**Heuristic Principles to Select Comparison and Relevant Question Pairs When Scoring Any CQT Format**

R

C

By Raymond Nelson1

Fourteen different named techniques met the criteria for inclusion in a me- ta-analytic survey of validated poly- graph techniques (American Poly-

graph Association, 2011). Some com- monly used question formats have multiple names. For example: the Test for Espionage and Sabotage (TES; Re-

Raymond Nelson is a psychotherapist, behavior scientist, trainer, and polygraph examiner who has conducted several thousand polygraph examinations. He has expertise in working with perpetrators and victims of sexual crimes and other abuse and violence. Mr. Nelson has expertise in statistics and data analysis and is one of the de- velopers of the OSS-3 scoring algorithm and the Empirical Scoring System. He is a researcher for Lafayette Instru- ment Company (LIC), a developer and manufacturer of polygraph and life-science technologies. Mr. Nelson is a past-President of the American Polygraph Association (APA), currently serving as an elected Director. Mr. Nelson teaches and lectures frequently throughout the United States and internationally, and has published numerous studies and papers on all aspects of the polygraph testing, including the psychological and physiological basis, test data analysis, faking/countermeasures, interviewing and question formation and test target selection. Mr. Nelson has been involved in policy development at the local, state, national and international levels in both polygraph and psychology, and has testified as an expert witness in court cases in municipal, district, appellate, superior and su- preme courts. Mr. Nelson is also the academic director of the International Polygraph Training Center (IPTC) and the Escuela National del Poligrafo (ENPOL). There are no proprietary or commercial interests and no conflicts of interest associated with the content of this publication. The views and opinions expressed in this publication are those of the author and not necessarily those of the APA, LIC or IPTC. Mr. Nelson can be reached at raymond. nelson@gmail.com.

search Division Staff, 1995a; 1995b) is also known as the Directed Lie Screen- ing Test (DLST; Nelson & Handler, 2012)

. The YouPhase technique (Depart- ment of Defense, 2006) is also referred to as a BiZone. And there are a num- ber of similar techniques referred to as a family of Zone Comparison Tech- niques (ZCT) or Zone of Comparison Techniques – a needless discussion that serves only to distract – including the Federal ZCT (Department of De- fense, 2006), Utah 3 question format (Kircher & Raskin, 1988), RCMP/CPC A-Series (Honts, 1996), and a small number of boutique formats. We also have the two versions of AFMGQT (Department of Defense, 2006), one of which is structurally virtually identical to the Utah 4 question format (Han- dler, 2006, Handler & Nelson, 2008; Raskin, Honts, Nelson & Handler, 2015) though it is used differently.

An example of the different names and rules associated with these named test formats can be seen in the selection of which comparison ques- tion to use when numerically scoring each relevant question. The Federal YouPhase/BiZone format, with two relevant questions at positions 5 and 7 in the question sequence, is common- ly scoring by comparing each relevant question to the adjacent comparison question in the question series, either preceding or following, depending on

which comparison question produced the greater change in physiological activity. In contrast, the Federal ZCT format, with three relevant questions at positions 5, 7 and 10, is evaluated by comparing question 5 to the ad- jacent comparison questions at posi- tions 4 and 6, while questions 7 and 10 are scored only to the preceding comparison questions at positions 6 and 9 in the question sequence. The Utah 3-question format – referred to by some as the Utah Zone or Utah ZCT has three relevant questions at positions 5, 8 and 11 that are evalu- ated with the preceding comparison questions at positions, 4, 7, and 9. The RCMP/CPC A-Series format – struc- turally somewhat similar to the Fed- eral ZCT though derived from the Utah 3-question format – has relevant questions at positions 5, 7 and 10, and there is anecdotal information that some examiners evaluate these using only the preceding comparison ques- tions at 4, 6 and 9, while others use the Federal rules for which relevant ques- tion number 5 is evaluated against the comparison questions at positions 4 and 6 depending on which compar- ison question produces the greater change in physiology.

Polygraph exams conducted using the AFMGQT V1 format are common- ly scored by evaluating the relevant question at position 4 with the com-

parison questions at position 3 and 5, depending on which comparison question produced the greater change in physiological activity, while the rel- evant question at position 6 is evalu- ated with the comparison questions at positions 5 and 7 and the relevant question at position 8 is evaluated with the comparison questions at po- sitions 7 and 9. The relevant question at position 10 is evaluate only with the comparison question at position 9. These rules differ slightly from those for the AFMGQT V2 format for which the relevant questions at positions 4 and 5 are both evaluated with the comparison questions at positions 3 and 6, depending on which compar- ison question produced the greater change in physiological activity, while the relevant questions at positions 7 and 8 are evaluated with the compari- son questions at positions 6 and 9. The Utah 4-question format, structurally similar to the AFMGQT V2, uses rele- vant question in position 5 and 6 that are evaluated with comparison ques- tions at positions 4 and 7, with the rel- evant questions in positions 8 and 9 evaluated with comparison questions at positions 7 and 10. DLST/TES format is scored with each of the three or four repetitions of relevant questions 1 and

2 evaluated with comparison ques- tions 1 or 2 depending which compar- ison question produced the greater change in physiology. Over-empha-

sis named techniques can create the mistaken impression that all of these named techniques are different.

All of these examination formats are, in fact, more similar than they are dif- ferent. All of these test formats con- form to the same underlying theory of polygraph testing: that greater chang- es in physiological activity are loaded at different types of test stimuli as a function of deception or truth-telling in response to the investigation tar- get stimuli. Over-emphasis on named techniques andassociated rules serves to distract attention away from things actually do affect the test performance characteristics, and serves to cause un-necessary confusion about under- lying scientific principles that form the basis of the necessary test structure.

**Two important considerations**

The traditional system of named tech- niques requires that trainees and working professionals devote atten- tional and memorization resources to master a dizzying array of names, question numbers and different rules. In reality the name of the technique is not important. What is important are two fundamental considerations: 1) is the technique to be used as a diag- nostic test or as a screening test, and

1. how many relevant questions are going to be employed.

**Diagnostic or screening test?**

The first consideration – diagnostic or screening test – is a simple func- tion of whether there is or is not a known problem. The existence of a known problem – incident or allega- tion – will mean that the purpose of the polygraph test will be to gain an accurate diagnosis of the examinee’s involvement or non-involvement in the known problem. Because the test result is intended to become a basis of evidence or information on which to make a decision that can affect the future rights and liberties of the ex- aminee, examiners are ethically obli- gated to conduct diagnostic tests in a manner that permits them to make one decision about deception and truth-telling concerning the known allegation or known incident. Test conducted in this way – with the goal of making a single classification – are known to have the smallest potential for error, and for this reason they can provide the greatest level of precision or accuracy.

Screening tests are all tests conducted in the absence of a known problem – known incident or known allegation – with the goal of identifying and rectify- ing unknown problems while they are easily managed (i.e., before they turn into larger problems that are more dif- ficult and costly to manage). Because

they are looking for possible problem where there are no known problems, screening tests are often more use- ful when they can address a number of possible problems. Although it is possible to conduct a screening test regarding a single issue, when there is a single issue of concern to the screening context – including screen- ing polygraph tests – they are often conducted as multiple-issue tests (i.e., tests that are interpreted with the as- sumption of independent criterion variance). Multiple-issue tests are in- tended to make multiple simultane- ous classifications, each of which is, in reality, a probability decision.

Because these tests involve multiple simultaneous probability decisions, they introduce the potential for mul- tiplicity effects that are well known to statisticians and researchers. These multiplicity effects will increase the total probability of error for the test as function of the number of probabili- ty decisions. The overall effect is that multiple-issue tests can have corre- spondingly less precision and accu- racy compared to single-issue exams. Statistical corrections can be used to reduce the influence of multiplicity ef- fects (Nelson, 2015), but these are im- perfect solutions and multiple-issue tests will still have inherently weak- er precision than a test designed to make a single probability decision. It

is for this reason that multiple issue test formats are not used in diagnos- tic testing contexts – when there is a known problem – for which the test is intended to be a basis for decision and action. In the diagnostic context it would be ethically questionable to select a testing format that is known to offer weaker precision than a readi- ly available alternative.

**How many relevant questions?**

The second consideration – how many relevant questions will be used in the test question sequence – will manifest itself in different effects for diagnos- tic and screening tests. This will be a function of the different types of deci- sion rules used to classify the results of diagnostic and screening tests.

Diagnostic tests – because of the eth- ical mandate for high precision when the test results are intended as a ba- sis for action in response to a known problem – will be interpreted with a single classification at the level of the test as a whole. In the diagnostic con- text, more relevant questions means a larger data set or more data on which to calculate a precise result. More data will contribute to smaller errors of measurement and greater overall pre- cision. Said differently, more relevant questions will contribute to greater precision for diagnostic polygraphs.

Screening tests – because they are in- tendedtosearchforpossibleproblems in the absence of any known problems

* are often most useful when the test sensitivity can be broadened to mul- tiple issues of concern to the screen- ing context. Although screening tests can be formulated to investigate a sin- gle issue of screening concern, most screening polygraphs are conducted as multiple-issue exams, for which the use of multiple investigation target is- sues will mean multiple classifications or decisions. Because every testing classification or decision is fundamen- tally a probability statement – includ- ing when these are reduced to cat- egorical conclusions – every classifi- cation will have some accompanying probability of error. Due the effects of multiplicity (i.e., multiple simulta- neous statistical comparisons), test- ing errors are cumulative for multiple probability decisions. The result is the test error rates can increase as a func- tion of the number of screening issues and decisions to be made. In other words, use of more relevant questions will contribute to reduced precision in for multiple issue screening poly- graphs .

A simple heuristic for understanding polygraph testing formats

The simple heuristic to understanding polygraph techniques is this:

t \*T UIF UFTU B EJBHOPTUJD UFTU PS

screening test, and how many rel- evant questions will be used?

At the present time there is no evi- dence of any effect or effectiveness for of the esoteric ideas associated with named polygraph techniques – including the use of time-bars, symp- tomatic questions, special types of relevant questions or any technical questions intended to achieve testing objectives that are generally related to questions about internal consis- tency and internal validity. All that is necessary to understand the relation- ship between polygraph test ques- tion format and test effectiveness is to known whether a test is intended for diagnostic or screening purposes, and how many relevant questions will be used in the test question sequence.

The answer to this simple heuristic question, along with the choice of scoring or transformation method, will provide all the information that is needed to select the correct tables from the available published norma- tive reference data (Nelson & Handler, 2015). Statistical reference tables can be used to optimize practical deci- sions about test sensitivity, specificity and tolerance for false-negative and false-positive errors. Ultimately, the answer to this heuristic question will also guide both the selection of deci-

sion rules and the selection of which comparison question to use when quantifying responses to the relevant target stimuli.

Information published by the Amer- ican Polygraph Association (2011) showed that polygraph testing for- mats interpreted at the level of the test as a whole have greater overall precision, with effectively similar test sensitivity, compared to test formats that are interpreted with an assump- tion of independent criterion vari- ance. Some work has been published on polygraph decision rules (Senter, 2003; Senter & Dollins, 2003) showing that interpretation at the level of the test as a whole provides greater over- all precision or accuracy than inter- pretation at the level of the individu- al questions. Having derived a simple evidence-based heuristic for concep- tualizing the issues that determine the relationship between the test format and test effectiveness (i.e., Diagnostic or screening test? And how many rele- vant questions?), we can now proceed to define a simple heuristic for select- ing the correct comparison question for any comparison question poly- graph technique.

**A simple heuristic to select comparison and relevant question pairs**

Although less work has been done on the matter of the selection of the com- parison question when quantifying responses to the relevant question, Raskin, Honts, and Kircher (2014) re- portedthatdiagnostictestresultswere optimized by evaluating each relevant stimulus with the preceding compar- ison question, while multiple issue exams were optimized by evaluating each relevant question with either the preceding or following comparison questions depending on which has produced the greater change in physi- ological activity. Unpublished analysis by this writer has found similar results . The following heuristic can be applied to virtually any of of the polygraph techniques described in the meta-an- alytic survey (American Polygraph As- sociation, 2011) to achieve the correct selection of relevant and comparison stimulus pairs:

* 1. Do not evaluate a relevant stim- ulus with a comparison question if there is a data artifact, instruction, or other type of test stimulus be- tween the presentation of relevant and comparison stimulus. This is subject to one exception (see #5 below) when the preceding com- parison question is unusable.
	2. Whenever possible, evaluate each relevant stimulus with the two comparison questions that are ad- jacent with (i.e., immediately pre- ceding or immediately subsequent to) the relevant question in the question sequence, selecting the comparison question that produc- es the greater change in physiolog- ical activity.
	3. Pairs of adjacent relevant ques- tions can be evaluated with the comparison questions that are im- mediately preceding or subsequent to the paired relevant questions. In other words, it is acceptable to there can be one additional rele- vant question between a relevant question and preceding or subse- quent comparison question.
	4. Evaluate the relevant question with the preceding comparison question when there is no compar- ison question immediately subse- quent to a relevant question in the question sequence.
	5. If the comparison question pre- ceding the relevant stimulus is un- usable then a relevant question can be evaluated with the subsequent comparison question even when there is a single neutral question in question sequence between the relevant question and subsequent

comparison question, but only as long as there are no data artifacts, instructions or other questions in between the relevant question and subsequent comparison question.

This heuristic, once learned, can be applied to any polygraph test format regardless of the name of the format.

**Summary**

In decades past, named polygraph techniques were affiliated with partic- ular schools, typically named after the founder or originator of the polygraph technique andschool. Polygraphtrain- ing during those early pioneering and development years may have func- tioned as a form of professional im- printing, whereby once having been trained at a particular school that used a particular techniques, a professional field examiner may have had a sense that the technique was very different from the polygraph techniques used by other examiners who were trained at other schools. Today we know that comparison question polygraph tech- niques are largely similar to each oth- er in their structure, use and interpre- tation. We also know that the name of a technique is far less important that its principles and procedures, most of which are increasingly similar for many polygraph techniques in use today.

At the present time there is little or no information suggesting that any named polygraph technique is actu- ally superior to others, though there are some practical differences that remain important and can be used to optimize testing accuracy or preci- sion. Those basic differences involve only two concerns: whether the poly- graph test is intended for diagnostic or screening purposes, and how many relevant questions are included in the test question sequence. These two is- sues will determine, together with the numerical transformation method, the selection of the statistical refer- ence model (or normative reference tables), and the selection of optimal decision rules.

In general, diagnostic polygraph with more relevant questions will have greater precision, simply because they can have more data for any given num- ber of repetitions of the test question sequence. Multiple-issue screening polygraphs with more relevant ques- tions – interpreted with an assump- tion of independent criterion variance

* involve more probabilistic decisions and therefore more opportunity for testing error and inconclusive results, and this may be loaded for examinees who are actually truthful. Use of statis- tical corrections can reduce but does not completely eliminate these multi- plicity effects.

The tradition of named polygraph techniques no longer serves the needs of the polygraph profession, and instead distracts attention away from these more important consider- ations. Named polygraph techniques, while they once served to help orga- nize conceptual information about defined polygraph techniques, have the effect today of causing people to hold onto an arcane and incorrect viewpoint that comparison question polygraph techniques are actually different simply because they have different names. More importantly, continued reliance on the tradition of named techniques serves to inhibit progress by maintaining burdensome memorization tasks that cannot be easily described in simple heuristics, cannot be easily generalized to other formats. In other words, the rules for each named format have, until this time, had to learned and memorized separately - without the option to apply generalizable knowledge from one named format to another. A sim- pler and more adaptive solution will be to begin to think about polygraph techniques in terms that are descrip- tive of the factors that govern both the use and effectiveness of the test, and to learn select relevant and com- parison question pairs using a simpler heuristic that can be generalized to any comparison question test format.

The selection and choice of which comparison question to use may re- main an important consideration when evaluating the relevant ques- tions in both diagnostic and screen- ing polygraphs. This should, however, be guided by a conscious awareness of the principles that govern the test effectiveness. The simple heuristic, de- tailed earlier, involves the evaluation of each relevant question with either the preceding or subsequent compar- ison question when possible, and with the preceding comparison question when not. Paired relevant questions are handled the same way. The only remaining caveat is that the occur- rence of data artifacts, instructions, or other types of test stimuli in between the relevant and comparison question should be a cause for concern caution.

The tradition of named polygraph techniques is no longer serving the needs and goals of the profession, and should be replaced by a concep- tual and procedural heuristic that can be more easily generalized to any polygraph technique. Remaining an- chored to arcane notions that cannot advance our profession further will only increase the likelihood that poly- graphtestingmay bereplaced by new- er credibility assessment technologies that are perceived as less burdened with the typical problematic percep- tions that polygraph testing is overly

complex, overly subjective, and overly vulnerable to human error. Polygraph field examiners and polygraph train- ing programs will be wise to encour- age professional field practitioners to think about and describe their work and field practices in terms that will not remain locked to arcane tradition,

and will better ensure the continued usefulness of the polygraph in a field practice context that will likely in- clude the increased juxtaposition of polygraph methodologies with other emerging scientific technologies for credibility assessment.

**References**

American Polygraph Association (2011). Meta-analytic survey of criterion accura- cy 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.

Department of Defense (2006). Federal Psychophysiological Detection of Decep- tion Examiner Handbook. Retrieved from <http://www.antipolygraph.org/> documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Re- printed in Polygraph, 40(1), 2-66.

Handler, M. (2006). The Utah Probable Lie Comparison Test. Polygraph, 35(3), 139- 146.

Handler, M., Nelson, R., (2008). Utah Approach to Comparison Question Polygraph Testing. European Polygraph, 2, 83-110.

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

Senter, S. M. & Dollins, A.B. (2003). New Decision Rule Development: Exploration of a two-stage approach. Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph, 37(2), 149-164.

Senter, S. M. (2003). Modified general question test decision rule exploration. Poly- graph, 32, 251-263.

Nelson, R. (2015). Bonferroni and Šidák corrections for multiplicity effects with sub- total scores of comparison question polygraph tests. Polygraph, 44(2), 162- 167.

Nelson, R. & Handler, M. (2012). Monte Carlo study of criterion validity of the Di- rected Lie Screening Test using the Empirical Scoring System and the Objec- tive Scoring System version 3. Polygraph, 41(3), 144-155.

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

Raskin, D. C., Honts, C. R., & Kircher, J. C. (2014). Credibility Assessment. Academic Press.

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

Research Division Staff (1995a). A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats. Re- port number DoDPI94-R-0008. DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 26(2), 79-106.

Research Division Staff (1995). Psychophysiological detection of deception accura- cyrates obtainedusingthetest for espionageandsabotage. DoDPI94-R-0009. DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 27, (3), 171-180.