Criterion Validity of the Empirical Scoring System and the Objective Scoring System, version 3 with the USAF Modified General Question Technique

Raymond Nelson, Benjamin Blalock and Mark Handler

Abstract

Using archival data, this study investigated the criterion accuracy of ESS scores with the USAF-MGQT format that is commonly used for multiple-facet diagnostic PDD testing and multiple-issue screening exams. Two inexperienced and one experienced examiners completed blind scoring tasks of an archival sample of confirmed field cases from the Department of Defense confirmed case archive. Sample cases were also scored with an automated version of the ESS, and the OSS-3 computer algorithm. Overall unweighted decision accuracy for manual ESS scores was 88.2%, with 18.3% inconclusives. Decision accuracy for the automated ESS model was 89.7% with 15.4% inconclusives. The OSS-3 computer algorithm produced 90.2% correct decisions with 1.0% inconclusives. Pearson correlations were strong for the scores of the study participants, (r = .931) between the manual ESS and automated ESS scores (r = .938). Pair-wise decision agreement was 80.4% including inconclusives, and perfect when inconclusives were excluded. Pair-wise agreement was perfect for the ESS and OSS-3 models, for this small-scale study. Multivariate analysis showed no significant main effects and no significant interaction effects between the mean total scores of manual and automated ESS models. The authors recommend continued interest in the USAF-MGQT format, the ESS in both manual and automated models, and the OSS-3 algorithm.

Introduction

The Modified General Question Technique (MGQT) is a family of related 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 United States Air Force Modified General Question Technique (USAF-MGQT) (Department of Defense, 2006) is a modern version of the MGQT that is regarded as compliant with generally accepted valid principles for psychophysiological detection of deception (PDD) test construction (Krapohl, 2006). The USAF-MGQT has become widely used in investigative multiple-facet contexts to investigate multiple roles or levels of involvement in a single known incident or allegation, and multi-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 effort to extend our present knowledge-base regarding the criterion accuracy 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, 2010; Krapohl, 2010; Nelson, Blalock, Oelrich & Cushman, in press; Nelson et al., 2011; Nelson, Krapohl & Handler, 2008) and the Objective Scoring System, version 3 (OSS-3) computer algorithm (Nelson et al., 2008).

Method

Blind evaluations were obtained from three scorers, including one American Polygraph Association (APA) certified primary instructor, trained at the Department of Defense, and two inexperienced trainees at a polygraph school accredited by the APA.

A matched random sample (N = 22) of USAF-MGQT examinations was selected from the confirmed case archive at the Department of Defense. Eleven confirmed truthful examinations were found that met the selection criteria of healthy adult criminal suspects, reportedly not taking psychotropic medications, who had examinations conducted using the USAF-MGQT consisting of three or more test charts. Eleven matching confirmed deceptive examinations were randomly selected from the confirmed case archive. Two versions of the USAF-MGQT exist (Department of Defense, 2006), version 1 and version 2 and there is no published evidence and no compelling hypothesis suggesting that the differences are substantive or would have any effect on criterion accuracy. Both versions of the USAF-MGQT were included in the sample, and the sample cases consisted of 2, 3 and 4 investigation targets as permitted by the procedures (Department of Defense, 2006). All examinations were subject to quality assurance review in the field, and examination results from the original examiners were not 100 percent accurate.

Data were scored using an automated version of the ESS TDA model, including automated measurement of physiological features, automated transformation of the integer point scores using procedures identical to those used when manual scoring with the ESS, and automated execution of decision rules using alpha = .05 for deceptive classifications and alpha = .1 for truthful classifications. These data were also scored using the OSS-3 computer algorithm with alpha = .05 for deceptive classifications and alpha = .1 for truthful classifications. The decision rule for the automated ESS model and OSS-3 was the spot-score-rule (SSR) (Light, 1999), identical to that used when manually scoring multi-facet and multi-issue examinations.

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 in which it is assumed the investigation 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.

Means, standard deviations, and statistical confidence intervals were calculated for a dimensional profile of criterion accuracy, 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. Distributions of scores were compared to scores obtained from another study (Nelson & Handler, in review) using unbalanced multivariate ANOVAs.

Results

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

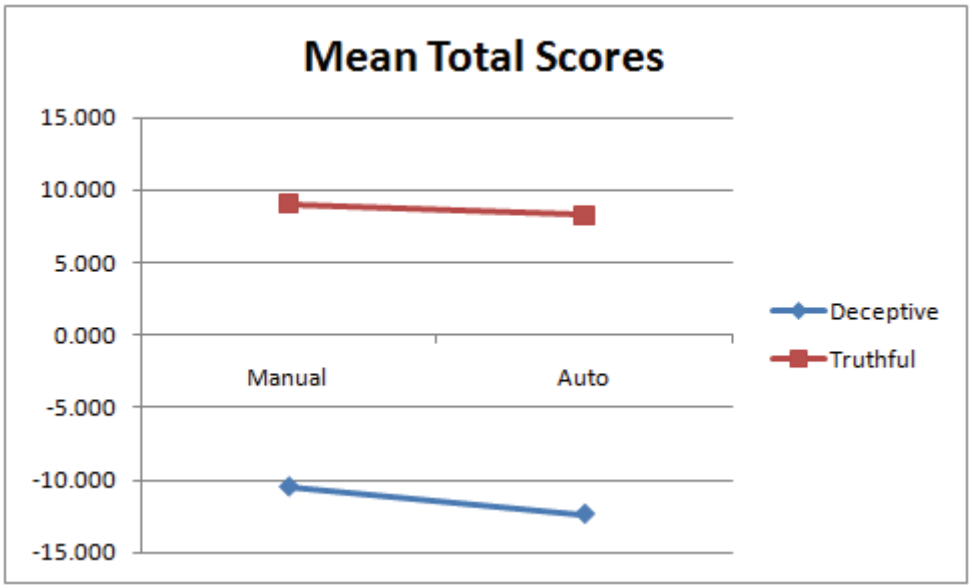
Manual ESS and Automated ESS Scores

The mean total ESS manual score for deceptive cases was -10.50 (SD = 10.04) and the mean total ESS manual score for truthful cases was 9.03 (SD = 9.97). The mean total automated ESS score for deceptive cases was -12.36 (SD = 11.99) and the mean total automated ESS score for truthful cases was 8.27 (SD = 8.45). Table 1 shows the results of a two-way ANOVA using the absolute (unsigned) mean totals, scoring method x case status, for mean total scores. Mean total scores are shown in Figure 1. Neither the main effects nor interaction were significant, indicating a good correspondence between the manual and automated ESS scores.

Table 1. Two-way ANOVA summary, scoring method x status for mean total.

Source	SS	df	MS	F	p	F crit .05
Method	3.370	1	0.153	0.001	0.970	4.085
Status	85.058	1	3.866	0.037	0.848	4.085
Interaction	18.882	1	18.882	0.182	0.672	4.085
Error	4152.097	40	103.802			
Total	107.310	43				

Figure 1. Means of confirmed deceptive and confirmed truthful total scores.



It is assumed that all subtotals of confirmed truthful cases are truthful subtotals. However, because multi-facet and multi-issue examinations are conducted with the assumption of independence among the question stimuli, in which it is assumed that a deceptive subject may not be lying to each question, subtotal scores for deceptive examinations may include both deceptive and non-deceptive subtotal scores. The mean ESS subtotal score for truthful cases was 4.53 (SD = 5.18). It was not possible to calculate a subtotal score for only those confirmed deceptive questions because confirmation was not available at the level of the individual questions. Instead, a grand mean subtotal score was calculated from the mean of all subtotals within the confirmed deceptive cases regardless of the assumption of independence and the possibility that examinees may have been truthful to some of the questions in the confirmed deceptive cases. The mean subtotal score for deceptive cases was -3.85 (SD = 4.73). Figure 2 shows the mean plots. Table 2 shows that neither the main effect nor the interaction of the unsigned mean scores was significant, indicating that manual and automated ESS subtotal scores approximate each other reasonably well.

Figure 2. Means of confirmed deceptive and confirmed truthful subtotal scores.

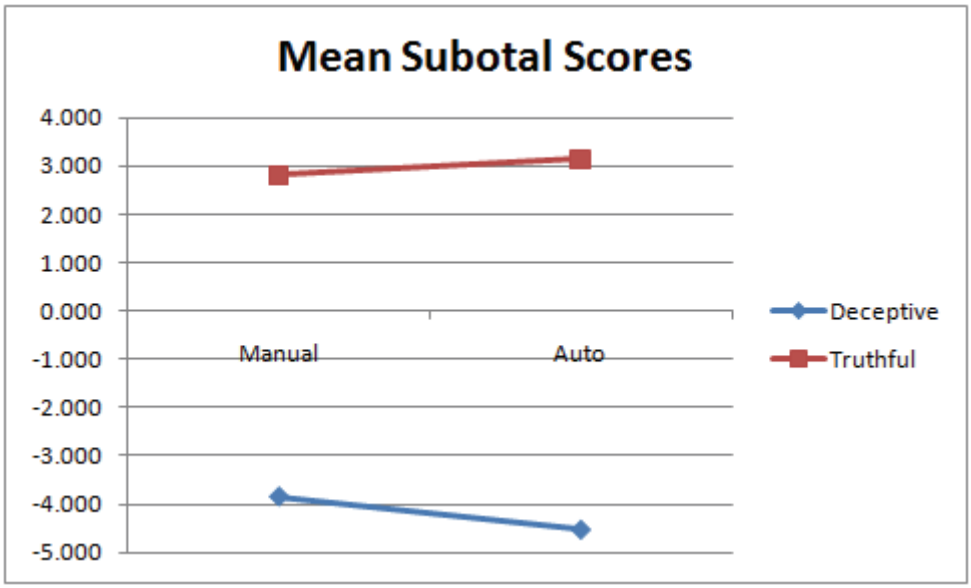


Table 2. Two-way ANOVA summary, scoring method x status, for subtotals.

Source	SS	df	MS	F	p	F crit .05
Method	2.758	1	0.125	0.006	0.937	4.085
Status	16.286	1	0.740	0.037	0.847	4.085
Interaction	0.351	1	0.351	0.018	0.895	4.085
Error	790.058	40	19.751			
Total	19.395	43				

A 2 x 3 x 4 ANOVA comparison (status x scorer, x question) of the subtotal scores is shown in Table 3, and resulted in no significant interaction between the three AVOVA factors, and no significant main effects for scorer or case status. The main effect for questions was approaching a statistically significant level (p = .065). This suggests that a main effect for individual questions might be found in a study of larger scale and greater statistical power.

Reliability

Pair-wise analysis of the manual ESS and automated ESS scores showed a strong correlation between the total scores of the study participants (r = .931, SEM = .038), and a strong pair-wise correlation between the total scores of the study participants and the automated ESS total score (r = .938 SEM = .193). The pair-wise proportion of decision agreement including inconclusive results was .804 (SE = .067). There was perfect correspondence between the decisions of the study participants when inconclusive results were excluded. Pair-wise decision agreement with the automated ESS model was .962 (SE = .028). There was perfect correspondence between the results of the automated ESS and OSS-3 algorithm results. We caution that perfect agreement between the algorithm and automated ESS should not be expected in a study of larger scale.

Table 3. ANOVA summary, status x scorer, x question for subtotals.

Source	SS	df	MS	F	p	F crit .05
Status	23.036	1	23.036	1.187	0.278	3.900
Scorer	15.898	2	7.949	0.409	0.665	3.052
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				

Criterion Validity

Table 4 shows the dimensional profile of criterion accuracy, including mean percentages, standard deviations, and statistical confidence intervals for ESS scores from the study participants, along with the criterion accuracy profile from an automated ESS model and the OSS-3 algorithm (Nelson, Handler, O'Burke & Morgan, submitted).

Table 4. Criterion Accuracy Profiles for Manual ESS, Automated ESS and OSS-3 Algorithm Scores (N = 22).

Mean, Standard Deviations and 95% Confidence Intervals for Manual ESS, Automated ESS and OSS-3 Algorithm SD and CI			
	ESS	Automated ESS	OSS-3
Unweighted Accuracy	.882 (.034) {.815 to .950}	.897 (.031) {.835 to .959}	.902 (.028) {.847 to .958}
Unweighted Inc	.183 (.038) {.108 to .258}	.154 (.035) {.084 to .223}	.010 (.010) {.001 to .030}
Sensitivity	.831 (.051) {.730 to .931}	.803 (.055) {.694 to .912}	.980 (.019) {.941 to .999}
Specificity	.616 (.069) {.479 to .752}	.710 (.065) {.581 to .84}	.806 (.056) {.695 to .917}
FN Error	.010 (.014) {.001 to .039}	.018 (.019) {.001 to .056}	.009 (.013) {.001 to .035}
FP Error	.175 (.054) {.069 to .281}	.158 (.053) {.053 to .262}	.183 (.054) {.075 to .290}
D Inc	.158 (.050) {.059 to .257}	.177 (.055) {.069 to .286}	.010 (.014) {.001 to .038}
T Inc	.208 (.057) {.096 to .320}	.130 (.047) {.037 to .223}	.010 (.015) {.001 to .040}
PPV	.826 (.053) {.721 to .931}	.837 (.052) {.735 to .940}	.842 (.048) {.748 to .936}
NPV	.983 (.024) {.935 to .999}	.974 (.026) {.921 to .999}	.988 (.016) {.955 to 1.021}
D Correct	.987 (.017) {.952 to .999}	.977 (.023) {.932 to .999}	.990 (.013) {.963 to 1.017}
T Correct	.778 (.067) {.647 to .909}	.817 (.060) {.698 to .937}	.814 (.055) {.706 to .923}

Discussion

These results indicate that the USAF-MGQT can differentiate confirmed deceptive cases from confirmed truthful cases at rates that are significantly greater than chance, when scored via the ESS, in both manual and automated models, and via OSS-3 scoring. These results also indicate a strong correlation between the scores and results of manual and automated TDA models, and suggest that the results and scores of manual and automated ESS models approximate each other well. Future research might investigate the potential for further automation of the ESS model.

Of course, no manual or automated TDA method can be expected to accurately interpret the results of a test that has been conducted improperly or ineffectively. The accuracy and reliability of both manual and automated TDA models in field settings will be substantially influenced by the quality of data and the effectiveness with which the examination is conducted and the data are collected. Generalization of study results to field settings is realistic only when the

examination is conducted competently and the data are of satisfactory quality, and sufficiently free of uninterpretable artifacts.

Because of the lack of confirmation data for individual questions, this analysis did not investigate the criterion accuracy of the subtotal scores or individual questions. Other studies have failed to support the hypothesis that multi-facet examinations 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 should continue to be the subject of future research. At the present time, the evidence supports the position that people pass or fail the test as a whole, not the individual questions. In a practical sense this means that a significant response indicative of deception to any individual test questions requires that the entire test result is classified as deceptive. There is presently no support for the practice of interpreting or reporting a non-deceptive result to other test questions when one or more test questions results in a significant reaction indicative of deception. However, it is accepted field practice, when multi-facet and multi-issue examinations are interpreted via the Spot Score Rule (Light, 1999), to interpret and report the individual question to which the examinee responded significantly, and to interpret the test as a whole as showing significant reactions indicative of deception. When this happens, favorable conclusions regarding individual test questions that do not produce a significant result indicative of deception are unsupportable, and no opinion can be rendered regarding those individual questions.

Limitations of the present study include the small cohort of scorers and the small sample size. Additionally, it is unknown how the sample cases came to be included in the confirmed case archive, other than the availability of ground-truth confirmation data. No small-scale study, and no single study, can be regarded as a definitive description of test performance under widely varying field circumstances.

Another limitation of this study involves the unknown generalizability of research evidence pertaining to multi-facet and multi-issue examinations. In terms of decision theory, the operational concern is the assumed independence of the issues represented by the test stimuli. If independence is assumed for both multi-facet and multi-issue examination models, then there should be no hypothesized statistical difference between the two models. If it is assumed that the target issues described by multi-facet test stimulus questions are non-independent, then important differences are hypothesized and the application of decision rules intended for independent stimuli would be inappropriate, calling instead for decision rules designed for single-issue examinations for which the test stimuli are non-independent. This should be the focus of future research. For the present time, we surmise that the test stimuli for both multi-facet and multi-issue examinations are assumed to be independent, that the application of similar decision rules to both models is warranted, and that statistical and decision theoretic evidence can be cautiously generalized between multiple-facet and multiple-issue approaches to target selection and question formulation.

Limitations notwithstanding, these results do increase our present knowledge regarding criterion accuracy of the USAF-MGQT, and do suggest continued interest in this test format for multi-facet investigative and multiple-issue screening uses, along with 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. 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. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph*, 39, 200-215.

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

Krapohl, D. (2010). Short report: A Test of the ESS with two-question field cases. *Polygraph*, 39, 124-126.

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, 131-139.

Nelson, R. & Handler, M. (submitted). Criterion validity of the United States Air Force Modified General Question Technique and Iraqi scorers.

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