilistic nature of much scientific information, and to reduce the human tendencies to expect infallibility and to exaggerate potential effect sizes. The inclusion of probability information along with categorical results or professional opinions is a general standard in forensic science following Daubert which held that good science is characterized by reasonable descriptions of known or potential error rates. Probability information can be described in several ways.

Reference to published standards

The simplest way to provide probabilistic information in support of a categorical polygraph test result is to refer to the APA Standards of Practice that defines the accuracy effect size requirements for polygraph techniques selected for various purposes. Evidentiary polygraphs are those that are conducted with the goal or intention that the test result will is to be used as a basis of information in a legal proceeding. Paired testing involves the conduct of two polygraph exams on two opposing witnesses by two different examiners who are both blind to the other test outcome. Investigative polygraphs are those conducted without the intention to use the test results in a legal context, such as those polygraphs conducted in applicant screening, security screening, and post-conviction monitoring programs.

Section 1.1.7.3.1 of the APA Standards of Practice state that polygraph techniques used for evidentiary exams shall provide an overall accuracy rate of .90 or greater, with an inconclusive rate of .20 or lower. Section 1.1.7.3.2 states that polygraph techniques used for paired testing are required have an overall accuracy rate of .86 or greater with an inconclusive rate of .20 or lower (the lower accuracy rate is permitted because the accuracy effects two different results from examiners, both blind to the other result, can be combined as independent probability events, with a resulting effect size that outperforms either test alone when the two blind concur that one witness was deceptive and the other was truthful). Accuracy requirements for investigative polygraph techniques are described in Section 1.1.7.3.3 of the APA Standards of Practice. These are required to provide an overall accuracy rate of .80 or greater with an inconclusive rate of .20 or lower. The following is an example of how to report compliance with Standard 1.8.3 for evidentiary polygraphs. The text can be modified for tests conducted for pair-testing and investigative purposes.

The test data were recorded using a test format and analysis method that conform to the requirement of the APA Standards of Practice Section 1.1.7.3.1, including an overall accuracy rate of 90% or greater.

When test accuracy information is expressed as a function of frequentist accuracy metrics observed in published studies, it does not describe the strength of information for any individual test result. Instead, the strength of any individual examination result is estimated as a function of the Standards of Practice, premised on the conduct of the examination using methods that are known to be consistent with the standards.

Frequentist accuracy and error rates for the test format

Another way to provide probability information for an evidentiary examination is to describe the frequentist test accuracy metrics. This includes simple accuracy metrics such as true-positive, true-negative, false-positive, and false-negative rates, and can also include aggregated metrics positive-predictive-value and negative-predictive-value, and related metrics such as the false-positive-index and false-negative-index. Description of the total percent correct and percent incorrect is another example. These metrics can be studied for comparison question techniques as a whole, or for unique formulations of the comparison question technique. The following is an example of how to report the for accuracy metrics for a four-question evidentiary test format. This text can be modified for other test formats.

The test data were recorded using a 4-Question single-issue polygraph technique (Raskin Technique). When analyzed using the ESS-M data analysis method, effect sizes for this technique have been summarized in an appendix (APA Editorial Staff, 2020) to the 2011 meta-analytic survey of validated polygraph techniques (APA, 2011). Overall accuracy for this method was reported in a reported at .944 (95% CI .897 to .987) with an inconclusive rate of .031 (.01 to .092). Test sensitivity was .923, and specificity *was .908. The reported false positive error rate was .046 and the false negative rate was .062.*

Importantly, these statistics should not be reported from a single study, but should be aggregated from a corpus of available research data as is commonly done in research literature reviews and meta-analytic research. Another important consideration is that research may be accomplished with balanced study groups, or with imbalance groups, and aggregated frequentist statistics may not be readily applicable to field settings in which the incidence rate or prior probability is imbalanced. A third important consideration is that individual professionals generally do not outperform the limits or capabilities of the underlying science – despite the fact that it is sometimes socially gratifying to make claims of personal accuracy rates that are extremely high. (This is similar to the way that most investors do not, over time, outperform the financial markets.)

When the exact causes of testing errors are not known, testing errors are regarded as analogous to random measurement errors. Of course, when the exact cause of an error is known it can be corrected or avoided, and test accuracy is thus increased. But it is important to always remember that scientific tests, of all types, are not expected to be infallible. This is because the purpose of any scientific test is to quantify a phenomena of interest that cannot be subject to physical measurement (still subject to random measurement error) or deterministic perfection (which would be immune to both random measurement error and human behavior). Scientific tests make use the statistical relationship between available proxy data and the phenomena of interest. Because test data and test results are not direct physical measurements, they are only expected to quantify the margin of uncertainty that surrounds a conclusion or test result.

Frequentist test accuracy metrics characterize all testing errors as random events – assuming that the tests are conducted correctly, and that errors are not attributable to procedural error. The likelihood of an error for any individual case is estimated as a function of the random error rate. In a complimentary manner, the likelihood of a correct test result for any individual case is also estimated as a function of the known effect sizes for a polygraph technique. A limitation of this method is that some frequentist accuracy metrics of interest may be non-resistant to influence from imbalanced priors. A more practical limitation of this method is that it does not provide intuitively useful information about the strengths or limitations of the data for any individual case.

Statistical classifiers

A statistical classifier is a reproducible statistical value for an individual polygraph test, and is used to make a categorical classification of the test result based on the strength of the statistical value. A number of different types of statistical classifiers are possible. A common simple classifier is the p-value, which can be calculated using either empirical or theoretical reference distributions. More complex statistical classifiers are also commonly used, including logistic functions, naive Bayes classifiers, support vector machines, k-nearest neighbor methods, and others. The following is an example of how to report a statistical classifier (p-value) for an evidentiary polygraph test result. This text can be modified for other test formats and other test outcomes.

Using the ESS, an evidence-based, norm-referenced, and standardized protocol for test data analysis, the grand total score of -5 equals or exceeds the required cutscore of -4 for deceptive classifications. The level of statistical significance, is calculated at p = .032, which is equal to or less than the required alpha boundary (a = .05), and indicates that only a small proportion of truthful persons (3.2%) can be expected to produce an equal or lower test score. These results support the conclusion that there is deception indicated by the physiological responses to the test stimulus questions during this examination.

Statistical classifiers are mathematical abstractions that provide an objective basis for decision-making at the individual case level. However, they do not, by themselves, describe the practical likelihood correct or incorrect classification at the level of an individual case or a group of cases. They do not attempt to provide information that can be easily or intuitively descriptive of the practical strength of information at the individual case level. However, statistical classifiers can be combined with other methods, especially Bayes Theorem, to provide posterior or outcome information that has greater practical and intuitive meaning.

Posterior or outcome probabilities

Posterior conditional probabilities, also known as posterior probabilities or posterior distributions, and outcome confidence levels are related concepts in probability and statistics. They are probabilities assigned to specific events or outcomes after observing and analyzing the test data. Posterior conditional probabilities provide a statistical basis for classification while also giving information that describes the practical strength of the information for each case. In the context of Bayesian Analysis, posterior probabilities are calculated using Bayes' theorem, which relates the conditional probability of an event A given evidence B to the prior probability of A and the likelihood of B given A. Importantly, both the construction and conduct of the polygraph test (including the, interview, sequence of test question, instrumentation, and sensors) and the analysis of the test data are premised on the basic theory of the polygraph test – that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception and truth-telling in response to relevant target stimuli.

Posterior probabilities are essential in Bayesian statistics, because they allow us to use new evidence or data to make objective and mathematical updates to the probabilities associated with different possible outcomes. Bayesian posterior probabilities are often expressed in odds form, and lead easily to the calculation of Bayes Factor – which describes, in odds form, the change or increase in the probabilistic strength of information in support of a particular classification or conclusion. Following is an example of how to report the posterior information in support of a categorical conclusion for an evidentiary polygraph exam. The text can be modified for other test outcomes.

Using the ESS-M, , an evidence-based, norm-referenced, and standardized protocol for test data analysis, the grand total score of 19 equaled or exceeded the required numerical cut-score (3). The posterior odds of truth-telling was 87 to 1, for which the posterior probability was .98. The lower limit of the 1-alpha Bayesian credible interval was 15 to 1, which exceeded the prior odds (1 to 1). This indicates a likelihood of greater than 95% that the posterior odds of truth exceed the prior odds. The posterior information for this examination was increased by a Bayes Factor of 87 times. These analytic results support the conclusion that there were no significant reactions indicative of deception in the loading of recorded changes in physiological activity in response to the relevant test stimuli during this examination.

Outcome confidence levels are another way to express the degree of certainty or confidence associated with a particular prediction or decision made by a probabilistic model, such as Bayesian network or other statistical classifier. These confidence levels are often expressed as decimal proportions between 0 and 1, or as a percentage, with higher values indicating greater confidence in the outcome or prediction. Outcome confidence levels are typically derived from posterior conditional probabilities, and can be useful for decision making and risk assessment.

In practice, both posterior conditional probabilities and outcome confidence levels are valuable for their ability to provide practical and intuitively useful descriptions of the strength of information in support of predictions and classifications made by probabilistic models. They can be used to set decision thresholds (e.g., classifying data points based on a minimum confidence level), and to provide additional information to users about the level of certainty that can be reasonably assigned to a classification or conclusion.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40(4)*, 196-305.

American Polygraph Association. (Aug. 25, 2023). *APA Standards of Practice*. Retrieved from https://www.polygraph.org on Nov. 7, 2023.

APA Editorial Staff. (2020). Addendum to the 2011 meta-analytic survey – the Utah four-question test ("Raskin technique") / ESS. *Polygraph & Forensic Credibility Assessment, 49(2)*, 73-78.

Daubert v. Merrel Dow Pharmaceuticals, 509 U.S. 579 (1993).

Honts C. R., Thurber, S., and Handler, M. (2021) A comprehensive meta analysis of the comparison question polygraph test. *Applied Cognitive Psychology 35(2)*, 411–427.

National Research Council. (2003). *The Polygraph and Lie Detection*. Washington, DC: The National Academies Press.

National Research Council. (2011). *Reference Manual on Scientific Evidence*. Washington, DC: The National Academies Press.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine 2016,* 49(5), 69-82.

Nelson, R. (2017). Multinomial reference distributions for the empirical scoring system. *Polygraph & Forensic Credibility Assessment, 46(2)*, 81-115.

Nelson, R. (2018). Guide for how to use the ESS-Multinomial reference tables in four steps. *APA Magazine, 51(2)*, 78-89.

Nelson, R. (2018). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. *Polygraph & Forensic Credibility Assessment, 47(2)*, 158-175.

Nelson, R. (2018). Five minute science lesson: Bayes' theorem and bayesian analysis. *APA Magazine 2018, 51 (5)*, 65-78.

Nelson, R. (2018). Practical polygraph – a tutorial (with graphics) on posterior results and credible intervals using the ESS-M Bayesian classifier. *APA Magazine, 51(4)*, 66-87.

Nelson, R. (2019). Explainer for the ESS-M narrative summary. *APA Magazine 2019, 52(6)*, 30-48.

Nelson, R. (2022). Multinomial reference distributions for seven-position Likert scores of comparison question polygraph tests. *Polygraph & Forensic Credibility Assessment, 51(1)*, 41-67.

Nelson, R. & Handler, M. (2012). Using normative reference data with diagnostic exams and the Empirical Scoring System. *APA Magazine, 45(3)*, 61-69.

Nelson, R. & Handler, M. (2015). Statistical reference distributions for comparison question polygraphs. *Polygraph, 44(1)*, 91-114.

Nelson, R. & Rider, J. (2018). Practical polygraph: ESS-M made simple. *APA Magazine, 51(6)*, 55-62.