Criterion Validity of the United States Air Force Modified General Question Technique and Three Position Scoring

Mark Handler and Raymond Nelson

Abstract

Criterion accuracy profiles were calculated for the USAF-MGQT format for PDD examinations. The dimensional profile of criterion accuracy was compared to results from seven-position scores and ESS scores from earlier studies. Criterion accuracy of three-position scores of the USAF-MGQT exams was significantly greater than chance, with no significant differences in test sensitivity to deception compared to the results of the seven-position and ESS TDA models in the previous studies. Test specificity was significantly weaker for the three-position model compared to the other TDA models and truthful case inconclusive results were significantly higher for three-position compared to the other models. Additional research is recommended to determine whether normative data and the use of statistically optimal cutscores would increase test specificity and decrease inconclusive results with the three position scoring model.

Introduction

The United States Air Force Modified General Question Technique (USAF-MGQT) (Department of Defense, 2006) is a variant of the Comparison Question Test (CQT) that is commonly used in multi-issue screening contexts and multi-facet diagnostic contexts when the relevant questions (RQs) are assumed to be independent.[1] Nelson, Handler, Morgan and O'Burke (2011) and Senter, Waller and Krapohl (2008) addressed the criterion validity of the USAF-MGQT using seven-position scoring (DoDPI, 2006). Nelson and Blalock (in press), and Nelson, Blalock and Handler (in press) investigated the criterion validity of USAF-MGQT examinations with ESS scores. The present study is an effort to extend our knowledge-base regarding the criterion accuracy of the USAF-MGQT when the three-position scoring model (DoDPI, 2006) is applied.

Method

USAF-MGQT data was obtained from a previous study (Nelson, Handler, Morgan & O'Burke, 2011) involving three experienced Iraqi National Police polygraph examiners from the Iraq Ministry of Interior's National Information and Investigation Agency (NIIA) and Iraq's Ministry of Defense's Directorate General for Intelligence and Security (DGIS) polygraph programs. Participants received their initial training through American Polygraph Association certified and U.S. Department of Defense instructors. It is estimated that the three examiners combined conducted in excess of 1,000 examinations in field settings in Iraq.

Acknowledgments

The authors thank Ben Blalock, Donald Krapohl and Edward Hoffman for their thoughtful reviews and comments to earlier drafts of this paper. The authors would grant unlimited use and duplication rights to any polygraph school accredited by the American Polygraph Association or the American Association of Police Polygraphists. Questions and comments are welcome at polygraphmark@gmail.com or raymond.nelson@gmail.com.

[1] *Independence* in scientific test theory refers to the notion that the variance of response to each individual stimulus is not affected by and does not affect the variance of response to other stimuli. Stimuli for single-issue examinations are assumed to be non-independent, and the results are therefore calculated at the level of the test as a whole. Stimuli for multi-issue and multi-facet examinations are assumed to be independent. Results for multi-issue and multi-facet examinations are therefore calculated at the level of the individual questions, though the results are reported at the level of the test as a whole.

Data for this study were a matched random sample of field examinations (N = 22) selected from the confirmed case archive at the Department of Defense. Examinations were conducted by U.S. federal and local law enforcement agencies using a variant of the USAF-MGQT consisting of two, three or four relevant question (RQs). Nine cases each had two and three RQs and the remaining four cases had four RQs. Eleven cases were confirmed as truthful via confession and evidence that inculpated an alternative suspect, while the remaining 11 samples were confirmed as deceptive via a combination of confession and extra-polygraphic evidence.

Seven-position scores for a confirmed case sample of USAF-MGQT examination were transformed to their three-position counterparts. Seven-position and three-position test data analysis models taught by the U.S. Department of Defense differ only in the use of linear ratios for assigning higher order scores when using the seven-position model. Evidence exists to support there are no statistically significant differences in the seven-position and three-position test data analysis (TDA) rubrics for assigning first order scores. Earlier studies (Harwell, 2000; Krapohl, 1999; Krapohl, 2010, Nelson, Handler, Blalock & Cushman, in press) have shown that seven-position scores can be converted to three-position values.

All examinations consisted of three test charts and followed the procedures described by the Department of Defense (DoDPI, 2006). Because the criterion status of the RQs in multi-facet and multi-issue examinations conducted with the USAF-MGQT format are assumed to be independent, categorical decisions are made at the level of the individual questions, using the spot-score-rule (SSR) (Light, 1999; Swinford, 1999). Grand total scores are not used with the SSR, and all categorical decisions are made using the subtotal scores for individual RQs.

USAF-MGQT examination test results are always classified at the level of the test as a whole even though categorical decisions are made using the subtotal scores for the individual RQs. As a practical matter and assuming data of normal interpretable quality, when using the SSR, the overall test result is classified as truthful when all individual RQs produce a truthful numerical score, and are classified as deceptive whenever one or more RQs produces a deceptive score. Deceptive results for diagnostic examinations of known incidents or known allegations are commonly reported as Deception Indicated (DI), while results of screening examinations, conducted in the absence of a known incident or known allegation, are commonly reported as Significant Reactions (SR). This distinction is important because screening examinations are not designed, and not intended, to provide diagnostic accuracy or diagnostic conclusions. Numerical scores that are insufficient to reach a decision are commonly reported as Inconclusive (INC) or No Opinion (NO), meaning that no evidence-based professional opinion can be rendered.

Previous research (Barland, Honts & Barger, 1989; Podlesny & Truslow, 1993) has not supported the hypothesis that independent RQs in psychophysiological detection of deception (PDD) examinations can provide question-level sensitivity and specificity. In other words, studies do not support the ability to determine deception to some RQs and truthfulness to other RQs within a single examination. These data do support the notion that whenever one or more individual RQs produce a deceptive result, no-opinion can be rendered regarding individual RQs that do not also produce a deceptive numerical score.

Three-position cutscores, as taught by the Department of Defense (2006), are not based on normative data or statistical analysis. In the absence of evidence to guide decisions about optimal cutscores for the three-position TDA model, the traditional solution has been to use cutscores for three-position TDA models which are identical to those used with seven-position TDA models. Traditional cutscores for independent RQs of multi-facet and multi-issue examinations, using the SSR, are -3 and +3. This means that any subtotal less than or equal to -3 will result in a deceptive classification of the overall test result, and that a truthful classification will result when every subtotal score is greater than or equal to +3.

Results

All tests were tested with level of statistical significance set at alpha = .05, except as labeled otherwise.

A dimensional profile of criterion accuracy was calculated, including: sensitivity to deception, specificity to truthfulness, false-negative and false-positive error rates, inconclusive rates for deceptive and truthful cases, positive predictive value (PPV), negative predictive value (NPV), the proportion of correct decisions for deceptive and truthful cases, and the unweighted accuracy (i.e., without inconclusives) for the proportion of correct decisions and inconclusive results. Monte Carlo methods were used to estimate statistical confidence intervals for the dimensional profile of criterion accuracy. Multi-variate ANOVAs were used to compare the profile of criterion accuracy with the accuracies of ESS and seven-position scores of USAF-MGQT examinations, as reported in previous studies (Nelson, Blalock & Handler, in press; Nelson Handler, Morgan & O'Burke, 2011). Statistical confidence intervals for the proportion of interrater agreement were calculated using bootstrapping methods. The dimensional profile of criterion accuracy for three-position USAF-MGQT scores can be seen in Table 1, along the criterion accuracy profiles of seven-position and ESS scores for USAF-MGQT examinations.

Table 1. Mean (SD) and {95% CI} for dimensional profile of criterion accuracy for three TDA models.

Accuracy Dimension	3-position	7-position	ESS
Unweighted Mean Accuracy	.739 (.059) {.622 to .856}	.754 (.043) {.669 to .838}	.882 (.034) {.815 to .950}
Unweighted Mean Inconclusives	.421 (.043) {.335 to .506}	.241 (.041) {.158 to .323}	.183 (.038) {.108 to .258}
Sensitivity	.780 (.060) {.661 to .899}	.809 (.055) {.701 to .917}	.831 (.051) {.730 to .931}
Specificity	.180 (.052) {.077 to .282}	.364 (.067) {.23 to .496}	.616 (.069) {.479 to .752}
FN Error	.010 (.013) {.001 to .037}	.010 (.014) {.001 to .038}	.010 (.014) {.001 to .039}
FP Error	.186 (.054) {.078 to .294}	.333 (.065) {.206 to .463}	.175 (.054) {.069 to .281}
D Inc	.209 (.059) {.092 to .325}	.182 (.053) {.075 to .284}	.158 (.050) {.059 to .257}
T Inc	.633 (.066) {.503 to .763}	.301 (.062) {.179 to .424}	.208 (.057) {.096 to .320}
PPV	.807 (.057) {.694 to .920}	.706 (.061) {.586 to .826}	.826 (.053) {.721 to .931}
NPV	.944 (.074) {.799 to .999}	.972 (.038) {.896 to .999}	.983 (.024) {.935 to .999}
D Correct	.986 (.017) {.951 to .999}	.987 (.017) {.952 to .999}	.987 (.017) {.952 to .999}
T Correct	.492 (.117) {.261 to .723}	.520 (.084) {.354 to .686}	.778 (.067) {.647 to .909}

The statistical confidence interval in Table 1 shows that the three-position TDA model is capable of providing criterion accuracy that is statistically significantly greater than chance ($p < .001$).

A series of multi-variate ANOVAs, criterion status x accuracy index, was calculated for the proportions of correct decisions with inconclusives (i.e., sensitivity and specificity), errors, and inconclusive results for the three-position, seven-position, and ESS scores of the USAF-MGQT sample. Table 2 shows that there was a significant interaction between model and case status for correct decisions with inconclusives, which precluded interpretation of the significant main effects for status and model. Figure 1 shows the plots of the proportions of correct decisions with inconclusive results. Post-hoc one-way ANOVAs showed that the difference in sensitivity to deception was not significant for the three TDA modes [$F (2,30) = 0.213$ ($p = .809$)]. However, the difference in test specificity to truthfulness was significant [$F (2,30) = 12.023$ ($p < .001$)]. Pair-wise contrasts showed that the increases in test specificity were significant for both the seven-position model [$F (1,20) = 4.707$ ($p = .042$)] and the ESS model [$F (1,20) = 25.465$ ($p < .001$)]. The difference in test specificity to truthfulness between the seven-position and ESS model was also significant [$F (1,20) = 6.685$ ($p = .016$)].

Table 2. 2 x 3 ANOVA summary for correct decisions with inconclusives.

Source	SS	df	MS	F	p	F crit .05
Model	0.656	1	0.030	8.442	0.005	4.001
Status	2.911	1	0.088	24.986	0.000	4.001
Interaction	0.413	1	0.413	116.933	0.000	4.001
Error	0.212	60	0.004			
Total	3.979	63				

Figure 1. Correct decisions including inconclusives.

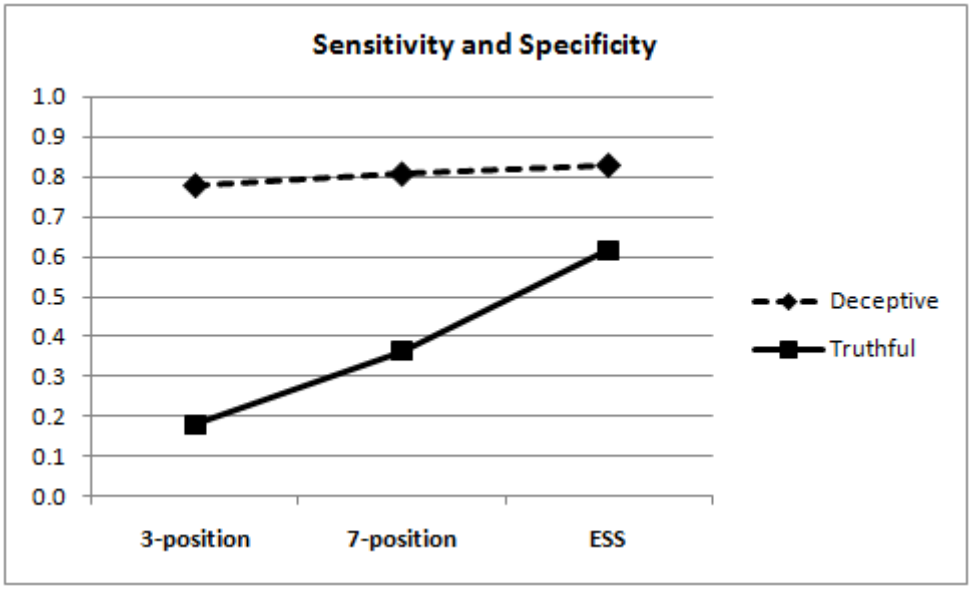


Table 3 shows there was a significant interaction of case status and TDA model for decision errors. Figure 2 shows the plots of the proportions of false-negative and false-positive errors. The significant main effect for case status cannot be interpreted without additional analysis. Post-hoc one-way ANOVA showed there was no significant difference in false-negative errors for the three TDA models. However, the difference in false-positive errors was approaching a significant level [F (2,30) = 2.321 (p = .116)]. Pair-wise contrasts were also not statistically significant at the .05 level but were approaching statistical significance for all contrasts.

Table 3. 2 x 3 ANOVA summary for errors.

Source	SS	df	MS	F	p	F crit .05
Sample	0.171	1	0.004	2.199	0.141	3.916
Status	0.569	1	0.009	4.873	0.029	3.916
Interaction	0.171	1	0.171	96.748	0.000	3.916
Error	0.223	126	0.002			
Total	0.912	129				

Figure 2. False-negative and false-positive errors.

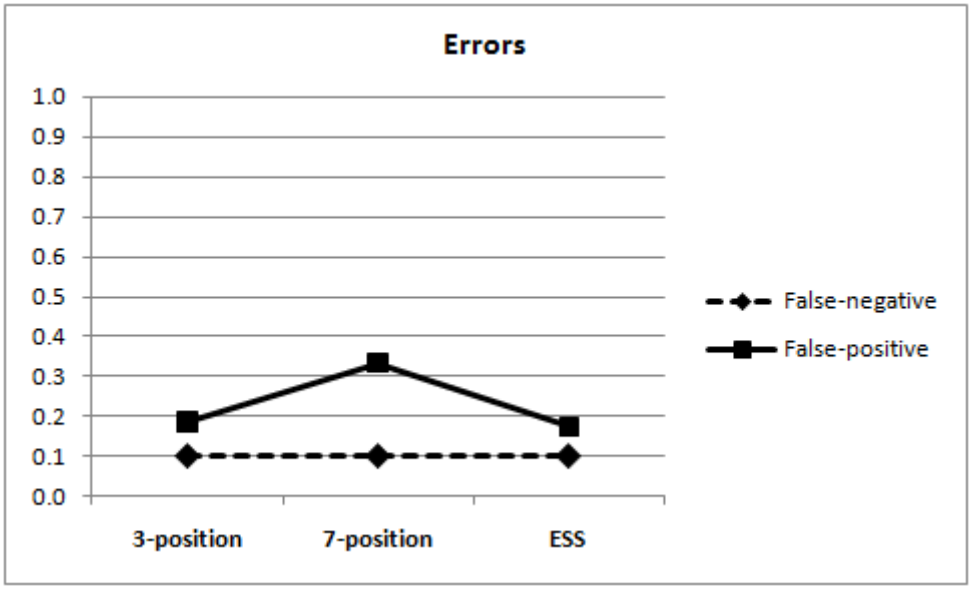


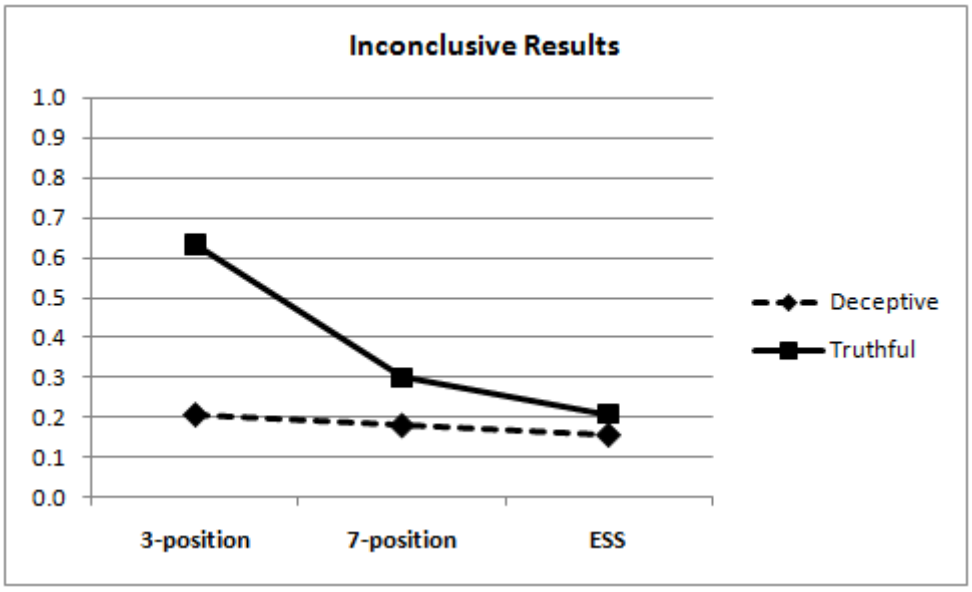
Table 4 shows there was a significant interaction between TDA model and case status for inconclusive results, which precluded interpretation of the significant main effects until additional one-way post-hoc ANOVAs were completed. Figure 3 shows the plots of inconclusive results for the three TDA models. Post-hoc one-way ANOVAs showed that the difference in inconclusive results for deceptive cases was not significant. However, the difference in inconclusive results for truthful cases was significant [F (2.30) = 13.080 (p < .001)]. Pair-wise contrasts showed that differences in inconclusives rates for the

three-position TDA model were significant at the .001 level for both the seven-position and ESS models, though the difference between the seven-position and ESS models was not significant.

Table 4. 2 x 3 ANOVA summary for inconclusive results.

Source	SS	df	MS	F	p	F crit .05
Sample	1.354	1	0.031	9.120	0.003	3.916
Status	1.289	1	0.020	5.792	0.018	3.916
Interaction	0.871	1	0.871	258.341	0.000	3.916
Error	0.425	126	0.003			
Total	3.514	129				

Figure 3. Inconclusive results.



Discussion

Results of this study support the validity of the hypothesis that three-position scores of USAF-MGQT examinations can produce criterion accuracy levels that are statistically significantly greater than chance. Test sensitivity to deception for the three-position model did not differ significantly from the seven-position or ESS models. However, test specificity to truthfulness and inconclusive rate for truthful cases did differ significantly for the three-position and other TDA models. Error rates were loaded on truthful cases (i.e., false-positives errors were more frequent than false-negative errors), and the difference in false-positive errors was approaching a significant level for the three TDA models. It is possible that observed differences in test-specificity, and related increases in error and inconclusive rates, were due to the use of sub-optimal cutscores that are not based on normative data and were originally intended for use with the seven-position TDA

model. It is also possible that statistically optimized cutscores, based on normative data, might increase test specificity and criterion accuracy for the three-position TDA model. This should be the focus of future research.

The three-point system is based on the bigger-is-better rule but it does not assign different weighting of the components to conform to what has been consistently noted in the literature, namely that electrodermal activity has shown to be the most powerful contributor to examination results (Ansley & Krapohl, 2000; Capps & Ansley, 1992; Harris, Horner & McQuarry, 2005; Kircher & Raskin, 1988; Kircher et al., 2005; Krapohl & McManus, 1999; Nelson et al., 2008; Raskin et al., 1988). The results of this study and other studies, suggest that the use of either the Empirical Scoring System or the Seven-Position System would be more advantageous in evaluating USAF-MGQT examinations. Furthermore, when there are more effective alternative TDA models available to PDD examiners in field settings it may be disadvantageous, and possibly unethical, to continue to use a scoring system that is not as noticeably effective.

Limitations of the present study include the small sample size, small cohort of scorers, and the unknown degree to which the random stratified sample will overestimate test accuracy. It is assumed that field samples, constructed of only those cases for which confirmation data can be obtained, may overestimate test accuracy. Another limitation is that this study did not address the issue of question-level specificity, which should be investigated in future studies.

Although these results support the validity of the three-position TDA model for USAF-MGQT examinations, it is difficult to advocate the use of the three-position model when more effective models are available. Inconclusive rates are significantly greater and unacceptably high for the three-position TDA model compared to both the seven-position and ESS models. It is possible that the use of statistically optimized cutscores could reduce the rate of inconclusives with the three-position model, and additional research is needed in this area.

References

Ansley, N. (1998). The validity of the modified general question test (MGQT). *Polygraph*, 27, 35-44.

Ansley, N., & Krapohl, D. J. (2000). The frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph* , 29, 169-176.

ASTM (2002). Standard Practices for Interpretation of Psychophysiological Detection of Deception (Polygraph) Data (E 2229-02). ASTM International.

Backster, C. (1963). Polygraph professionalization through technique standardization. *Law and Order*, 11, 63-65.

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). The validity of detection of deception for multiple issues. *Psychophysiology*, 26, 13 (Abstract).

Bell, B. G., Raskin, D. C., Honts, C. R., & Kircher, J.C. (1999). The Utah numerical scoring system. *Polygraph*, 28(1), 1-9.

Blalock, B., Cushman, B., & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38, 281-288.

Capps, M. H., & Ansley, N. (1992). Analysis of federal polygraph charts by spot and chart total. *Polygraph*, 21, 110-131.

Department of Defense Polygraph Institute (2006). Test Data Analysis: DoDPI numerical evaluation scoring system. Reprinted in *Polygraph*, 40(1).

Handler, M., Nelson, R., Goodson, W., & Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39(4), 200-215.

Harris, J., Horner, A., & McQuarrie, D. (2000). An evaluation of the criteria taught by the Department of Defense Polygraph Institute for Interpreting Polygraph Examinations. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272.

Harwell, E. M. (2000). A comparison of 3 and 7-position scoring scales with field examinations. *Polygraph*, 29, 195-197.

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

Kircher, J. C., Kristjansson, S.D., Gardner, M. K., & Webb, A. (2005). Human and computer decision-making in the psychophysiological detection of deception. University of Utah.

Krapohl, D. J. (1998). A comparison of 3 and 7-position scoring scales with laboratory data. *Polygraph*, 27, 210-218.

Krapohl, D. J., & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. *Polygraph*, 35(1), 55-63.

Krapohl, D. J. (2010). Short Report: A Test of the ESS with Two-Question Field Cases. *Polygraph*, 39, 124-126.

Krapohl, D. J. (2005). Polygraph decision rules for evidentiary and paired testing (Marin Protocol) applications. *Polygraph*, 34, 184-192.

Krapohl, D. J. (2006). Validated polygraph techniques. *Polygraph*, 35(3), 149-155.

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

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

Light, G. D. (1999). Numerical evaluation of the Army zone comparison test. *Polygraph*, 28, 37-45.

Nelson, R., Blalock, B., Oelrich, M., & Cushman, B. (2011). Reliability of the Empirical Scoring System with expert examiners. *Polygraph*, 40(3).

Nelson, R., Handler, M., Blalock, B., & Cushman, B. (In review). Blind Scoring of confirmed Federal You-Phase examinations by experienced and inexperienced examiners: Criterion validity with the Empirical Scoring System and the seven-position model.

Nelson, R. & Handler, M. (2010). Empirical Scoring System: NPC Quick Reference. Lafayette Instrument Company. Lafayette, IN.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40(2).

Nelson, R. & Krapohl, D. (In press). Criterion validity of the Empirical Scoring System with experienced examiners: Comparison with the seven-position evidentiary model using the Federal Zone Comparison Technique.

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.

Nelson, R., Handler, M., Morgan, C., & O Burke, P. (2012). Criterion validity of the United States Air Force Modified General Question Technique and Iraqi scorers. Polygraph, 41(1).

Podlesny, J. A. & Truslow, C. M. (1993). Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology*, 78, 788-797.

Raskin, D.C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). Validity of control question polygraph tests in criminal investigation. *Psychophysiology*, 25, 476.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology*, 37, 542-547.

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

Senter, S. M., & Dollins, A.B. (2002). 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.

Senter, S. M., & Dollins, A.B. (2008a). Optimal decision rules for evaluating psychophysiological detection of deception data: an exploration. *Polygraph*, 37(2), 112-124.

Senter, S. M., & Dollins, A.B. (2008b). Exploration of a two-stage approach. *Polygraph*, 37(2), 149-164.

Senter, S., Waller, J., & Krapohl, D. (2008). Air Force Modified General Question Test validation study. *Polygraph*, 37(3), 174-184.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28, 10-27.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/-6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists*, 7, 28-47. Reprinted in *Polygraph*, 20(2), 70-79.