Extended Analysis of Senter, Waller and Krapohl's USAF MGQT Examination Data with the Empirical Scoring System and the Objective Scoring System, version 3

Raymond Nelson and Benjamin Blalock

Abstract

Using archival data from a previous laboratory study, the authors provide an extended analysis of validation data for the USAF MGQT format for PDD testing. Multi-variate analysis found no significant differences between total and subtotal scores of the laboratory sample and those from a sample of field investigation cases using the same technique. Criterion accuracy profile are shown for the ESS and OSS-3 TDA models, including mean, standard deviations, and statistical confidence intervals for decision accuracy with and without inconclusive results, errors for deceptive and truthful cases, and inconclusive rates for deceptive and truthful cases.

Introduction

The United States Air Force Modified General Question Technique (USAF-MGQT) (DoDPI, 2006) is a modern variant of the family of Comparison Question Techniques (CQT) that have come into existence as modifications of the General Question Technique (Reid, 1947) and the Zone Comparison Technique (Backster, 1963). The USAF-MGQT conforms to generally accepted valid principles for psychophysiological detection of deception (PDD) test construction (Krapohl, 2006), and exists in two closely related versions for which there is no published evidence and no compelling hypothesis suggesting that the differences are substantive or would have any effect on criterion accuracy. The USAF-MGQT is often used in multiple-facet investigative contexts, to investigate multiple roles or levels of involvement in a single known incident or allegation, and multiple-issue screening contexts, for which it is conceivable that an examinee may be involved in one or more distinct behavioral concerns while completely uninvolved in others.

Senter, Waller & Krapohl, (2008), using a mock roadside-bombing scenario, reported a mean 7-position blind-scoring criterion accuracy level of .849, excluding inconclusive results. The present study is an investigation into the use of the USAF-MGQT when scored via an evidence-based scoring protocol, the Empirical Scoring System (ESS) (Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson & Hicks, 2011; Krapohl, 2010; Nelson, Blalock, Oelrich & Cushman, 2011a; Nelson et al., 2011b; Nelson, Krapohl & Handler, 2008) and the Objective Scoring System, version 3 (OSS-3) computer algorithm (Nelson, Krapohl & Handler, 2008).

Method

Archival data were obtained from the Senter et al. (2008) laboratory study (N = 69), and were scored via an automated version of the ESS and with the OSS-3 computer algorithm. The Senter et al. (2008) article provides a complete account of the methodology of their study, and only a brief description is reported here. The sample consisted of confirmed laboratory examinations, for which 36 of the examinations were conducted on programmed truthful persons, while 33 examinations were conducted on programmed deceptive persons. All examinations were conducted with the USAF-MGQT. All of the examinations consisted of two investigation target questions, one pertaining to direct involvement and the other pertaining to secondary or indirect involvement in the mock incident, along with three comparison questions.

We scored the data using an automated version of the ESS TDA model, including automated measurement of physiological features, automated transformation of the integer point scores and automated execution of decision rules. ESS scores were assigned by comparing each investigation target question to the preceding or subsequent comparison question that elicits the greater reaction for each component sensor. Decision cutscores were set at alpha = .05 for deceptive classifications and alpha = .1 for truthful classifications.

Cutscores corresponding to these alpha levels were -3 and +1, meaning that any subtotal score of -3 or lower would be statistically significant for deception ($p < .05$), while test results in which all subtotal scores are +1 or greater would be statistically significant for truth-telling ($p < .1$). Bonferonni correction to the alpha cutscore for deceptive classifications is not used with PDD examinations for which it is assumed the investigate target questions are independent. However, an inverse of the Sidak correction for independent issues is used to correct for the deflation of alpha that occurs when calculating the normative probability that an examinee would produce a statistically significant truthful result to all investigation targets while lying to one or more of the independent issues.

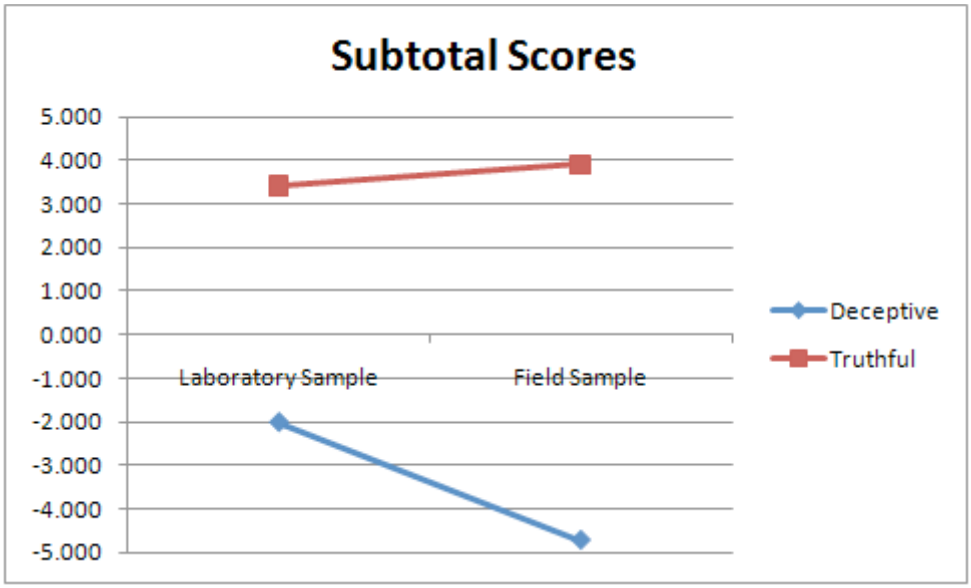
Sample data were also scored using the OSS-3 computer algorithm with alpha = .05 for deceptive classifications and alpha = .1 for truthful classifications. Whereas manual and automated scoring procedures for the ESS involve the assignment of integer scores while comparing the strengths of reaction for each relevant question to the stronger of neighboring comparison questions, the OSS-3 algorithm compares each relevant question to the average of all comparison questions, for each component sensor, to assign numerical scores in the form of standardized logged ratios. The decision rule for the automated ESS model and OSS-3 was the spot-score-rule (SSR) (Department of Defense, 2006), and is identical to that used when manually scoring multiple-facet and multiple-issue examinations.

Results

All statistical results were evaluated with a level of significance set at alpha = .05.

ESS Scores. The mean subtotal score for deceptive cases was -2.00 (SD = 5.03), and the mean subtotal for truthful cases was 3.42 (SD = 3.47). A two-way unbalanced ANOVA (sample x case status) was used to compare the distribution of ESS total scores with those from a study based on a sample of confirmed field cases (N = 22) using the USAF-MGQT (Nelson, Blalock & Handler, 2011). There was a significant interaction effect for sample distributions and subtotal scores ($F_{1,87} = 541.557$, $p < .001$). Figure 1 shows that the field sample (N = 22) produce subtotal scores of greater absolute value than the laboratory sample (N = 66), when subjected to automated measurement. Post-hoc unbalanced one-way ANOVAs showed that differences where not statistically significantly for in the subtotal scores. These results suggest that the sampling distributions of deceptive and truthful laboratory scores do not differ significantly even though the field sample scores were further from zero.

Figure 1. Mean subtotal scores for laboratory and field USAF MGQT samples.



Criterion Validity. Table 1 shows the mean percentages, standard deviations, and statistical confidence intervals for a dimensional profile of criterion accuracy for the laboratory sample of USAF-MGQT examinations, including: sensitivity, specificity, inclusive results for deceptive and truthful cases, false-positive and false-negative errors, positive predictive value, negative predictive value, percent of correct decisions for the deceptive and truthful cases, and the unweighted means of the percent correct and inconclusive results for deceptive and truthful cases. Data are shown for ESS scores and for the OSS-3 computer algorithm.

Table 1. Criterion Accuracy Profiles for ESS and OSS-3 Scores of USAF-MGQT Examinations (N = 66).

Mean, Standard Deviations and 95% Confidence Intervals for Manual ESS, Automated ESS and OSS-3 Algorithm SD and CI		
	Automated ESS	OSS-3
Unweighted Accuracy	.838 (.040) {.758 to .918}	.856 (.038) {.78 to .931}
Unweighted Inc	.153 (.033) {.087 to .220}	.154 (.036) {.082 to .226}
Sensitivity	.511 (.071) {.371 to .65}	.667 (.068) {.532 to .802}
Specificity	.940 (.034) {.872 to .999}	.782 (.058) {.667 to .898}
FN Error	.211 (.059) {.094 to .329}	.148 (.050) {.050 to .246}
FP Error	.029 (.024) {.001 to .078}	.092 (.042) {.009 to .175}
D Inc	.277 (.064) {.151 to .402}	.184 (.055) {.075 to .292}
T Inc	.030 (.024) {.001 to .078}	.125 (.045) {.035 to .215}
PPV	.945 (.045) {.857 to .999}	.879 (.054) {.773 to .984}
NPV	.815 (.052) {.712 to .918}	.840 (.053) {.736 to .944}
D Correct	.707 (.077) {.554 to .859}	.817 (.061) {.697 to .937}
T Correct	.969 (.025) {.919 to .999}	.894 (.048) {.800 to .988}

Discussion

These results support the hypothesis that both ESS and OSS-3 scores of USAF-MGQT examinations can differentiate confirmed truthful from confirmed deceptive cases at rates that are significantly greater than chance. Of course, generalization of study results to field settings is realistic only when the examination is conducted competently, and only as long as the examination data are of satisfactory quality, sufficiently free of uninterpretable artifacts.

The primary limitation of the present study involves the small cohort of scorers and the relatively small sample size. Despite these limitations, we argue that the cohort of scorers including one experienced examiner and two inexperienced trainees might be expected to represent and generalize to real-world field settings more effectively than study results based on the scores of a single expert scorer. Results from the multivariate analysis found no significant differences with the truthful and deception distributions of laboratory and field sample scores, though the field sample did produce scores of stronger absolute magnitude (i.e., further from zero).

Past studies have not shown support for the hypothesis that multi-facet examination can effectively differentiate deception from truth at the level of the individual question (Barland, Honts & Barger, 1989a; Barland, Honts & Barger, 1989b; Podlesny & Truslow, 1993), and this study could not attempt to differentiate truth telling from deception at the level of the individual question. Instead test results for the sample cases were classified at the level of the test as a whole, despite the fact that decision rules involved the subtotals for individual test questions. As is often the case, additional research is needed in this area.

These results suggest continued interest in the USAF-MGQT format and continued interest in the ESS and OSS-3 TDA models.

References

Backster, C. (1963). Standardized polygraph notepack and technique guide: Backster zone comparison technique. Cleve Backster: New York.

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

Barland, G. H., Honts, C. R. & Barger, S.D. (1989b). Studies of the accuracy of security screening polygraph examinations. DTIC AD Number A304654. Department of Defense Polygraph Institute.

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 (2006). Federal Psychophysiological Detection of Deception Examiner Handbook. Reprinted in Polygraph, 40(1), 2-66.

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2011). Empirical Scoring System: A Cross-cultural Replication and Extension Study of Manual Scoring and Decision Policies. Polygraph, 39, 200-215.

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

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

Nelson, R., Handler M., Morgan C., & O'Burke P. J. (2012). Short Report: Criterion Validity of the United States Air Force Modified General Question Technique and Iraqi Scorers. Polygraph. 41(1), 18-28.

Nelson, R., Blalock, B., Oelrich, M. & Cushman, B. (2011). Reliability of the Empirical Scoring System with Expert Examiners. Polygraph, 40(3), 131-139.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. Polygraph, 40(2) 67-78.

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.

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.

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

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