Short Report: Criterion Validity of the United States Air Force Modified General Question Technique and Iraqi Scorers

Raymond Nelson, Mark Handler, Chip Morgan & Pat O'Burke

Introduction

The Modified General Question Technique (MGQT) has become a de facto family of polygraph techniques resulting from various modifications of the General Question Technique (Reid, 1947) and the Zone Comparison Technique (Backster, 1963). The United States Air Force Modified General Question Technique (USAF-MGQT) (DoDPI, 2006) is a modern variant of the Comparison Question Test (CQT) that has become widely used due to its efficient structure, based on generally accepted valid principles for CQT test construction, and its capability to adapt easily to the requirements of both multi-facet investigative needs and multi-issue screening contexts. Some studies have described the criterion validity of older variants of the MGQT (Ansley, 1998; Krapohl, 2006; Krapohl & Norris, 2000; Senter, 2003). At the present time only one published study has addressed the criterion validity of the USAF-MGQT (Senter, Waller & Krapohl, 2008). Senter et al. (2008) reported a mean blind-scoring criterion accuracy level of .849, excluding inconclusive results. The present study is an effort to extend our knowledge-base regarding the criterion accuracy of the USAF-MGQT.

Method

Participants

A cohort of three experienced Iraqi polygraph examiners from the National Information and Investigative Agency (NIIA) and Director General for Intelligence and Security (DGIS) Polygraph Programs participated in the present study. All examiners had been trained by certified instructors from the American Polygraph Association, and the US Department of Defense. It is estimated that the three examiners combined had conducted in excess of 1,000 examinations in field settings in Iraq, where they routinely used the 7-position test data analysis (TDA) model.

Data

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. All examinations were conducted using the USAF-MGQT, which exists in two closely related variants, each capable of including two to four distinct investigation targets. Three of the cases included four relevant questions and nine cases each had three and two relevant questions. Eleven cases were confirmed as truthful via confession and evidence that inculpated an alternative suspect. The remaining 11 cases were confirmed as deceptive via a combination of confession and extrapolygraphic evidence. All examinations consisted of three test charts and were conducted by US Federal and local law enforcement agencies, following the procedures described by the Department of Defense (DoDPI, 2006).

We are extremely grateful to the USF-I ITAM-MoD Intelligence Division and ITAM Police Missions who sponsor training for Akram Sabri Jwad Al NDawi, Mohammed Ahmed Mufeed Kider, Rabea Minhal Araf Al Rubaii and Mohammed Abdul Jabar Al Dulaymi. Without the commitment of these dedicated professionals none of this work would have been accomplished.

The authors thank Mr. Donald Krapohl for his thoughtful review and comments to earlier drafts of this paper. Questions and comments are welcome at raymond.nelson@gmail.com or polygraphmark@gmail.com.

Procedure

Each participant worked independently and scored all 22 cases, providing numerical scores only. Scoring tasks were completed during a two-day period, while the participants attended advanced and continuing training at an APA accredited polygraph school. Participants were provided a two-hour review of the 7-position features and transformation methods (DoDPI, 2006) prior to commencing the scoring task. The proctor and scorers were blind to the criterion status of the cases. Results were calculated using decision rules and cutscores described by the Department of Defense. Classification of the overall test result for manual scoring of the USAF-MGQT examinations was achieved using only the sub-total scores for the distinct relevant questions. A truthful classification was made if all subtotal scores were +3 or greater, and a deceptive classification was made if any subtotal score was -3 or lower. All other score combinations were classified as inconclusive. All classifications were made at the case level, not by individual relevant question.

Data were later evaluated using an automated model of the Empirical Scoring System (ESS) (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2010; Krapohl, 2010; Nelson, Handler, Shaw, Gougler, Blalock, Russell, Cushman & Oelrich, 2011; Nelson, Blalock, Oelrich & Cushman, 2011; Nelson & Krapohl, 2011; Nelson, Krapohl & Handler, 2008), an evidence-based model for manual TDA. Data were further evaluated using the Objective Scoring System, version 3 (OSS-3) (Nelson Krapohl, & Handler, 2008), a sophisticated free, open-source, and cross-platform computer algorithm for statistical analysis of all types of PDD examination results.

Deceptive classifications for ESS results were made if one or more subtotal scores was -3 or lower (alpha = .05) without Bonferroni correction. Truthful classifications were made if all subtotal scores were +1 or greater (alpha = .1) (Nelson & Handler, 2010). Unlike other manual TDA models, ESS results are based on statistical probabilities calculated using normative data, using statistical decision theory and hypothesis testing methods to control error rates and inconclusive results. To decrease inconclusive results among cases that cannot be classified as deceptive, an inverse of the Sidak correction for independent issues was used to correct for the deflation of alpha which results from the fact that false-negative errors can occur only when an examinee produces a statistically significant truthful score to all relevant questions while lying to one or more questions. The OSS-3 algorithm calculates statistical results without using integer scores or integer cutscores. Truthful classifications were made with alpha = .1 while deceptive classifications were made with alpha = .05. For comparison, OSS-3 results were tabulated using the spot-score decision rule, similar to that used in the 7-position manual and ESS TDA models. Bonferonni correction was not used when applying the spot-score decision rule to the conceptually distinct relevant questions within each confirmed case.

Results

Alpha was set at = .05 for all statistical treatments. The proportion of pair-wise decision agreement, excluding inconclusives, was unanimous ($p < .001$). For the 7-position scoring model, the mean deceptive subtotal score was $M = -2.995$ ($SD = 4.727$), and the mean truthful subtotal score was $M = 2.365$ ($SD = 3.879$).

A three-way ANOVA (scorer x criterion status x question), showed none of the main effects or interaction effects were significant at the .05 level (see Table 1). However the main effect for the individual questions was approaching a statistically significant level ($p = .065$).

A two-way ANOVA (Table 2) shows there was no significant main effect or interaction of the unsigned strength of the subtotal scores for case status and scorers.

Table 1. Three-way ANOVA summary for subtotal scores: scorer x criterion status x question.						
Source	SS	df	MS	F	p	F crit .05
Scorer	15.898	2	7.949	0.409	0.665	3.052
Status	23.036	1	23.036	1.187	0.278	3.900
Question	142.953	3	47.651	2.454	0.065	2.660
Scorer x Status	26.490	2	13.245	0.682	0.507	3.052
Status x Question	92.800	3	30.933	1.593	0.193	2.660
Scorer x Question	13.785	6	2.297	0.118	0.994	2.155
Scorer x Status x Question	117.388	6	19.565	1.008	0.422	2.155
Error	3145.120	162.000	19.414			
Total	3577.470	185				

Table 2. Two-way ANOVA summary for subtotal scores: scorer x criterion status.						
Source	SS	df	MS	F	p	F crit .05
Scorer	2.079	1	0.094	0.005	0.943	4.001
Status	7.544	1	0.229	0.012	0.912	4.001
Interaction	11.395	1	11.395	0.618	0.435	4.001
Error	1106.734	60	18.446			
Total	21.018	63				

Monte Carlo methods were used to calculate statistical confidence intervals for several dimensions of criterion accuracy, including: sensitivity to deception, specificity to truthfulness,[1] inconclusive deceptive and truthful cases, false-negative and false-positive errors, positive predictive and negative predictive value,[2] percent of correct deceptive and truthful cases excluding inconclusives. Table 3 shows the accuracy profiles, including mean scores, standard deviations, and statistical confidence interval for test results using the 7-position model, an automated version of the ESS, and the OSS-3 computer algorithm using the spot score rule.

[1] Sensitivity and specificity are calculated as the percent correct among cases with inconclusives tabulated as errors). Although not as flattering as percent correct excluding inconclusives, these statistics provide a useful indicator of classification efficiency.

[2] Positive and negative predictive value are calculated as the ratio of correct deceptive decisions to all deceptive decision, and the ratio of correct truthful decisions to all truthful decisions. Although non-resistant to difference in base-rates, these statistics provide useful information about generalizability of test results to field settings.

Table 3. Mean, (SD) and {95% CI} for dimensional profile of criterion accuracy for three TDA models.

Accuracy Dimension	7-position	ESS (automated)	OSS-3
Unweighted Mean Accuracy	.754 (.043) {.669 to .838}	.897 (.031) {.835 to .959}	.902 (.028) {.846 to .958}
Unweighted Mean Inc.	.241 (.041) {.158 to .323}	.154 (.035) {.084 to .223}	.013 (.012) {.001 to .037}
Sensitivity	.809 (.055) {.701 to .917}	.803 (.055) {.694 to .912}	.980 (.019) {.941 to .999}
Specificity	.364 (.067) {.23 to .496}	.71 (.065) {.581 to .84}	.800 (.057) {.687 to .912}
FN Error	.010 (.014) {.001 to .038}	.018 (.019) {.001 to .056}	.009 (.014) {.001 to .037}
FP Error	.333 (.065) {.206 to .463}	.158 (.053) {.053 to .262}	.181 (.054) {.075 to .288}
D Inc	.182 (.053) {.075 to .284}	.177 (.055) {.069 to .286}	.009 (.014) {.001 to .037}
T Inc	.301 (.062) {.179 to .424}	.130 (.047) {.037 to .223}	.017 (.019) {.001 to .055}
PPV	.706 (.061) {.586 to .826}	.837 (.052) {.735 to .940}	.844 (.047) {.751 to .936}
NPV	.972 (.038) {.896 to .999}	.974 (.026) {.921 to .999}	.988 (.017) {.954 to .999}
D Correct	.987 (.017) {.952 to .999}	.977 (.023) {.932 to .999}	.990 (.014) {.962 to .999}
T Correct	.520 (.084) {.354 to .686}	.817 (.060) {.698 to .937}	.814 (.055) {.706 to .923}

Table 4 shows the two-way ANOVA summaries for criterion status x scoring model for correct decisions without inconclusives. Figure 1 shows the mean plots for correct decisions without inconclusives for the three TDA models.

Table 4. ANOVA summary for accuracy without inconclusives: criterion status x scoring model.

Source	SS	df	MS	F	p	F crit .05
Model	0.313	1	0.014	5.807	0.019	4.001
Status	1.182	1	0.036	14.626	0.000	4.001
Interaction	0.329	1	0.329	134.130	0.000	4.001
Error	0.147	60	0.002			
Total	1.824	63				

Figure 1. Mean plot of correct decisions with inconclusives.

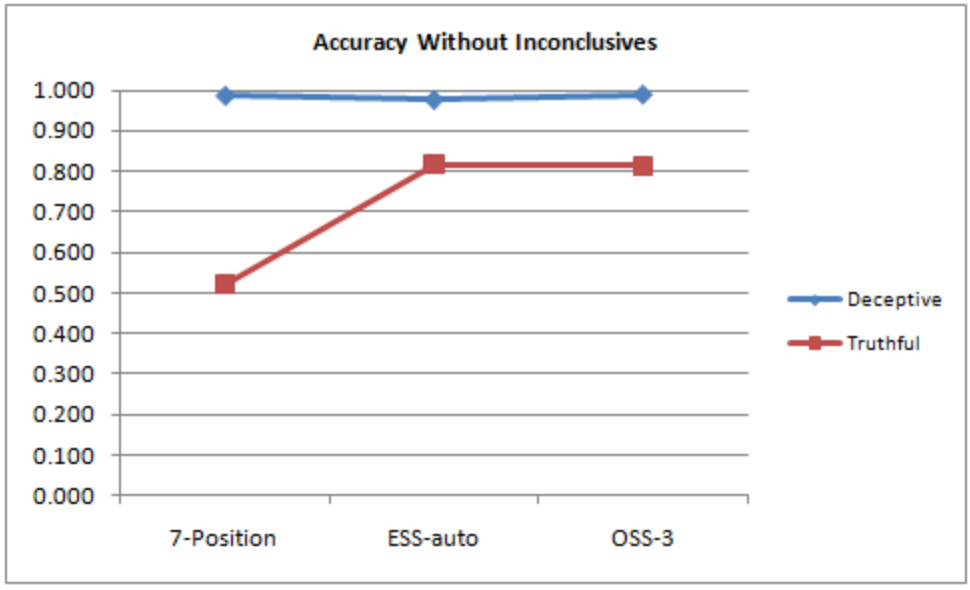


Table 5 shows the two-way ANOVA summaries for criterion status x scoring model for decision errors. Figure 2 shows the mean plots for errors for the three TDA models.

Table 5. ANOVA summary for decision errors: criterion status x scoring model.

Source	SS	df	MS	F	p	F crit .05
Model	0.096	1	0.004	2.436	0.124	4.001
Status	0.742	1	0.022	12.598	0.001	4.001
Interaction	0.106	1	0.106	59.593	0.000	4.001
Error	0.107	60	0.002			
Total	0.943	63				

Figure 2. Mean plot of decision errors for three TDA models.

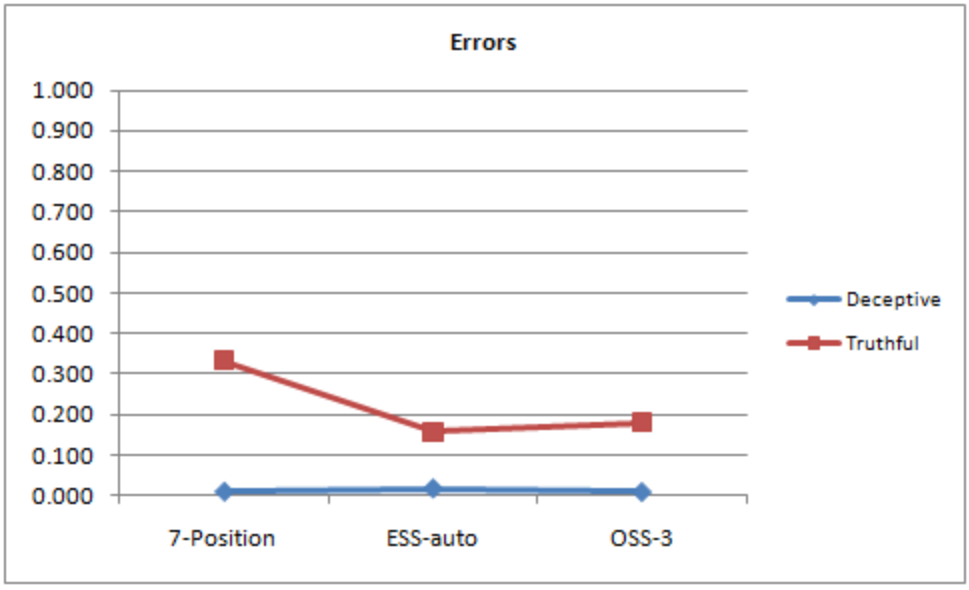
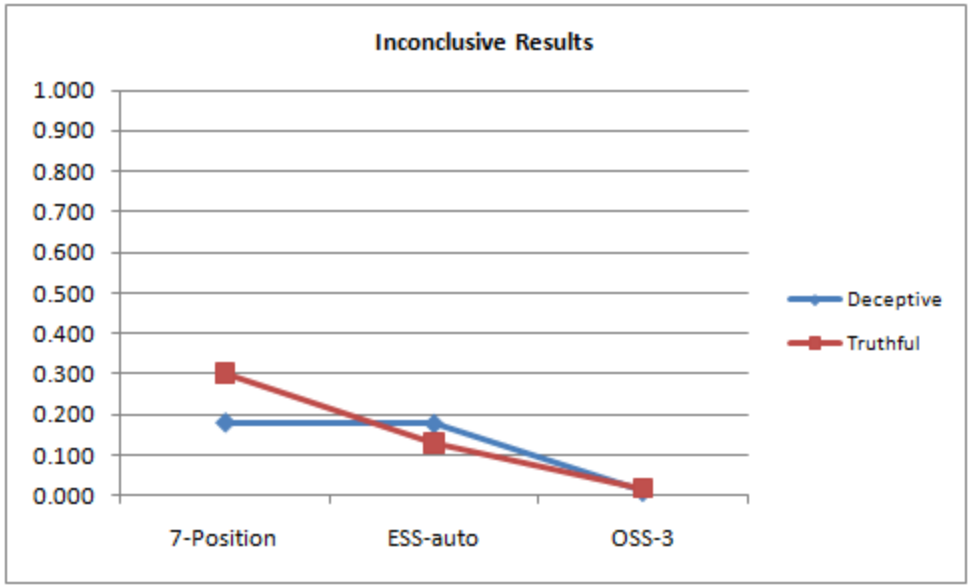


Table 6 shows the two-way ANOVA summaries for criterion status x scoring model for inconclusives. Figure 3 shows the mean plots for inconclusives for the three TDA models.

Table 6. ANOVA summary for inconclusives: criterion status x scoring model.

Source	SS	df	MS	F	p	F crit .05
Model	0.582	1	0.026	12.758	0.001	4.001
Status	0.013	1	0.000	0.185	0.669	4.001
Interaction	0.082	1	0.082	39.409	0.000	4.001
Error	0.124	60	0.002			
Total	0.676	63				

Figure 3. Mean plot of inconclusive results for three TDA models.



Discussion

The nearly significant main effect for question in the three-way ANOVA (Table 1) is not surprising because polygraph questions in single-issue multiple-facet examinations are intended to represent distinct behavioral concerns. Results of the two-way shown in Table 2 suggest that the three participants approached the scoring task in a similar manner for the truthful and deceptive cases.

There were significant main effects for case status, along with an interaction of case status x scoring model for decision accuracy, as shown in Table 4 and Figure 1. All scoring models were more effective at correctly identifying deceptive cases than truthful cases with the USAF-MGQT cases, and the 7-position method was less effective with truthful cases than the automated ESS and OSS-3 algorithm. Table 5 and Figure 2 illustrate there was a significant main effect

for case status along with a significant interaction of case status x scoring model for errors. Consistent with the general trend in the literature, all scoring models produced more false-positive results for case status x scoring model. The 7-position decision model produced more false-positive errors than the other modes. The reasons for this cannot be determined, but may be related to the use of sub-optimal cutscores. There was a significant main effect for scoring model, along with a significant interaction of scoring model x case status for inconclusive results. The 7-position scoring model produced more inconclusives than the automated ESS truthful cases, and the OSS-3 algorithm produced fewer inconclusives with both truthful and deceptive cases.

The cohort of three experienced Iraqi scorers, using the 7-position TDA model, produced an unweighted mean decision accuracy level without inconclusives 75.4% (95% CI = 66.9% to 83.8%). Mean inconclusives for the human scorers was 24.1% (15.8% to 32.3%). Although this inconclusive rate would appear to be high at first glance, it is well within the realm of mathematical estimates of inconclusive rates when considering the decision rules applied to the number of distinct subtotal scores within each examination.

An automated model of the ESS produced an unweighted decision accuracy rate of 89.5% (83.5% to 95.9%) along with 15.4% inconclusive results (8.4% to 22.3%). The OSS-3 computer algorithm scored the sample with an unweighted mean accuracy level of 90.2% (84.6% to 95.8%).[3] It is unclear whether the increase in criterion accuracy for ESS results was due to the structure of the scoring model or to automation. This should be the focus of future research.

Statistically significant differences exist among the TDA methods, along with significant differences in how the TDA models handle the truthful and deceptive cases. There were statistically significant main effects and interaction for decision accuracy without inconclusives, as shown in Table 4, for the results of the 7-position, automated ESS, and OSS-3 scoring models. Figure 1 shows that the mean decision accuracy without inconclusives for the 7-position model was weaker than that for the ESS and OSS-3. Although the exact cause cannot be known, it is possible that the difference is due to weaker specificity of the 7-position model resulting from the use of suboptimal cutscores, not based on normative data that have not been investigated for their statistical significance. There was a significant main and interaction effect for errors, shown in Table 5, for the three decision models resulting from significantly higher rate of false-positive errors for the 7-position model along with a fewer false-negative than false-positive errors for all three scoring models. In addition, there was a significant main effect and interaction for inconclusives, shown in Table 6, for the scoring models. Truthful cases produced a greater number of inconclusive results than deceptive cases, and the OSS-3 algorithm produced fewer inclusive results than the other models for both deceptive and truthful cases.

Limitations of the present study include the small cohort of scorers and small sample size and little access confirmation criteria for case status. It is possible that different decision rules and statistically optimized cutscores might produce important and desirable differences in some dimensions of criterion accuracy. Test sensitivity, inconclusives, and false-positive error rates may all be influenced by the spot-score decision rule, which requires a case to be classified as deceptive if any subtotal score result is significant for deception, and also requires that case is classified as inconclusive, if none of the subtotals is significant for deception but any of the subtotals is not significant for truthfulness. Optimized decision rules and cutscores for the USAF-MGQT should be the focus of future

[3] OSS-3 results using the two-stage decision rules (Senter 2002, 2003; Senter & Dollins, 2008a, 2008b), not shown in Table 3, produced zero errors and zero inconclusives. The authors caution against any expectation that any TDA protocol will produce perfect accuracy and zero inconclusives.

research. Previous studies have shown the Grand Total Rule (GTR) to provide significant improvements over the SSR in several dimensions of criterion validity (see Handler et al., 2010; Nelson et al., in press; Nelson et al., 2008). Other studies have shown that strategic use of combinations of the GTR and SSR can also optimize some dimensional aspects of criterion accuracy (Senter, 2003; Senter & Dollins, 2003; Senter & Dollins, 2008a; 2008b).

In summary, these data support the validity of the hypothesis that the USAF-MGQT can provide a high level of criterion accuracy, and can differentiate truthful from deceptive examinees at rates that are statistically significantly greater than chance ($p < .01$) using the 7-position, ESS and OSS-3 TDA models. Further research is warranted regarding decision rules, normative data, and statistically optimal cutscores for manually scoring the USAF-MGQT. Continued interest in the ESS and 7-position models is also recommended, in both research and field practice settings. Finally, continued interest is warranted in the USAF-MGQT as a field testing protocol.

References

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

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.

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.

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.

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

Kircher, J. & Raskin, D. (2002). Computer methods for the psychophysiological detection of deception. In Murray Kleiner (Ed.), *Handbook of Polygraph Testing: Academic Press*.

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.

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

Nelson, R. & Krapohl, D. J. (Manuscript accepted). 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., Handler, M., Shaw, P., Gougler, M., Blalock, B., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40, 67-78.

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011). Reliability of the Empirical Scoring System with expert examiners. *Polygraph*, 40, 131-139.

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

Nelson, R., Krapohl, D. J. & 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.

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.