Does Spot Scoring and Relevant and Comparison Question Order Help or Hurt the Examinee? A Computer Analysis of Ground Truth Verified Army and Air Force MGQT and Federal ZCT Exams

Keith Hedges[1] and George Deitchman[2]

Abstract

The Army and Air Force variants of the Modified General Question Test (MGQT) and Federal Zone Comparison Test (ZCT) have been widely used for investigations in private, governmental and police work. The Army MGQT has had the reputation as being a deceptive test. Previous review of field testing has shown the Army MGQT calls many non-deceptive examinees deceptive. The use of spot scores for each question on the Army MGQT test protocol may be one of the causes for this problem, a problem not seen with the Federal Zone when total scores are used to classify examinations as deceptive or non-deceptive. When positive spot scores are required to classify ZCT examinations as non-deceptive the false positive rate again increases. This study uses computer scoring and comparisons between the same relevant and control questions as performed by human scorers. The frequency of negative scores at each relevant question was compared to ground truth known deceptive and non-deceptive examinees and there was a notable deceptive bias. The analysis reveals serious difficulties for the non-deceptive examinee with traditional Army MGQT. The Air Force MGQT suffered the same problem to a much lesser extent. Additionally, when comparison questions preceded relevant questions in test design and when the more tracing comparisons were available per test decision, results improved. This appears to support the observed balance and strength of the Federal Zone and very similar Utah formats.

Introduction

The Army and Air Force MGQT and Federal Zone Comparison Test of deception are types of Comparison Question Tests (CQT) that are used extensively to assist in the resolution of criminal investigations and other serious matters. A CQT may consist of symptomatic questions (SY) (Backster, 1976; Hess, 1976), irrelevant questions (IR), a sacrifice relevant question (SR), relevant questions (R) and finally the probable-lie comparison (PLC) or directed-lie comparison (DLC) question. The decision concerning a subject's truthfulness or deception to a relevant issue (e.g. "Did you shoot that man?" "Did you steal that missing money?") is dependent upon the reaction relationship of the relevant and comparison questions. Proposed in the early 1960s Backster's theory of Psychological Set states in essence the subject will focus and react to the stimulus presenting the greatest psychological threat to his or her wellbeing, either the relevant or probable-lie comparison questions. The most recent theory is the subject is expected to react most strongly to those questions, whether relevants or comparisons which have the greater differential salience for the subject (Senter, Weatherman, Krapohl & Horvath,

[1] Keith Hedges completed a 576 hour polygraph school at the Penn Valley Community College in Kansas City, Missouri in 1982. He began measuring and researching polygraph tracings while working on a Ph.D. at Southern Illinois University, Carbondale, IL. He developed a computerized program to confirm examiner scoring in 1997 and marketed the program until selling it in 2005. He considers the American Polygraph Association David L. Motsinger Award as his greatest achievement.

[2] George Deitchman, Ph.D., LMFT is a therapist residing in Jacksonville, Florida. He has been treating sex offenders since the early 1990s. He became a polygraph examiner in 2009 when he attended the American International Institute of Polygraph. He provides treatment for adults and conducts evaluations and risk assessments for adults and adolescents. He has instructed physiology and psychology as part of the basic polygraph curriculum.

2010). An examination containing stronger psychophysiological reactions to the comparison questions than the relevant questions is classified as a truthful or No Deception Indicated (NDI) result while the reverse relationship is considered a deceptive or Deception Indicated (DI) result.

The methods used to determine if the relevant or comparison questions are producing the strongest reactions vary from one polygraph school to the next. Some compare relevant question reactions to the stronger adjacent comparison question component tracings, others compare to the preceding comparison question or even the weaker adjacent bracketing comparison question tracing. The recorded physiology normally consists of two respiration tracings (thoracic and abdominal), the electrodermal tracing and the cardiograph tracing.

Two polygraph examiners can evaluate the same relevant and comparison question reactions and depending on the examiner using the strongest or weakest comparison question the end results could present opposing opinions concerning the subject's veracity. The strength and weakness of polygraph examiner numerical scoring is its semi-objective nature.

Is there a cure for the above problem? One potential approach is a computer scoring program. A computer can be programmed to make consistent objective measurements and reaction comparisons. Often field polygraph examiners correctly state that a computer cannot identify some of the reaction criteria they were taught at their respective polygraph schools or in training seminars. If a computer scoring program identified and compared the polygraph tracings with sufficient accuracy, the field examiner would be more likely to use the computerized analysis as a quality control tool.

A computer program is useless without a skilled examiner. A computer program can do many things well, but it cannot do at least two critical tasks in analyzing the polygraph tracings; only the examiner can detect artifacts and eliminate tracings unsuitable for program analysis and only the examiner can determine if the charts are of sufficient quality for program analysis. Perhaps the best relationship is a kind of marriage between the polygraph examiner and program analysis of reactions.

This paper will address the accuracy of three CQT techniques: 1) Federal Zone Comparison Technique (ZCT), 2) Army Modified Question Test (MGQT), 3) and what the authors will operationally define as the Air Force Modified General Question Test (AF) Versions 1 and 2.

The Federal ZCT has a question sequence of IR1, SR2, SY3, C4, R5, C6, R7, SY8, C9, and R10. R5 is evaluated against the strongest tracing reactions occurring between C4 and C6, R7 is evaluated against C6 and R10 is compared to C9.

The Army MGQT has a question sequence of IR1, IR2, R3, IR4, R5, C6, IR7, R8, R9, and C10 during the first two charts of an examination, and is typically followed by a mixed sequence chart: IR4, IR1, R5, C6, IR2, R3, C10, R9, C6, R8, C10. On the first two charts R3 and R5 are scored against C6, with R8 and R9 being scored against C10. The third chart was scored with R5 being compared to C6 and the remaining relevant questions being scored against the strongest bracketed comparison question. Around 1983 the U.S. Army training school changed the order of how the questions for the first two charts could be evaluated; R3 and R5 were still evaluated against C6 but now C6 could also be used to evaluate R8 as well. In addition the Army further changed the requirements for positioning of questions on the third chart and began allowing for the weakest relevant question to be placed in the number 3 position; however, the scoring rules remained the same.

The AF MGQT has two versions, each consisting of two, three, or four relevant questions. In Version 1 the four question sequence is IR1, SR2, C3, R4, C5, R6, C7, R8, C9, and R10. The second and third charts can be mixed however all relevant questions are evaluated against the stronger of the two adjacent comparison questions. In Version 2 the four question sequence is IR1, SR2, C3, R4, R5, C6, R7, R8, and C9. Again the second and third chart may be mixed however in this version both relevant questions are compared against the stronger of the two bracketing

comparison questions. The AF Version 1 format was operationally considered valid if the charts contained alternating comparison and relevant questions and the examination originated from the DoDPI data disk MGQT directory. The sequences fitting this definition are noted below:

Question Sequence	N of cases
CRCRC	24
CRCRCR	14
CRCRCRC	31
CRCRCRCR	19
CRCRCRCRC	11

The three and four question sequences sometimes ended with a comparison question in some charts but not in others in the same examination. These examinations were analyzed twice, once without the final comparison and again including the final comparison.

The AF Version 2 was considered valid if the examination originated in the DoDPI data disk MGQT directory and a comparison question was followed by two relevants and then another comparison. The sequences fitting this definition are noted below.

Question Sequence	N of cases
CRRC	14
CRCRRC	29
CRRCRC	14
CRRCRRC	11

The Federal ZCT consists of three relevant and three probable-lie comparison questions. The Army MGQT consists of four relevant and two probable-lie comparison questions, the AF MGQT Version 1 consists of four relevant and four comparison questions, and the AF Version 2 consists of four relevant questions and three comparisons. With the Army MGQT RRCRRC sequence, the probable-lie comparison questions are repeated to fulfill the RCRCRCRC mixed sequence. Regardless of the technique or sequence, the R and C questions are all differently worded. All techniques are normally repeated three times and each repetition is referred to as a chart.

Krapohl (2006) reviewed the literature and analyzed the differences between the Army MGQT and Federal ZCT tests and noted average accuracy; it is reproduced as Table 1 for easy reference. The Army MGQT's accuracy rate was 61% after excluding 21% inconclusive results and Federal ZCT's accuracy was 89% after excluding 16% inconclusives. The Federal ZCT had an accuracy rate of 97% for deceptive subjects and 82% for non-deceptive subjects. There was no information presented for the Air Force MGQT test. The Federal ZCT was 1% below the American Society for Testing and Materials (ASTM, 2005) standard of 90% for evidentiary testing. The Army MGQT results were 19% below the level of accuracy required by ASTM (80% excluding inconclusives) standard for investigative polygraph techniques. The Utah Probable-Lie Test (PLT) (Handler, 2006) met the ASTM standard with a 91% correct classification rate. When used in the three question format the Utah PLT has the same "CRCRCR" sequence as the Federal ZCT.

Table 1. Krapohl's (2006) rank order of polygraph techniques by accuracy (excluding inconclusives)

Technique	Accuracy without Inconclusives	Inconclusive Rate
Utah ZCT	91%	12%
Federal ZCT	89%	16%
TES	88%	2%
RI	83%	0%
Reid	83%	6%
CIT	80%	0%
MGQT	61%	21%

It is no surprise the Federal ZCT surpassed Army MGQT accuracy. The ZCT contains one more comparison question and one less relevant question per chart at least during the first two charts of the examination.

The non-deceptive Federal ZCT is classified by an overall positive (grand total) and all positive spot scores. The Utah (Handler, 2006), Backster and Matte (Matte, 1996) Zones all classify non-deceptive examinations based upon the overall total without regard to spot scores.

The Federal ZCT and Army MGQT are both criminal test formats focusing on a single incident. That is where the similarity ends though. Relevant questions in the Federal ZCT normally focus on the primary and often secondary suspected actions. The Army MGQT's format permits relevant questions focusing on different actions concerning the issue at hand. By focusing the relevant questions more narrowly (one behavior or action), the same basic behavior in question is presented more times per test. The Army MGQT is handicapped by scoring each individual relevant question (spot) where any negative relevant question score precludes a non-deceptive classification. The Air Force MGQT Version 1 test resembles the Federal ZCT in many ways, primarily containing additional comparisons to the relevant questions, and comparison questions preceding relevant questions. The results are still handicapped by scores to the relevant questions being classified on a question-by-question "spot" basis. Again, any negative scoring relevant question precludes a non-deceptive decision. The Army MGQT relevant questions may be less focused and cover related but not identical issues in each chart sequence. The test data analyzed was stripped of the actual questions and it was not possible to analyze question formulation or how the relevant questions were presented to the examinee. The following scoring examples will clarify the advantages and disadvantages of the three CQT under discussion.

Method

Relevant / Comparison Question Comparisons

The Federal ZCT relevant and comparison question sequence is as follows:

IR1 SR2 SY3 C4 R5 C6 R7 SY8 C9 R10

R5 is compared to the strongest reacting tracings in either C4 or C6. R7 is compared to the tracings in C6 and R10 is compared to the tracings in C9. The number of Federal ZCT R and C tracing comparisons total 12 per chart, R5 = 6, R7 = 3 and R10 = 3. This is a total of 36 tracing comparisons for a three-chart examination. Federal ZCT classifications will be made based on the grand total (one) score for all R questions as is the case in the Utah, Backster and Matte tests. Traditional scoring of the Federal ZCT dictates a grand total score of -6 or less for a deceptive result or a -3 or less at any one spot will result in a deceptive classification. A grand total score of +6 or greater with all spots having a positive score are required a non-deceptive classification.

The Army MGQT relevant and comparison question sequence is as follows:

IR1 IR2 R3 IR4 R5 C6 IR7 R8 R9 C10.

For the first two charts R3 and R5 tracings are compared to C6 tracings. R8 and R9 tracings are compared to C10 tracings. Sometime in 1983 C6 was also allowed to evaluate the R8 tracing as well.

For the third chart beginning in 1983 the Army MGQT mixed sequence might be:

IR4, IR1, R9, C6, IR2, R3, C10, R5, C6, R8, C10

The least reactive relevant question is placed as the first R question. With the exception of the first relevant on a mixed chart all remaining R's are compared to the strongest bracketed comparison question.

The number of Army MGQT R and C tracing comparisons total 12 in examinations scored before 1983 and 15 during and after 1983 in each of the first two charts, R3 = 3, R5 = 3, R8 = 3 or R8 = 6 post 1983 and R9 = 3. The above listed mixed sequence chart R and C tracing comparisons total 21, i.e. R5 = 6, R3 = 6, R9 = 3 and R8 = 6. This is a total of 45 tracing comparisons for a pre-1983 three-chart examination and a total of 51 tracing comparisons for a post-1983 examination. The data were analyzed with the post 1983 method. The Army MGQT classifications are based on each of four R scores. Under traditional scoring a negative three at any spot will result in an automatic deceptive classification and a non-deceptive classification must have a +3 score at each spot. Scores falling between -2 and a +2 are classified inconclusive. Any negative spot score will preclude a non-deceptive classification in this project.

There are two versions of the AF MGQT. They are distinctly different in structure. Version 1 utilizes alternating comparison and relevant questions while Version 2 utilizes alternating comparison and paired relevant questions. A sample sequence of a Version 1 four relevant examination is IR1, SR2, C3, R4, C5, R6, C7, R8, C9, R10 and a sample sequence of a Version 2 four relevant question examination is IR1, SR2, C3, R4, R5, C6, R7, R8, C9. The number of AF MGQT R and C tracing comparisons for the Version 1 sample sequence of IR1, SR2, C3, R4, C5, R6, C7, R8, C9, R10 total 21 per chart, R4 = 6, R6 = 6, R8= 6, R10 = 3. This is a total of 63 tracing comparisons for a three-chart four relevant question examination. The number of tracing comparisons for a Version 2 examination is R4 = 6, R5 = 6, R7 = 6 and R8 = 6. This is a total of 24 tracing comparisons per chart and 72 for a three-chart examination. Army MGQT spot score requirements apply to the Air Force Versions. Any negative spot score precludes a non-deceptive classification.

The above examinations were scored on a three-position rather than a seven-position scale. The authors doubt the Deceptive and Non-deceptive seven-position scale requirements are valid for the three-position scale.

The ZCT cases were scored based on the seven-position scale total score requirements and again utilizing the requirements adjusted for the three-position scale. The three-position scale is capable of a maximum of three points for a given relevant spot while the seven-position scale is capable of up to nine points. The three-position scale will therefore be adjusted to 3 / 9 = .33 and .33 * 6 = 2. Therefore an adjusted inconclusive zone of plus or minus two was used for the total score requirements.

The advantages and disadvantages become rapidly apparent considering the above information. In this project a three-chart Federal ZCT utilizes 36 tracing comparisons to obtain one total and three spot scores, the Army MGQT utilizes 51 (post-1983) tracing comparisons for four individual spot scores, the Air Force MGQT Version 1 utilizes 63 tracing comparisons for four spot scores, and Version 2 utilizes 72 tracing comparisons for four spot scores.

The Federal ZCT comparison tracing to total score ratio is 36:1 while the three spot scores average ratio is 36:3. The AF MGQT Version 1 comparison tracing total score ratio is 63:4, the Version 2 ratio is 72:4 and the Army MGQT tracing total score ratio is 51:4 (post-1983). Dividing the total number of tracings by the number of scores used to

obtain a DI or NDI test classification clarifies the data even further. The Federal ZCT based on a single total is 36/1 = 36. The Federal ZCT spot scores average 36/3 = 12. The AF MGQT Version 1 averages 63 / 4 = 15.75, and Version 2 averages 72 / 4 = 18.5, and the Army MQGT is 51/4 = 12.75 (post-1983).

Based upon the above information we can make some predictions. The Federal ZCT total score will be the most accurate as it has the most data as judged by the number of comparisons to score ratio at 36 comparisons for a single total score. The Federal ZCT total and positive spot score NDI requirement will result in lower non-deceptive classification accuracy averaging just 12 comparisons per spot. The AF MGQT Versions 1 and 2 should have higher correct spot classifications than the ZCT as the AF MGQT spot scores rely upon 15.75 (Version 1) and 18.5 (Version 2) comparisons per spot respectively. The Army MGQT will suffer the most in non-deceptive classifications at just 12.75 (post-1983) comparisons per spot score.

The purpose of this paper is to objectively test Krapohl's (2006) findings by a simple computer program designed to compare relevant and comparison questions in much the same way as a polygraph examiner. The comparison criteria will be different but the question comparisons will be the same. AF MGQT Version 1 examinations containing an extra comparison question were analyzed both with and without this question.

All tracing measurements were given the same weight scored on a three-position scale. The only possible scores will be +1, -1 and 0. The scored reaction differences will require a visibly perceptual difference. Traditional required cutoff scores were examined and altered to maximize their utility.

If Krapohl's (2006) review of the literature reflects the actual accuracy of the Federal ZCT and Army MGQT, our initial prediction is the Federal ZCT total relevant scores will produce the greatest accuracy. However, if any relevant question score is negative and the negative score precludes a non-deceptive classification, non-deceptive accuracy will diminish. The AF MGQT Versions 1 and 2 are predicted to place second in accuracy due to the advantage of additional comparisons and the Army MGQT is predicted to produce about 60% accuracy.

The following steps must be taken to attain this objective:

1. Create measurements for the respiration, electrodermal and cardiograph recordings/ tracings and a computer program to extract these measurements from computerized polygraph examinations.

2. Create a computer program that meets or exceeds ASTM standards for evidentiary Federal ZCT polygraph examinations. The program's data treatment will replicate some features of a previous program by the first author marketed under the name Identifi. To avoid infringement upon the rights of Lafayette Instrument Company some specific tracing comparison data will be left vague. These values will be identified in the text as undisclosed values.

3. Test the program on Federal ZCT examinations for accuracy based on the total score of the relevant questions. A total positive score will be classified non-deceptive and a total negative score will be classified as deceptive.

4. Test the Federal ZCT examinations again using relevant question spot scores to assess accuracy without an inconclusive zone. This is to focus on the impact of the requirement for a positive score at each spot for a non-deceptive classification. Any examination containing a negative relevant question score will be classified deceptive.

5. Run the program on Army MGQT and AF MGQT examinations to obtain classification comparisons for the three CQT's and comparison to the Krapohl findings.

Tracing Measurements

The data extraction program was written in 2007 by Fred Vater (deceased). The following criteria for each measurement were used by Vater during program creation:

Timm's (1981) respiration line length measurement was subjected to the abdominal respiration tracing from question onset through 15 seconds. If less than 15 seconds was available for measurement the measurement was discarded. The shorter line length is deemed the stronger of C and R lengths compared.

The electrodermal tracing degree of reaction was measured from a baseline established at the point of tracing rise to the highest point achieved by the reaction. The electrodermal tracing was considered eligible for measurement provided the rise began between question onset and six seconds following question end and there was a minimum of 15 seconds between question onset for the question being measured and the question onset of the following question. The larger rise was deemed the stronger of C and R rises compared.

The cardiograph area under the curve was measured between the diastolic tips and a baseline established at the point of rise if the tracing rise began between question onset and not later than six seconds following question end. In Vater's data extraction program of 2007, if the diastolic tips returned to or fell below a previously established baseline and recovered in less than 1.97 seconds, a new baseline was established at the recovery point and the additional area(s) were summed for a total area below the diastolic tips. Measurement of any question ended 15 seconds after question onset. In the event a rise in the cardiograph tracing surpassed 15 seconds, measurement ended at 15 seconds, a vertical line was made to baseline and the area was calculated. The largest area was deemed the stronger of the cardiosphygmographic areas compared.

Data Treatment and the Computer Program

The above measurements were extracted from all questions in an examination. The abdominal respiration tracing was converted to percentages by chart. All abdominal respiration measurements were summed and each question in the chart was divided by the sum and multiplied by 100 to convert the values from decimals to values greater than zero. The shorter respiration line length is considered stronger. The Relevant and Comparison question respiration tracings were compared by subtracting the Comparison from the Relevant. If the difference was greater than a specific undisclosed positive number it was scored a +1. If the number was negative and was less than a specific undisclosed number it was scored a -1. Differences falling between the undisclosed numbers were scored zero. In the electrodermal and cardiograph measurements larger values are more powerful. The electrodermal tracing required a one chart division difference to be considered perceptible. The Relevant and Comparison electrodermal tracings were compared by subtracting the Relevant tracing from the Comparison tracing. A positive difference of greater than or equal to one chart division was scored +1. A negative difference greater than or equal to one chart division was scored -1. Differences of less than one chart division were scored zero. The cardiograph was scored by adding the Relevant question value to the selected Comparison question tracing value and dividing the Comparison value by the summed Relevant and Comparison values. If the result was less than or equal to an undisclosed value less than .5, a score of -1 was assigned. If the result was greater than or equal to an undisclosed value greater than .5, a score of +1 was assigned. If the result was between the undisclosed values required for a +1 or -1, a score of zero was assigned.

Examination Selection

The examinations used in this study were obtained from a CD labeled Department of Defense Polygraph Institute, Psycho-physiological Detection Field Data, Accurate 03May02. The Federal Zone Comparison Test and MGQT examinations were extracted from the \JHUAPL\ZONE\$$*.* files and \JHUAPL\MGQT\$$*.* files. The directories were sorted into sub-directories for the ZONE and MGQT directories by ground truth, i.e., \JHUAPL\ZONE\DI and \JHUAPL\MGQT\ NDI. The Army and Air Force MGQT examinations were combined in the \JHUAPL\MGQT directory. There were no individual Army or Air Force MGQT directories. A Federal ZCT examination was considered valid if the first

Comparison was labeled "C4" and the Relevant and Comparison question sequence was CRCRCR. Furthermore, the relevant questions were labeled R5, R7 and R10. The Army MGQT examinations were considered valid if the first chart sequence had relevant and comparison questions in the order of RRCRRC and relevant question labels were R3, R5, R8 and R9. The AF Version 1 format was operationally considered valid if one of the first two charts contained alternating comparison and relevant questions and the examination originated from the DoDPI data disk MGQT directory. Version 2 examinations were considered valid if one of the first two charts contained at least one set of paired relevant questions. Each AF examination contained at least two and a maximum of four relevant questions. The Federal ZCT, Army MGQT and AF MGQT formats required a minimum of two correctly sequenced charts.

Results

Federal ZCT

The total score was used to classify deceptive and non-deceptive examinations. A total score greater than zero was classified non-deceptive and a total score less than zero was classified deceptive. The scores classified a total of 442 Deceptive (M = -6.57, SD = 5.63) and 161 Non-deceptive (M = 4.70, SD = 4.95) examinations. There were 382 (86%) deceptive and 136 (84%) non-deceptive examinations correctly classified. The traditional conclusive examination totals of +/-6 were calculated and the inconclusive zones were in excess of 50%. The requirements of a +/-2 total score was tested, the scores produced 359 (92%) correct deceptive and 140 (88%) correct non-deceptive results. The revised inconclusive zone produced an average correct classification rate of (92% + 88%) / 2 = 90%. The average error rate was 10% (30 deceptive and 17 non-deceptive) and the average inconclusive rate was 13% (53 deceptive and 4 non-deceptive).

The percentages of negative scoring relevant question spots for both deceptive and non-deceptive cases are listed in Table 3. As can be seen, only 53% of non-deceptive ground truth tests did not have any negative spot scores.

Table 2. Federal ZCT with Grand Total with no spot scoring using (-2 < INC < +2)

	Deceptive (442)	Truthful (161)	
DI decision	359 (92.3%)	17 (10.8%)	PPV = .955 (true positive rate)
NDI decision	30 (7.7%)	140 (89.2%)	NPV = .824 (true negative rate)
	0.923 sensitivity (hit-rate)	0.892 specificity (pass-over rate)	

Proportion Deceptive correctly identified is Sensitivity. Proportion Truthful correctly identified is Specificity. Proportion of DI decisions that are correct is Positive Predictive Value (PPV). Proportion of NDI decisions that are correct is Negative Predictive Value.

Table 3. Number of negative scoring spots per examination by Ground Truth DI/NDI

					N
# Neg. Spots	0	1	2	3	
DI:	14 (3%)	72 (16%)	163 (37%)	193 (44%)	442
NDI:	85 (53%)	58 (36%)	16 (10%)	2 (1%)	161

Army MGQT Results

The program classified 167 deceptive and 45 non-deceptive examinations. Any negative relevant question spot score precluded a non-deceptive classification. There were 165 (99%) correct deceptive and 30 (67%) incorrect non-deceptive examinations containing negative spot scores. The percentage of negative scoring relevant question spots for both deceptive and non-deceptive cases are listed in Table 4.

Table 4. Army MGQT number of negative scoring spots per examination by Ground Truth DI/NDI

# Neg. Spots	0	1	2	3	4	*N*
DI:	2 (1%)	5 (3%)	40 (24%)	41 (25%)	79 (47%)	167
NDI:	15 (33%)	5 (11%)	15 (33%)	9 (20%)	1 (2%)	45

The relevant question means and standard deviations are listed to illustrate a revised inconclusive zone would probably not result in a significant increase in non-deceptive conclusions as many examinations would be classified inconclusive. The relevant question data is displayed in the Table 5.

Table 5. Army MGQT Means and Standard Deviations by DI/NDI Ground Truth

Relevant Question	DI		NDI	
	M	*SD*	*M*	*SD*
R3	-3.21	2.28	-.60	2.86
R5	-2.90	2.55	.29	3.13
R8	-2.22	2.37	.60	2.70
R9	-1.99	3.06	1.89	3.14

The Army MGQT scoring was revised by defining a non-deceptive classification as an examination containing spot scores greater than or equal to zero, a -1 in any spot was inconclusive and a -2 or less in any spot was scored deceptive. This scoring gives an advantage to non-deceptive examinations. The modified scoring classified 159 (95%) deceptive examinations correctly, 2 (1%) incorrectly and 6 (4%) inconclusive. The scoring classified 15 (33%) non-deceptive examinations correctly, 26 (58%) incorrectly and 4 (9%) inconclusive. The average accuracy rate was 64% (excluding inconclusives).

Air Force MGQT Version 1 Results

The scoring classified 70 deceptive and 28 non-deceptive examinations. Any negative relevant question score precluded a non-deceptive classification. Excluding all comparison questions at the end of a chart from analysis produced 65 deceptive (93%) and 14 non-deceptive (50%) examinations correctly classified. The average correct classification rate was 72%.

Table 6. Air Force MGQT Version 1 number of negative scoring spots per examination by Ground Truth DI/NDI

# Neg. Spots	0	1	2	3	4	*N*
DI:	5 (7%)	12 (17%)	25 (36%)	18 (26%)	10 (14%	70
NDI:	14 (50%)	10 (36%)	4 (14%)	0	0	28

The modified inconclusive zone used in the Army MGQT section was applied to the Air Force MGQT Version1 examinations. The scoring produced 61 (87%) correct deceptive classifications, 5 (7%) errors and 4 (6%) inconclusives. The scoring produced 14 (50%) correct non-deceptive classifications, 7 (25%) errors and 7 (25%) inconclusives. The average correct classification rate was 69% (excluding inconclusives).

Table 7. Air Force MGQT Version 1 decisions with values greater than or equal to zero for non-deceptive decisions, -1 inconclusive and -2 or less for deceptive decisions (-2 < INC < 0)

	Deceptive (70)	Truthful (28)	
DI decision	61 (92.4%)	7 (33.3%)	PPV = 0.897 (true positive rate)
NDI decision	5 (7.6%)	14 (66.7%)	NPV = 0.737 (true negative rate)
	0.924 sensitivity (hit-rate)	0.667 specificity (pass-over rate)	

The relevant question means and standard deviations indicate an unbalanced inconclusive zone might increase accuracy. The relevant question data is displayed in Tables 8 and 9.

Table 8. DI Examinations

Relevant Question	*N* of cases	DI	
		M	*SD*
1ST Relevant	70	-2.26	2.98
2nd Relevant	70	-2.97	3.26
3rd Relevant	50	-1.94	2.88
4th Relevant	27	-1.93	2.60

Table 9. NDI Examinations

Relevant Question	*N* of cases	NDI	
		M	*SD*
1ST Relevant	28	1.29	2.32
2nd Relevant	28	2.04	2.22
3rd Relevant	24	.71	1.94
4th Relevant	2	.50	2.12

The AF MGQT Version 1 was also analyzed including the comparison questions ending the charts. The program classified 70 deceptive and 28 non-deceptive examinations. Any negative relevant question score precluded a non-deceptive classification. There were 66 deceptive (93%) and 19 non-deceptive (68%) examinations correctly classified. Using the final comparison question increased non-deceptive classifications by 9%. The average correct classification rate was 81%. The question format with the final comparison question was also scored with the Army MGQT modified inconclusive zone. The modified inconclusive zone produced 61 (94%) deceptive, 4 (6%) errors and 5 (7%) inconclusives. The scoring produced 19 (73%) non-deceptive, 7 (27%) errors and 2 (7%) inconclusives. The average correct classification rate was 84% (excluding inconclusives). See Table 10.

Table 10. Air Force MGQT Version 1 decisions with non-deceptive values greater than or equal to zero, -1 inconclusive and less than or equal to -2 for deceptive decisions (-2 < INC < 0)

	Deceptive (70)	Truthful (28)	
DI decision	61 (93.8%)	7 (26.9%)	PPV = 0.897 (true positive rate)
NDI decision	4 (6.2%)	19 (73.1%)	NPV = 0.826 (true negative rate)
	0.938 sensitivity (hit-rate)	0.731 specificity (pass-over rate)	

Air Force MGQT Version 2 Results

The program classified 39 deceptive and 29 non-deceptive examinations. Any negative relevant question score precluded non-deceptive classification. Scoring classified 34 deceptive (87%) and 24 non-deceptive (83 %) examinations correctly. The average correct classification rate was 85 %.

Table 11. Air Force MGQT Version 2 number of negative scoring spots per examination by Ground Truth DI/NDI

# Neg. Spots	0	1	2	3	4	*N*
DI:	5 (13%)	6 (15%)	15 (38%)	10 (26%)	3 (8)%	39
NDI:	24 (83)	5 (17%)	0	0	0	29

Version 2 was also subjected to the Army MGQT modified inconclusive zone. The scoring produced 30 (86%) correct deceptive classifications, 5 (17%) errors and 4 (10%) inconclusives. The scoring produced 24 (86%) correct non-deceptive classifications, 4 (14%) errors and 1 (3%) inconclusives. The average correct classification rate was 86% (excluding inconclusives).

Table 12. Air Force MGQT Version 2 decisions with non-deceptive examinations scoring zero or greater, -1 inconclusive and minus two or less deceptive (-2 < INC < 0)

	Deceptive (39)	Truthful (29)	
DI decision	30 (85.7%)	4 (14.3%)	PPV = 0.882 (true positive rate)
NDI decision	5 (14.3%)	24 (85.7%)	NPV = 0.828 (true negative rate)
	0.857 sensitivity (hit-rate)	0.857 specificity (pass-over rate)	

Discussion

1. The test data analyzed was from charts confirmed by outside means (ground truth). The data comes from criminal investigative tests, possibly multi-facet, but most likely not multi-issue testing. The original questions have been stripped out and were not available for inspection. The impact of test questions which address multi-facet or multi-issue testing was not studied in this research.

2. A computerized scoring program was developed to analyze test data by a comparison of respiration movement line length, vertical deflection of the EDA and area under the diastolic curve line, much as a human scorer would. The resulting values were used to predict Federal ZCT DI and NDI decisions. These predictions were compared to ground truth with a resulting average accuracy rate of 90%. This result differs from Krapohl's finding of 89% by 1%. Using any negative relevant question spot score to preclude a NDI decision clearly causes a decrease in NDI Federal ZCT accuracy. The difference between permitting any negative scoring spot to preclude a non-deceptive result (53% correct non-deceptive result) and use of a simple total score inconclusive zone of plus or minus two (88% correct non-deceptive) is nothing short of dramatic.

3. The Army MGQT relies on vertical scoring of each relevant spot score, and gives a high level of false positive decisions. This is due in effect to putting too much value on too few presentations of a single question and having fewer comparisons. The Army MGQT also uses question order where the relevant questions precede the comparison questions. This question order may promote larger responses to relevant questions, whereas, a preceding comparison question may reduce the relevant question response somewhat. The preceding comparison question is favored by many other CQT formats. In addition, a high number of false positive decisions had negative spot scores, leading to many ground truth NDI subjects being called DI. Negative spot scores resulted in the correct classification of only 33% non-deceptive examinations and a simple inconclusive zone increased this number to just 37%. Using any negative spot score to preclude a non-deceptive classification produced an average accuracy rate of 66%

and a simple inconclusive zone increased this number to just 68%. The average classification of the two methods was 67%. This average differs from Krapohl's 61% (without inconclusives) by 6%.

4. The scoring classified AF MGQT examinations with greater accuracy than the Army MGQT. AF Version 1 excluding the final comparison question and any negative spot score precluding a non-deceptive classification, classified 93% DI and 50% NDI results in an average correct classification rate of 72%. The use of an inconclusive zone classified 92% DI and 67% NDI results in an average correct classification rate of 80%.

The AF Version 1 including the final comparison question and any negative spot precluding a non-deceptive classification, classified 93% DI and 68% NDI correctly, resulting in an average 81% correct classification rate. The addition of an inconclusive zone classified 94% DI and 73% NDI correctly, resulting in an average 84% classification rate. It would appear use of the final comparison question contributes to positive spot scores with no decline in DI classifications and an 18% increase in NDI classifications.

5. The AF MGQT Version 2 spot scores also classified examinations with greater accuracy than the Army MGQT. AF Version 2 and any negative spot score precluding a non-deceptive classification, classified 87% DI and 83% NDI results in an average correct classification rate of 85%. The use of an inconclusive zone classified 86% DI and 86% NDI results in an average correct classification rate of 86%.

As predicted, AF MGQT Version 1 and 2 NDI spot classifications accuracy was greater than those of the Federal ZCT. There is insufficient data to form any conclusions related to which AF MGQT Version or sequence is best. It should be noted that in Version 1 24% of the examinations contained two relevant questions, 45% of the examinations contained three relevant questions and 30% of the examinations contained four relevant questions. In Version 2 19% of the examinations used two relevant questions, 66% of the examinations used three relevant questions and 15% of the examinations used four relevant questions.

6. The authors would be extremely hesitant to attempt generalizations and recommendations for the Army MGQT if not for the consistency of Krapohl's findings and the scored classifications. Due to the extremely small accuracy differences we cannot recommend the Army MGQT for field examinations. It is difficult to produce non-deceptive results with any scoring technique as is illustrated by the Krapohl ZCT test accuracy review. Even if the amount of error in replication of non-deceptive Army MGQT classifications is corrected by the error increase noted for the ZCT the difference is still lacking in power. The noted Krapohl ZCT result of 82% for non-deceptive examinations minus our result of 53% using negative spot scores equals 29% and our Army MGQT non-deceptive classification rate of 33% plus 29% equals 62%. A correct non-deceptive classification rate of 62% still does not permit us to recommend the Army MGQT for field use. Perhaps this is the reason the Federal Polygraph School (NCCA) no longer teaches the Army MGQT.

7. The authors do not have supporting information for the AF MGQT and due to the small number of NDI cases are extremely hesitant to make recommendations related to the AF MGQT test. The results of this study show promise for the format and we believe if the results of the present study can be replicated by future research it may indeed be a viable test.

8. What can be said about the Federal ZCT results? The Federal ZCT total score method used in this study outperforms the Air Force and Army MGQT tests in field testing. The Utah PLC may be slightly more accurate but the Federal ZCT has more longevity. The similarities between the two formats in terms of comparison questions preceding the relevant questions may be significant. The lack of the ZCT's symptomatic questions does not seem to be detrimental to the Utah PLC. Also the focus of both tests on three questions aimed at one issue may explain their similar accuracies. This simple computer program classified 90% of the Federal ZCT examinations correctly. As was earlier stated, the best combination is a type of marriage between the polygraph examiner and computer program.

References

Backster, C. (1976). Outside-issue factor. *Handout from polygraph examiner course PE-68*. Backster School of Lie Detection, San Diego, CA.

Handler, M.D. (2006). Utah probable lie comparison test. *Polygraph*, 35, 139-148.

Hess, C. (1976). The "symptomatic" question. Lecture notes from polygraph examiner course PE-68. Backster School of Lie Detection, San Diego, CA.

Krapohl, D.J. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph*, 29, 185-194.

Krapohl, D. (2006). Validated polygraph techniques. *Polygraph*, 35, 149-155.

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

Senter, S., Weatherman, D., Krapohl, D. & Horvath, F. (2010). Psychological set or differential salience: a proposal for reconciling theory and terminology in polygraph testing. *Polygraph*, 39, 109-117.

Timm, H.W. (1982). Analyzing deception from respiration patterns. *Journal of Police Science & Administration*, 10, 47-51.