Statistical Reference Distributions for Comparison Question Polygraphs

Raymond Nelson and Mark Handler

Abstract

Statistical reference tables are shown for the distributions of numerical scores of diagnostic and screening polygraphs. Discussion is provided regarding the importance of statistical reference data to help understand the polygraph test results as probabilistic information as opposed to deterministic observation or physical/mechanical measurement of deception.

Keywords: *reference tables, normative data, validated polygraph techniques, test data analysis, evidence-based polygraph*

The American Polygraph Association (2011) published the results of a meta-analytic survey of contemporary polygraph test formats and polygraph scoring methods that are supported by published descriptions of test administration and test data analysis procedures[1]. Information in that report included statistical confidence intervals for expected precision and error rates for diagnostic and screening exams. It also included information on test sensitivity, specificity, false-positive and false-negative error rates, positive and negative predictive values, inconclusive rates and the unweighted average accuracy excluding inconclusive results. Means and standard deviations of the distributions of scores were included in an appendix. Since that time there has been increased appreciation for the importance and value of evidence-based practices and validated polygraph techniques. There is also raised awareness of the potential problems that can accompany the use of un-validated or experimental methods in field settings.

Information in the appendix of the 2011 report could be used to calculate the level of significance for an individual test result. However, many professionals prefer not to do statistical calculations in field settings. They rely instead either on computer programs to compute any required statistical and mathematical calculations, or on published statistical reference tables that include calculated statistical results for the complete range of possible test scores.

Computers can offer the convenience of rapid calculations and automated professional work-flow, such as the construction of reports. Reference tables offer a simple solution for manual test data analysis and can facilitate a simple and intuitive understanding of the probabilistic meaning of an individual test score, relative to other possible test scores. Whether calculated manually, by computer algorithm, or through published reference tables, statistical reference data are an important aspect of the basis of validity for any scientific test.

Reference tables included in the present publication pertain to techniques in the 2011 report on validated polygraph techniques.

[1] The meta-analytic survey is not intended to take the form of an official list or field practice policy regarding scientific questions about polygraph validity.

Authors Note:
The view and opinions expressed in this publication are those of the authors and do not necessarily represent those of the American Polygraph Association, or any other agency or business entity with whom they are affiliated. Address correspondence to raymond.nelson@gmail.com or polygraphmark@gmail.com.

They include 7-position, 3-position and ESS scores for both diagnostic and screening test formats. The practical and scientific purpose of statistical reference data is to answer questions about the level of significance or margin of uncertainty surrounding a test result. This information is a potentially important resource for researchers, policy makers, program managers, and field practitioners within the polygraph profession.

Testing: classification and inference

Test results can be thought of as addressing two concern: classification and inference. Classification refers to the simple categorical results that are useful to most people. In the most general form, categorical results are described as either positive or negative, and these refer to whether or not the phenomenon of interest is determined to be present or absent. In the polygraph context, classification refers to whether test results support a conclusion of deception or truth-telling. Inference on the other hand, refers to the ability to quantify the margin of uncertainty surrounding a test result. Some testing contexts may be unconcerned about making a categorical conclusion, while others may not be concerned about probabilistic error estimation.

In the polygraph profession both categorical and inferential test results are important and useful. Categorical results provide a convenient mechanism to understand and respond to different test outcomes in consistent and structured ways that avoid the temptation for biased or arbitrary decisions. At the same time, there is an inherent need to remain aware of, and account for, the potential for testing error. Neglecting our obligation to quantify the margin of uncertainty and potential for testing error invites problems in the form of misguided and frustrated expectations for perfect test accuracy, or charlatanry in the form of claims of perfect or near-perfect testing accuracy.

Probabilistic vs deterministic

It has been well-accepted for several decades that there is no unique physiological lie response (Lykken, 1959; Kircher & Raskin, 1982; Raskin, Honts & Kircher, 2014). In practical terms this means that it is not possible to achieve either perfect deterministic observation or physical/mechanical measurement of deception – Pinocchio's growing nose does not exist. A consequence of this is that all test results, including the scientific detection of deception, will be a statistical and probabilistic endeavor regarding an amorphous construct –deception or truth-telling. Tests attempt to evaluate amorphous phenomena that cannot be observed or mechanically measured through the identification and measurement of proxy features correlated with the phenomena of interest. We can then combine these features to optimize the effectiveness of the test result. Obviously, a test is not needed when a deterministic observation or physical measurement is possible.

Because all tests are imperfect, all test results can be thought of as non-deterministic and therefore probabilistic. All test results are probability statements, including when simplified to categorical results[2]. A consequence of this is that the use of the polygraph, or any imperfect test, will be accompanied by a need for administrative and professional practice policies that are properly informed about the test capabilities. This includes test sensitivity, specificity, false-positive, and false-negative error rates.

As test results are sometimes used as a basis of evidence to add incremental validity to decisions that affect individual persons, field practitioners and referring professionals are obliged to become prepared to discuss inferential and probabilistic information about the effectiveness of a test method in general, and also the probabilistic meaning of each individual test result.

[2] Tests are used and needed only when there is a need or desire to measure or evaluate an amorphous phenomenon that cannot be subjected to deterministic solution or physical/mechanical measurement.

Statistical reference distributions, such as those included herein, can provide a practical resource and tabular illustration to help answer important questions and explain concepts regarding our present knowledge and evidence pertaining to polygraph test accuracy.

Polygraph test accuracy

Nelson and Handler (2013) described the history of scientific reviews of studies of polygraph accuracy since 1973 and showed that test accuracy, though overestimated in the past, appears to have converged over time among studies from both within and outside the polygraph profession (American Polygraph Association, 2011; Honts & Peterson, 1997; National Research Council, 2003; Office of Technology Assessment, 1983). Attempts to portray polygraph as capable of near perfect accuracy are in opposition to both the weight of the evidence and sound scientific theory. There is sufficient evidence today, however, to support the conclusion that the polygraph is a valid scientific test with accuracy that is significantly greater than chance.

Discussions about polygraph test accuracy in general, however, may be unsatisfying when discussing the results of a single examination. To more fully appreciate the meaning of a single test result it is useful to compare the numerical and probabilistic test result to what is normally expected from members of a particular group.

Normative data

From a statistical perspective, normative data are the mean, standard deviation and distribution shape of the numerical scores of guilty and innocent persons. Statistical means, or averages, provide information about the similarity of the scores among the members of the group. Standard deviations are a statistical measure of dispersion, and provide information about how the scores differ among members of a group. These statistics can be used to calculate the expected proportion of a group relative to the individual's test score, and this can be used to make probabilistic inferences about the probability that a conclusion or hypothesis is incorrect.

In the most rigorous sense, normative data will be based on information from multiple large-scale randomly selected samples. This way the different sampling results can be expected to converge to the testing population. This is the common method of estimating population characteristics that cannot themselves be subject to deterministic observation or exact measurement.

In a more general sense, normative data are any source of reference knowledge or information that are considered representative of a group. Normative data may then be used as a baseline against which subsequently collected data can be compared. Normative reference data can help to establish program and field practice rules that seek to constrain testing errors to within margins of uncertainty based on the program objectives. From a practical perspective, statistical reference data can help field polygraph examiners and program managers to achieve desired probabilistic test accuracy and error rates by using the optimal cutscores selected a priori.

Statistical information about the reference populations of scores with different polygraph testing techniques helps make better decisions at the level of the individual test and better field practice policies at the level of the test in general. There are a growing number of polygraph professionals, referring professionals and policy makers who appreciate and make use of this important information. A complete description of the operational procedures for using normative reference data is the subject of other publications (Nelson & Handler, 2012), but these procedures can be summarized in four steps: 1) locate the normative reference data for the testing technique, 2) determine the alpha boundaries and numerical cutscores, 3) conduct and score the test and calculate the test error statistic, and 4) interpret the test result (i.e., translate the numerical and statistical test result into categorical test results that can be described in human language useful to referring professionals).

Numerical and probability cutscores

An important practical use for reference tables is to determine numerical cutscores that correspond to desired probability cutscores. These are often expressed as an alpha level that indicates a tolerance for error in the context of a non-deterministic, probabilistic test result. Another practical use for reference tables is to determine the level of significance for an individual test result. Alpha boundaries are commonly set at .05, indicative of a 5% tolerance for error. Nothing prevents the use of a more restrictive alpha boundary of .01 in circumstances in which a very low error rate is desired, with the understanding that more test results will be classified as not statistically significant (i.e., inconclusive or no-opinion). Similarly, nothing prevents the use of a less restrictive alpha boundary of .10, indicative of a 10% tolerance for error, under circumstances that may benefit more from a reduce rate of inconclusive results. In actuality, nothing prevents a decision boundary at .25 or any arbitrary probability, though this is not common practice.

In practice, the selection of an alpha boundary is an administrative decision that is not generally decided by field practitioners. Instead, field practitioners should become familiar with the simple procedure for selecting a numerical cut-score from a reference table with regard for the required alpha level indicative of the desired accuracy rate and tolerance for error. Field practitioners should also become familiar with the use of statistical reference tables to determine the level of significance or probability of error associated with an observed test score.

Evaluating test accuracy

Accuracy the polygraph is evaluated at both the level of the test technique itself, in terms of the generally expected precision and error of the technique, and at the level of the individual examination. Evaluation of individual examinations may include general quality assurance activities, concerned with procedural compliance with field practice standards, and can also involve quantification of the margin of uncertainty surrounding the observed test result (i.e., what is the level of significance or probability of error associated with an individual examination result).

To be useful in practical ways, it is important probabilistic estimations have some intuitive or understandable connection to the test data, and to the psychological and physiological mechanisms that play a role in the recorded test data. Intuitive, in this sense, means that human experts can understand and explain not only the simple categorical classification, but also the broader contextual meaning regarding how a test result or decision was achieved. This includes the data or evidence on which it is based, and how the information is obtained, quantified, transformed, reduced and interpreted.

The alternative to an intuitive understanding would be a black-box model, for which we are given a result and expected to accept it without an opportunity to scrutinize or understand the supporting information. It is ethically preferable to use models that offer intuitive understanding, and avoid black-box models, whenever we are engaging in human decision making. Reference tables, while they do not offer information about the psychological and physiological foundations, provide a visual and intuitive understanding of the meaning of different possible test scores within the distributions of possible scores.

Reference tables

Reference Tables are shown in Appendices A-P. In formulating these tables we have forgone the use of named techniques, a feature of the meta-analytic survey that was met with criticism because of the tendency for named techniques to foster more confusion than accurate understanding. Instead we have provided reference tables associated with two important fundamental issues: 1) whether a technique is used for diagnostic testing[3]

[3] Diagnostic test are those which there is a known allegation, known incident or known problem, and for which the test stimulus questions are all associated and for which response variance is therefore non-independent.

or for screening purposes[4]; and 2) the number of relevant stimulus questions

The number of questions has a direct effect on the total number of stimulus-response iterations and scored segments of data, and will also introduce well-known multiplicity effects that cause unintended distortion of desired alpha-decision boundaries and error rates[5].

Most manual scores are nonparametric in that they do not rely on linear measurements and do not rely on linear assumptions about recorded physiological responses. Even so, numerical scores have been found to conform to linear assumptions and the normal distribution shapes of Gaussian models. A Gaussian-Gaussian signal discrimination model can be described with regard to most presently available manual scoring protocols for comparison question polygraph tests (Barland, 1985; Krapohl & McMannus, 1999; Nelson, Krapohl, & Handler, 2008; Wickens, 2002).

Reference distributions for deceptive test scores are calculated as the cumulative distribution function for the mean and standard deviation of truthful scores. Reference distributions for truthful scores are calculated as the inverse cumulative distribution function for the deceptive mean and standard deviation. In this way, discussions of statistical significance can take the familiar form of lower-tail p-values for both deceptive and truthful classifications. These reference distributions provide information on seven-position, three-position and Empirical Scoring System scores for both diagnostic and screening exams with two, three, and four relevant questions.

Diagnostic and error variance

Polygraph data – like virtually all data in science and testing – is a combination of diagnostic variance (i.e., response data that is explained by the test or experimental stimulus condition) and error variance (i.e., data that is not explained by the test or experimental stimuli). It is sometimes referred to more casually as signal and noise. Error variance can be further decomposed to reducible error variance and irreducible error variance.

Reducible sources of error variance can include undiscovered or unmeasured diagnostic features that are not included in the test model. Reducible sources of error, if they can be identified, are a potential opportunity to improve the accuracy of a test. However, even if it were possible to identify, extract, and quantify all available diagnostic variance from the test data, the test result would still have some margin of uncertainty due to irreducible sources of error. This is because by definition any non-deterministic model will have some non-zero quantity of irreducible error variance.

An ideal test, a deterministic test, would produce the same numerical, mathematical, and categorical result every time the test is repeated for the same subject, topic and other testing conditions. Because the polygraph is not a deterministic test, results are inherently probabilistic. This leaves both test developers and field practitioners with the obligation to accurately inform referring professionals about the empirical basis and probabilistic meaning of the test results.

[4] Screening tests are those for which there is no known allegation or incident, and for which multiple-issue exams can provide the scientific advantage of increased test sensitivity to a wider range of important target issues. In field practice, screening exams are sometimes conducted as single-issue exams.

[5] Multiplicity effects in data analysis – including the inflation of alpha and false-positive error rates when making deceptive classifications using multiple subtotal scores of diagnostic exams or deflation of alpha and increased inconclusive rates when making truthful classification of multiple-issue screening exams – can be reduced using common statistical corrections.

It is important to recognize that the complete range of possible factors that contribute to irreducible error and unexplained variance cannot be defined and enumerated. If it were possible to explain and account for (i.e., quantify, either deterministically, mechanically, or probabilistically) all sources of error variance then it would also be feasible to control those factors and there would no longer be any sources of error variance. Attempts to portray polygraph test results with absolute confidence – as incapable of error – are not consistent with reality. Statistical reference data can help to increase understanding of the probabilistic nature of polygraph test results

Conclusion

Publication of these reference tables is an important step for the polygraph profession. A probabilistic view of polygraph does not depend explicitly on any particular theory or hypothesis about the psychological and physiological mechanisms that underlie responses to polygraph stimuli. A probabilistic view does reject any particular construct, and therefore is not inherently atheoretical. Statistical reference data merely describe what we have observed, and can expect to observe, under similar testing circumstances. This does not preclude or negate the value and importance of the development of continued knowledge regarding the bases and causes of psychological and physiological mechanisms-the construct validity of the test.

These reference data pertain to particular testing procedures involving the type of test (i.e., diagnostic or screening) and the number of questions (which may introduce multiplicity effect and the need for statistical corrections, depending on the choice of decision rules). This means that procedural compliance with published procedures and established standards of practice will remain an important concern. Changes in, or additions to, sensor technology, signal processing methodology, testing procedures, and even feature extraction (i.e., score assignment) can be expected to bias these observed distributions such that inference may be untenable and the reasonable prediction of classification error and precision rates might become difficult.

It will important to verify the performance characteristics of proposed changes or improvements before attempting to apply them to presently available reference data. Evaluation of the merits and effectiveness of proposed changes will require the availability of information describing the procedures for both test administration (for quality assurance and standards compliance purposes) and also for test data analysis – including statistical information about the precision of test results. Absence of information about both the expected precision and error rates, and the expected distributions of test scores will limit the ability to properly evaluate new developments.

Among the greatest hazards will be that an absence of information about the expected precision and error of test results will foster a condition in which professional expertise becomes a form of inscrutable esoteric knowledge subject to confirmation bias. Published standards and procedures, along with published statistical reference information, can help the polygraph profession to advance its effectiveness and usefulness.

As always, there is a need for continued research and more information in this and other areas. Although the scope of the 2011 meta-analytic survey was large, the included reference tables are still based on a relatively small amount of available data. Additional studies are needed to further understand the limits and utility of these distributions. Research is also needed to advance our understanding of the details psychological and physiological mechanisms that affect responses to polygraph questions. It is our hope that polygraph professionals and scientists will be able to appreciate and make use of the information contained in these reference tables.

References

Abdi, H. (2007) Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.) *Encyclopedia of Measurement and Statistics,* Thousand Oaks, CA, Sage.

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40(4),* 196-305. [Electronic version] Retrieved August 20, 2012, from http://www.polygraph.org/section/research-standards-apa-publications.

Barland, G.H. (1985) A method of estimating the accuracy of individual control question polygraph tests. In: Anti-terrorism; forensic science; psychology in police investigations: *Proceedings of IDENTA-'85*, p 142-147. The International Congress on Techniques for Criminal Identification.

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

Honts, C. R. & Peterson, C.F. (1997). *Brief of the committee of concerned social scientists as amicus curiae United States v Scheffer.* Available from the author.

Horvath, F. & Palmatier, J. (2008). Effect of two types of control questions and two question formats on the outcomes of polygraph examinations. *Journal of Forensic Sciences, 53* (4), 1-11.

Kircher, J. C. & Raskin, D. C. (1982). Is there a "specific lie pattern" of autonomic responses? *Psychophysiology, 19, 569.* (Abstract)

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

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43,* 385-388.

Matte, J. A. & Reuss, R.M. (1989). A field validation study of the quadri-zone comparison technique. *Polygraph, 18*, 187-202.

National Research Council (2003). *The Polygraph and Lie Detection.* National Academy of Sciences.

Nelson, R. (2012). Monte Carlo study of criterion validity of Backster You-Phase. *Polygraph. 41* (1), 44-53

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., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph, 40,* 67-78.

Nelson, R. & Handler, M. (2013). A brief history of scientific reviews of polygraph accuracy research. *APA Magazine, 46* (6), 22-28.

Nelson, R., Krapohl, D. & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph, 37*, 185-215.

Office of Technology Assessment (1983). *The validity of polygraph testing: A research review and evaluation.* Printed in *Polygraph, 12*, 198-319.

Raskin, D.C., Honts, C.R., Nelson, R., and Handler, M. (2015). Monte Carlo estimates of the validity of four relevant question polygraph examinations. *Polygraph 44* (1), 1-27.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (Eds.). (2014). Credibility assessment: Scientific research and applications. Oxford: Academic Press.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford.

Appendix A

Two-question Event-specific Exams / Backster 7-position Scoring Method

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-12	8	6	10

Deceptive scores		Truthful scores	
Score	p-value	Score	p-value
0	.274	1	.052
-1	.242	2	.040
-2	.212	3	.030
-3	.184	4	.023
-4	.159	5	.017
-5	.136	6	.012
-6	.115	7	.009
-7	.097	8	.006
-8	.081	9	.004
-9	.067	10	.003
-10	.055	11	.002
-11	.045	12	.001
-12	.036	13	.001
-13	.029	14	.001
-14	.023	15	<.001
-15	.018		
-16	.014		
-17	.011		
-18	.008		
-19	.006		
-20	.005		
-21	.004		
-22	.003		
-23	.002		
-24	.001		
-25	.001		
-26	.001		
-27	.001		
-28	<.001		

Means and standard deviations are from Nelson (2012)

Appendix B

Two-question Event-specific Exams / Empirical Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-6	6	6	6

Deceptive scores		Truthful scores	
Score	p-value	Score	p-value
0	.159	1	.122
-1	.122	2	.091
-2	.091	3	.067
-3	.067	4	.048
-4	.048	5	.033
-5	.033	6	.023
-6	.023	7	.015
-7	.015	8	.010
-8	.010	9	.006
-9	.006	10	.004
-10	.004	11	.002
-11	.002	12	.001
-12	.001	13	<.001
-13	<.001		

Means and standard deviations are truncated integers as reported previously in Nelson *et al.*, (2011).

Appendix C

Three-question Event-specific Exams / Empirical Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-9	8	8	7

Deceptive scores		Truthful scores	
Score	p-value	Score	p-value
0	.127	1	.106
-1	.099	2	.085
-2	.077	3	.067
-3	.058	4	.052
-4	.043	5	.040
-5	.032	6	.030
-6	.023	7	.023
-7	.016	8	.017
-8	.011	9	.012
-9	.008	10	.008
-10	.005	11	.006
-11	.003	12	.004
-12	.002	13	.003
-13	.001	14	.002
-14	<.001	15	.001
		16	<.001

Means and standard deviations are truncated integers as reported previously in Nelson *et al.*, (2011).

Appendix D

Multiple-issue Exams / Empirical Scoring System

Sub-total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-2	3	2	3

Deceptive Scores		Truthful Scores				
Score	p-value	Score	p-value	2 RQs	3 RQs	4 RQs
0	0.252	1	.159	0.083	0.056	0.042
-1	0.159	2	.091	0.047	0.031	0.024
-2	0.091	3	.048	0.024	0.016	0.012
-3	0.048	4	.023	0.011	0.008	0.006
-4	0.023	5	.010	0.005	0.003	0.002
-5	0.010	6	.004	0.002	0.001	0.001
-6	0.004	7	.001	0.001	<.001	<.001
-7	0.001	8	<.001	<.001		
-8	<.001					

P-values for truthful classifications of multiple issue exams are statistically corrected using the Šidák correction for the number of relevant questions.

Means and standard deviations are truncated integers as reported previously in Nelson *et al.*, (2011).

Appendix E

Two-question Event-specific Exams / Federal 7-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-7	5	5	5

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.159	1	.055
-1	.115	2	.036
-2	.081	3	.023
-3	.055	4	.014
-4	.036	5	.008
-5	.023	6	.005
-6	.014	7	.003
-7	.008	8	.001
-8	.005	9	.001
-9	.003	10	<.001
-10	.001		
-11	.001		
-12	<.001		

Normative parameters are from combined studies using Federal 7-position scores, as reported in American Polygraph Association (2011).

Appendix F

Three-question Event-specific Exams / Federal 7-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-8	9	7	8

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.191	1	.159
-1	.159	2	.133
-2	.130	3	.111
-3	.106	4	.091
-4	.085	5	.074
-5	.067	6	.060
-6	.052	7	.048
-7	.040	8	.038
-8	.030	9	.030
-9	.023	10	.023
-10	.017	11	.017
-11	.012	12	.013
-12	.009	13	.010
-13	.006	14	.007
-14	.004	15	.005
-15	.003	16	.004
-16	.002	17	.003
-17	.001	18	.002
-18	.001	19	.001
-19	.001	20	.001
-20	<.001	21	.001
		22	<.001

Means and standard deviations are from combined studies using Federal 7-position scores, as reported in American Polygraph Association (2011).

Appendix G

Multiple Issue Exams / Federal 7-position Scoring System

Sub-total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-2	4	3	3

Deceptive Scores		Truthful Scores				
Score	p-value	Score	p-value	2 RQs	3 RQs	4 RQs
0	.159	1	.227	.121	.082	.062
-1	.091	2	.159	.083	.056	.042
-2	.048	3	.106	.054	.037	.028
-3	.023	4	.067	.034	.023	.017
-4	.010	5	.040	.020	.014	.010
-5	.004	6	.023	.011	.008	.006
-6	.001	7	.012	.006	.004	.003
-7	<.001	8	.006	.003	.002	.002
		9	.003	.002	.001	.001
		10	.001	.001	.001	<.001
		11	.001	<.001	<.001	
		12	<.001			

P-values for truthful classifications of multiple issue exams are statistically corrected using the Šidák correction for the number of relevant questions.

Means and standard deviations are from combined studies using Federal 7-position scores, as reported in American Polygraph Association (2011).

Appendix H

Two-question Event-specific Exams / Federal 3-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-5	3	3	4

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.227	1	.023
-1	.159	2	.010
-2	.106	3	.004
-3	.067	4	.001
-4	.040	5	<.001
-5	.023		
-6	.012		
-7	.006		
-8	.003		
-9	.001		
-10	.001		
-11	<.001		

Means and standard deviations are from combined studies using Federal 3-position scores, as reported in American Polygraph Association (2011).

Appendix I

Three-question Event-specific Exams / Federal 3-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-7	5	5	5

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.159	1	.055
-1	.115	2	.036
-2	.081	3	.023
-3	.055	4	.014
-4	.036	5	.008
-5	.023	6	.005
-6	.014	7	.003
-7	.008	8	.001
-8	.005	9	.001
-9	.003	10	<.001
-10	.001		
-11	.001		
-12	<.001		

Means and standard deviations are from combined studies using Federal 3-position scores, as reported in American Polygraph Association (2011).

Appendix J

Multiple Issue Exams / Federal 3-position Scoring System

Sub-total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-1	2	2	2

Deceptive Scores		Truthful Scores				
Score	p-value	Score	p-value	2 RQs	3 RQs	4 RQs
0	.159	1	.159	.083	.056	.042
-1	.067	2	.067	.034	.023	.017
-2	.023	3	.023	.011	.008	.006
-3	.006	4	.006	.003	.002	.002
-4	.001	5	.001	.001	.001	<.001
-5	<.001	6	<.001	<.001	<.001	

P-values for truthful classifications of multiple issue exams are statistically corrected using the Šidák correction for the number of relevant questions.

Means and standard deviations are from combined studies using Federal 3-position scores, as reported in American Polygraph Association (2011).

Appendix K

Three-question Event-specific Exams – Utah 7-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-10	7	9	8

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.130	1	.058
-1	.106	2	.043
-2	.085	3	.032
-3	.067	4	.023
-4	.052	5	.016
-5	.040	6	.011
-6	.030	7	.008
-7	.023	8	.005
-8	.017	9	.003
-9	.012	10	.002
-10	.009	11	.001
-11	.006	12	.001
-12	.004	13	.001
-13	.003	14	<.001
-14	.002	15	
-15	.001	16	
-16	.001		
-17	.001		
-18	<.001		

Means and standard deviations are from combined studies using Utah scores, as reported in American Polygraph Association (2011).

Appendix L

Four-question Event-specific Exams – Utah 7-position Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-11	9	13	10

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.097	1	.091
-1	.081	2	.074
-2	.067	3	.060
-3	.055	4	.048
-4	.045	5	.038
-5	.036	6	.029
-6	.029	7	.023
-7	.023	8	.017
-8	.018	9	.013
-9	.014	10	.010
-10	.011	11	.007
-11	.008	12	.005
-12	.006	13	.004
-13	.005	14	.003
-14	.003	15	.002
-15	.003	16	.001
-16	.002	17	.001
-17	.001	18	.001
-18	.001	19	<.001
-19	.001		
-20	<.001		

Means and standard deviations are as reported in Raskin, Honts, Nelson and Handler (2015).

Appendix M

Four-question Event-specific Exams – Empirical Scoring System

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-12	10	11	9

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.111	1	.097
-1	.091	2	.081
-2	.074	3	.067
-3	.060	4	.055
-4	.048	5	.045
-5	.038	6	.036
-6	.029	7	.029
-7	.023	8	.023
-8	.017	9	.018
-9	.013	10	.014
-10	.010	11	.011
-11	.007	12	.008
-12	.005	13	.006
-13	.004	14	.005
-14	.003	15	.003
-15	.002	16	.003
-16	.001	17	.002
-17	.001	18	.001
-18	.001	19	.001
-19	<.001	20	.001
		21	<.001

Means and standard deviations were reported in Raskin, Honts, Nelson and Handler (2015).

Appendix N

MSU-MGQT (5 Question[6]) – 7-position scores

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-12	17	11	12

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
0	.180	1	.222
-1	.159	2	.205
-2	.139	3	.189
-3	.122	4	.173
-4	.106	5	.159
-5	.091	6	.145
-6	.078	7	.132
-7	.067	8	.120
-8	.057	9	.108
-9	.048	10	.098
-10	.040	11	.088
-11	.033	12	.079
-12	.028	13	.071
-13	.023	14	.063
-14	.019	15	.056
-15	.015	16	.050
-16	.012	17	.044
-17	.010	18	.039
-18	.008	19	.034
-19	.006	20	.030
-20	.005	21	.026
-21	.004	22	.023
-22	.003	23	.020
-23	.002	24	.017
-24	.002	25	.015
-25	.001	26	.013
-26	.001	27	.011
-27	.001	28	.009
-28	.001	29	.008
-29	<.001	30	.007
		31	.006
		32	.005
		33	.004
		34	.003
		35	.003
		36	.002
		37	.002
		38	.002
		39-43	.001
		44	<.001

Means and standard deviations are from Horvath and Palmatier (2008).

[6] We are not aware of anyone using five relevant questions in contemporary field practice. Nor are we aware of any accredited polygraph training program that is presently teaching this technique. The 5th relevant question in the studies on this technique ("Were you assigned to be a guilty person during this research?") is thought to be of unknown ecological and external validity. This information is included for completeness because the available studies on the MSU-MGQT satisfied the requirements for inclusion in the APA (2011) report.

Appendix O

Integrated Zone Comparison Technique[7,8]

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-21	12	19	4

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
13	.067	-5	.091
12	.040	-4	.078
11	.023	-3	.067
10	.012	-2	.057
9	.006	-1	.048
8	.003	0	.040
7	.001	1	.033
6	.001	2	.028
5	.000	3	.023
4	.000	4	.019
3	.000	5	.015
2	.000	6	.012
1	.000	7	.010
0	.000	8	.008
-1	.000	9	.006
-2	.000	10	.005
-3	.000	11	.004
-4	.000	12	.003
-5	.000	13	.002
-6	.000	14	.002
-7	.000	15	.001
-8	.000	16	.001
-9	.000	17	.001
-10	.000	18	.001
-11	.000	19	<.001
-12	.000		
-13	<.001		

Means and standard deviations are from studies on the Integrated Zone Comparison Techniques, as reported by American Polygraph Association (2011).

[7] This boutique technique involves the use of a proprietary scoring system. Accuracy rates reported in studies on this technique were reported as approaching perfection, and were shown in the 2011 meta-analytic survey to be an outlier to the distribution of other results. Studies supporting this technique have been described as substantially methodologically flawed, and it is considered unlikely that the reported accuracy rates will be achieved in field settings. Although a complete discussion of the statistical errors is beyond the scope of this publication, readers can refer to the 2011 report for more information on the publication citations and discussion about the limitations of the reported findings. Inclusion of information on this technique is not intended to be an endorsement or criticism of the technique. Instead a summary of the reported information is included here so that readers can more fully understanding the issues and controversies, and for completeness of inclusion of all polygraph techniques that were included in the 2011 meta-analytic survey.

[8] Cutscores initially recommended by the developer of the Integrated Zone Comparison Technique (Gordon & Cochetti, 1987) were +18 and -18 for truth-telling and deception, and were subsequently reported as +13 and -13. It is unclear why these cutscores were recommended, as information in the published on this technique suggest that a deceptive cutscore of +5 should be expected to achieve the same near-zero false-positive error rate as -13 or -18.

Appendix P

Matte Quadri-track Zone Comparison Technique[9,10,11]

Grand total scores			
Guilty cases		Innocent cases	
Mean	SD	Mean	SD
-9.1484	2.8433	6.0017	3.099

Deceptive Scores		Truthful Scores	
Score	p-value	Score	p-value
6	.500	-9	.479
5	.373	-8	.343
4	.259	-7	.225
3	.166	-6	.134
2	.098	-5	.072
1	.053	-4	.035
0	.026	-3	.015
-1	.012	-2	.006
-2	.005	-1	.002
-3	.002	0	.001
-4	.001	1	<.001
-5	<.001	2	<.001
		3	<.001

Means and standard deviations are from Matte and Reuss (1989).

[9] This boutique technique involves the use of a proprietary scoring system. Accuracy rates reported in studies on this technique were reported as approaching perfection, and were shown in the 2011 meta-analytic survey to be an outlier to the distribution of other results. Studies supporting this technique have been described as substantially methodologically flawed, and it is considered unlikely that the reported accuracy rates will be achieved in field settings. Inclusion of information on this technique is not intended to be an endorsement or criticism of the technique. Instead a summary of the reported information is included here so that readers can more fully understanding the issues and controversies, and for completeness of inclusion of all polygraph techniques that were included in the 2011 meta-analytic survey. Although a complete discussion of the statistical errors is beyond the scope of this publication, information provided by the developers suggests that 95% of truthful persons can be expected to produce 3-chart totals of +9 or greater, while 95% of deceptive persons can be expected to produce 3-chart total scores of -19 or lower. Readers can refer to the 2011 report for more information on the publication citations and discussion about the limitations of the reported findings.

[10] Published procedures for this technique involve the average total score per chart instead of the more common grand total score. This will require the summation of all scores for all charts and division of the result by the number of charts. We note a procedural inconsistency with statistical and mathematical theory which holds that average scores can be subject to linear multipliers or divisors, but standard deviations are not subject to linear multiplication or division. The standard deviation of three charts is not a simple linear multiplier of the standard deviation of one chart or the average of charts. Instead the variance, calculated as the variance as the square of the standard deviation, can be subject to linear multiplication, after which the standard deviation can be recalculated as the square root of the result.

[11] Information is shown for truthful scores to +3, beyond the limit of necessity, only because the developers have recommended cutscores of -5 and +3 per chart. It is unclear why these cutscores were chosen, as a cutscore of +1 would compute to the same result based on information published by the developers.