Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

VOLUME 49 **2020** **NUMBER 1**

Contents

Published Semi-annually

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

APA Officers for 2019 – 2020

Subscription information: *Polygraph* is published semi-annually by the American Polygraph Association. Editorial Address is Editor@polygraph.org. Subscription rates for 2020: One year $150.00. Change of address: APA National Office, P.O. Box 8037 Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

Instructions to Authors

Scope

The journal *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* publishes articles about the psychophysiological detection of deception, and related areas. Authors are invited to submit manuscripts of original research, literature reviews, legal briefs, theoretical papers, instructional pieces, case histories, book reviews, short reports, and similar works. Special topics will be considered on an individual basis. A minimum standard for acceptance is that the paper be of general interest to practitioners, instructors and researchers of polygraphy. From time to time there will be a call for papers on specific topics.

Manuscript Submission

Manuscripts must be in English, and may be submitted, along with a cover letter, on electronic media (MS Word). The cover letter should include a telephone number, and e-mail address. All manuscripts will be subject to a formal peer-review. Authors may submit their manuscripts as an e-mail attachment with the cover letter included in the body of the e-mail to:

Editor@polygraph.org

As a condition of publication, authors agree that all text, figures, or other content in the submitted manuscript is correctly cited, and that the work, all or in part, is not under consideration for publication elsewhere. Authors also agree to give reasonable access to their data to APA members upon written request.

Manuscript Organization and Style

All manuscripts must be complete, balanced, and accurate. Authors should follow guidelines in the *Publications Manual of the American Psychological Association*. The manual can be found in most public and university libraries, or it can be ordered from: American Psychological Association Publications, 1200 17th Street, N.W., Washington, DC 20036, USA. Writers may exercise some freedom of style, but they will be held to a standard of clarity, organization, and accuracy. Authors are responsible for assuring their work includes correct citations. Consistent with the ethical standards of the discipline, the American Polygraph Association considers quotation of another's work without proper citation a grievous offense. The standard for nomenclature shall be the *Terminology Reference for the Science of Psycho-physiological Detection of Deception* (2012) which is available from the national office of the American Polygraph Association. Legal case citations should follow the *West* system.

Manuscript Review

An Associate Editor will handle papers, and the author may, at the discretion of the Associate Editor, communicate directly with him or her. For all submissions, every effort will be made to provide the author a review within 4 weeks of receipt of manuscript. Articles submitted for publication are evaluated according to several criteria including significance of the contribution to the polygraph field, clarity, accuracy, and consistency.

Copyright

Authors submitting a paper to the American Polygraph Association (APA) do so with the understanding that the copyright for the paper will be assigned to the American Polygraph Association if the paper is accepted for publication. The APA, however, will not put any limitation on the personal freedom of the author(s) to use material contained in the paper in other works, and request for republication will be granted if the senior author approves.

Monte Carlo Study of Multiple Issue Polygraph Techniques with Two, Three, and Four Questions

Raymond Nelson[1], Mark Handler[2], Stuart Senter[3]

Abstract

Monte Carlo methods and multivariate analysis of variance (ANOVA) were used to study criterion accuracy of multiple-issue PDD examinations with two, three, and four relevant questions (RQs) – such as those conducted using the USAF MGQT – when scored with the seven-position, three-position, and Empirical Scoring System (ESS) methods. Test sensitivity to deception exceeded chance (.5) for all scoring conditions with two, three and four RQs. Some differences were observed for different treatments, with inconclusive rates decreasing with the number of RQs for criterion deceptive cases and increasing with the number of RQs for criterion truthful cases. Test specificity to truth-telling was significantly greater than chance only for the 2 RQ model with ESS scores. No significant differences were found in false-positive or false-negative rates for seven-position, three-position or ESS scores with two, three or four RQs. However, the likelihood of testing error increased with the number of RQs for criterion truthful cases while decreasing for criterion deceptive cases. Excluding inconclusive results, the unweighted average decision accuracy for criterion deceptive and criterion truthful cases exceeded chance, and no significant differences were observed in unweighted accuracy for the three scoring methods with two, three, and four RQs. It was not possible in this study to determine whether this difference was due to the scoring method or to the use of a norm-referenced cutscores and multiplicity correction for ESS cutscores compared to traditional cutscores.

Introduction[4]

Multiple-issue polygraphs are commonly used in polygraph screening – in the absence of a known allegation or incident, using two, three, and four relevant questions (RQs). The United States Air Force Modified General Question Test (USAF MGQT) (Department of Defense, 2006; Nelson, Blalock & Handler, 2011; Nelson, Handler, Morgan & O'Burke, 2012; Senter, Waller & Krapohl, 2008) – for which two versions exist in field practice – is an example of a polygraph test that can be used with two, three and four RQs. Other multiple-issue polygraph formats also exist. Multiple issue polygraphs can be thought of as a contemporary variant of the of the comparison question technique described by Reid (1947) and Summers (1939). The defining characteristic of multiple-issue polygraphs, including the USAF MGQT and other formats – is that the relevant questions (RQs) are assumed to be independent[5].

[1] Raymond Nelson is research specialist with the Lafayette Instrument Company (LIC) and an elected member of the APA Board of Directors. The views expressed in this work are those of the author and not the LIC or the APA. Mr. Nelson is a psychotherapist, polygraph field examiner, developer of the OSS-3 scoring algorithm, and is the author of several publication on various polygraph topics. For information contact raymond.nelson@gmail.com.

[2] Mark Handler is an experienced police examiner and polygraph researcher who helped develop the Objective Scoring System, version 3 and the Empirical Scoring System. His email address is polygraphmark@gmail.com.

[3] Stuart Senter is employed at the National Center for Credibility Assessment (NCCA). The views expressed in this work do not reflect those of the NCCA.

[4] The authors are grateful to Joseph Stainbeck IV, who reviewed an earlier version of this manuscript.

The authors found no published studies that describe criterion accuracy of this technique while varying or comparing the numbers of RQs. The present study is an exploratory effort to extend our knowledge base regarding differences in criterion accuracy that may be observed as a function of the number of RQs. The hypothesis was that the multiple-issue polygraphs, with two, three, and four RQs can achieve classification accuracy rates that are greater than chance (50 %) when evaluated with the 7-position, 3-position and ESS methods. This can also be stated in terms of testing errors: wherein the hypothesis is that multiple-issue polygraphs with two, three and four RQs can achieve false-positive and false-negative error rates that are significantly less than chance.

Method

Monte Carlo methods were used to calculate confidence intervals for criterion accuracy of multiple-issue polygraph examinations with two, three, and four RQs, including test sensitivity, specificity, false-positive and false negative error rates, along with unweighted decision accuracy and inconclusive rates. Data were scored and interpreted using the seven-position and three-position test data analysis (TDA) methods (Department of Defense, 2006; Harwell, 2000; Krapohl, 1998; Van Herk, 1990) and the Empirical Scoring System (ESS; Blalock, Cushman & Nelson, 2009; Handler, Nelson, Goodson, & Hicks, 2011; Krapohl, 2010; Nelson, Blalock & Handler, 2011; Nelson, Blalock, Oelrich & Cushman. 2011; Nelson & Handler, 2010; Nelson et al., 2011; Nelson & Krapohl, 2011; Nelson, Krapohl, & Handler, 2008). Monte Carlo models were constructed for the three different scoring methods, and each of these was evaluated using two, three, and four RQs. In addition to these nine models, three addition Monte Carlo models were defined to evaluate the effectiveness of the seven-position, three-position and ESS scoring methods while randomly varying the number of RQs.

The Monte Carlo space consisted of N = 100 simulated multiple-issue examinations, for which the criterion status of each RQ was set independently by comparing a random number to a fixed base rate. Separate Monte Carlo models were created for examinations with two, three and four RQs, and the number of RQs was uniform within each Monte Carlo space. Each Monte Carlo space was simulated 10,000 times to create three Monte Carlo distributions of results – for two, three, and four RQs – that could be studied for decision accuracy, errors and inconclusive results. Each Monte Carlo distribution would be evaluated with the seven-position, three-position, and ESS scoring methods.

Subtotal scores were simulated by standardizing random numbers to seeding parameters that were the means and standard deviations of the subtotal scores provided by the participants in the Krapohl and Cushman (2006) study after transforming the seven-position subtotal scores of the guilty and innocent cases to three-position scores and then to ESS scores[6]. Krapohl (2010) and Robertson (2012) showed that transformed ESS scores are capable of extracting similar physiological data as compared to 7-position and 3-position manual scores.

[5] Independence, in scientific testing, refers to the assumption that the criterion variance or external state of each individual test stimulus is not affected by and does not affect the criterion variance of other test stimuli. Criterion variance is related to but distinct from response variance. As a practical matter, both multi-facet and multi-issue examinations are assumed to be composed of independent stimuli, and both types are therefore scored and interpreted using question sub-total scores, though the independence of sub-total scores of multi-facet examinations has not been supported by previous studies.

[6] The Federal ZCT cases in Krapohl and Cushman (2006) consisted of three relevant questions that refer to the examinee's involvement a single known allegation or incident. Traditional usage of the Federal ZCT included two relevant questions that describe the examinee's behavior, while the third relevant question is used to describe the examinee's knowledge of incriminating details of the incident or allegation. However, all relevant questions are interpreted uniformly or non-independently when using the Federal ZCT, and no extant publications have described effect sizes for the independent treatment or interpretation of Federal ZCT questions. One of the Marin sample cases included only two relevant questions. A total of 299 subtotal scores, regarded as uniformly innocent or guilty, were used for the Monte Carlo seeds of the multiple-issues cases in the Monte Carlo model. Whereas the traditional usage of the Federal ZCT involves both the grand total and subtotal scores, only the subtotal information was used for seeding parameters for the present Monte Carlo study.

Table 1 shows the input seed parameters, the subtotal means and standard deviations, for the Monte Carlo sample scores. The design of this Monte Carlo space meant that the criterion state was random, independent, and known for each RQ in the Monte Carlo space, and the number of RQs could be manipulated to evaluate the effect sizes.

Table 1. Subtotal means and standard deviations.

	Deceptive Mean	Deceptive SD	Truthful Mean	Truthful SD
7-position	-2.827	4.504	3.556	3.766
3-position	-1.886	3.161	2.427	2.557
ESS	-3.031	4.535	3.265	3.661

For each Monte Carlo space, the base rate for deception and truth-telling for individual RQs was calculated using the inverse of the Šidák correction (Abdi, 2007; Šidák, 1967) for multiple statistical comparisons under a condition of independent variance (Abdi, 2007). Base rates for individual questions were as follows; two RQs = .293, three RQs = .206, and four RQs = .159. For each RQ in each case a random uniform number was compared to the base-rate, and the criterion state was set to truthful if the base-rate was less than the random number. This ensured a base rate for each Monte Carlo distribution that converged at .5 while randomly setting the criterion state for each RQ and while allowing variation in the observed incidence rate of deception and truth-telling for each iteration of the cases in the Monte Carlo Space. For each exam in each of the Monte Carlo spaces the criterion state of each case in the Monte Carlo space was set to deceptive if the criterion state of one or more of the RQs was deceptive. The criterion states of the cases were set to truthful if the criterion status of all RQs was truthful.

Traditional cutscores were used for the for the seven-position and three-position TDA methods: test results were classified as deceptive when any subtotal score was -3 or lower, and test results were classified as truthful when all subtotal scores were greater than or equal to +3. It can be noted that these traditional cutscores are not based on normative data, but were derived through experience and heuristic study and are similar to cutscores that are derived from statistical procedures (Nelson, *et al.*, 2011; Nelson, 2017; Nelson & Rider, 2018).

Cutscores for ESS scores of USAF MGQT exams are based on statistical reference distributions for individual subtotal scores of guilty and innocent persons (Nelson *et al.*, 2011, Nelson, 2017, Nelson & Rider, 2018). The main difference between ESS cutscores and traditional cutscores is that ESS cutscores are determined using a Šidák correction to account for the multiplicity effects that are expected as a result of the procedural requirement that all subtotal scores are statistically significant for truth-telling in order to classify a test result as truthful. ESS cutscores were -3 and +1, meaning that test results would be classified as deceptive if when any subtotal score was -3 or lower and would be classified as truthful when all subtotal scores are +1 or greater.

All cases in the Monte Carlo space were evaluated using the subtotal score rule (SSR; Department of Defense, 2006a, 2006b; Capps & Ansley 1992; Senter Waller & Krapohl; 2008) for which the overall test result is inherited from the lowest question/subtotal score – whereas the question level results of event-specific diagnostic exams are inherited from the overall test result [See Nelson, Blalock & Handler, 2019 for more information]. PDD test results are categorized at the level of the test as a whole regardless of whether the decision is made using grand total or subtotal scores. In practical terms, the procedural rubric for the SSR is that test results are classified as

indicative of deception – commonly using the term *significant reactions* – whenever any subtotal score equals or exceeds the cutscore for deceptive classifications, and are classified as indicative of truth-telling – using the term *no significant reactions* – when all subtotal scores equal or exceed the cutscore for truthful classifications. Examination results are classified as inconclusive or no opinion (i.e., not statistically significant for deception or truth-telling) when none of the sub-total scores equals or exceeds the cutscore for deceptive classification while less than all sub-total scores equal or exceed the cutscore for truthful classifications.

Previous research (Barland, Honts & Barger, 1989; Podlesney & Truslow, 1993; Department of Defense, 1995a; 1995b) has not supported the hypothesis of test sensitivity or specificity at the level of the individual RQs, and field practices dictate that examiners are not permitted to render decisions of both deception and truth-telling within a single examination. For this reason, there was no attempt to determine deception to some RQs and truth-telling to other RQs within the individual cases in the Monte Carlo space.

Results

Criterion accuracy was calculated for each of the three USAF MGQT conditions (i.e., two, three, and four RQs) for the three test data analysis methods (i.e., seven-position, three-position, and ESS). Accuracy indices of interest included the following: test sensitivity to deception, test specificity to truth-telling, false-negative and false-positive error rates, and inconclusive rates for deceptive and truthful cases. Positive predictive value (PPV; calculated as true positives divided by all positive results), negative predictive value (NPV; calculated as true negatives divided by all negative results), the proportions of correct decisions without inconclusive results for deceptive and truthful cases, along with the unweighted average of the proportions of correct decisions and inconclusive results for the deceptive and truthful cases. All statistical analyses were completed with a level of significance set at alpha = .05. These may be found in Appendices A through D.

Decision accuracy for USAF MGQT exams with two, three and four RQs.

Test accuracy effects were evaluated using a Monte Carlo hypothesis test. This method involves the use of Monte Carlo methods to calculate the statistical confidence interval (Efron & Hastie, 2016; Efron & Tibshirani, 1986; 1993) which is then compared with the null-hypothesis or chance value (i.e., .5). Results are interpreted as not statistically significant when the chance value is not contained within the confidence interval, or when the limits of the 1 – alpha confidence interval exceed the chance value.

Monte Carlo confidence intervals were calculated as the alpha/2 = .025th and 1-alpha/2 = .975th percentile of 10,000 iterations of a Monte Carlo space consisting of n = 100 simulated multiple-issue exams. Separate Monte Carlo simulations were conducted for multiple-issues examinations with two, three and four RQs. Nine different Monte Carlo simulations were completed. In addition, a 10th Monte Carlo simulation was calculated with the number of RQs randomized from two to four.

For each Monte Carlo simulation, criterion accuracy was calculated for each iteration of the Monte Carlo space, including test sensitivity, specificity, false-positive and false negative error rates, along with positive-predictive-value, negative-predictive-value, unweighted decision accuracy and inconclusive rates for deceptive and truthful cases. the all observed data. The mean standard deviation was also calculated for each dimension of criterion accuracy, so that factorial ANOVAs could also be computed for number of RQs x scoring method x criterion state.

Results are shown in Appendices A, B and C for multiple-issue polygraphs two, three and four RQs. Appendix D shows the results while varying the number of RQs for the cases within each iteration of the Monte Carlo space.

Sensitivity and specificity for USAF MGQT exams with two, three and four RQs.

The method described by Cohen (2002) was used – along with the mean sample sizes in the Monte Carlo space (n=50 for deceptive

case and mean n=50 for truthful cases), and the Monte Carlo means and standard deviations – to calculate a three-way ANOVA (criterion status x TDA method x number of RQs) for decision accuracy including inconclusive results (i.e., test sensitivity and specificity). Table 2 shows the three-way ANOVA summary, and Figure 1 shows the mean plot for test sensitivity and specificity. The three-way interaction was significant $F(4,882) = 5.705$, $p < .001$). This result indicated that differences may exist for in the effectiveness of three-position, seven-position and ESS scoring methods with criterion deceptive and criterion truthful exams with two, three or four relevant questions.

Table 2. Three-way ANOVA summary for accuracy (number of RQs x TDA method x criterion state).

Source	SS	df	MS	F	p	F crit .05
# RQs	0.048	2	0.024	2.684	.069	3.006
Status	4.368	1	4.368	486.435	<.001	3.852
Model	8.240	2	4.120	458.787	<.001	3.006
# RQs x Status	28.203	2	14.102	1570.261	<.001	3.006
Status x Model	4.629	2	2.314	257.719	<.001	3.006
# RQs x Model	0.170	4	0.042	4.731	.001	2.382
# RQs x Status x Model	0.204	4	0.051	5.669	<.001	2.382
Error	7.921	882	0.009			
Total	53.784	899				

Figure 1. Mean plot for test sensitivity and specificity for three-position, seven-position and ESS scoring methods.

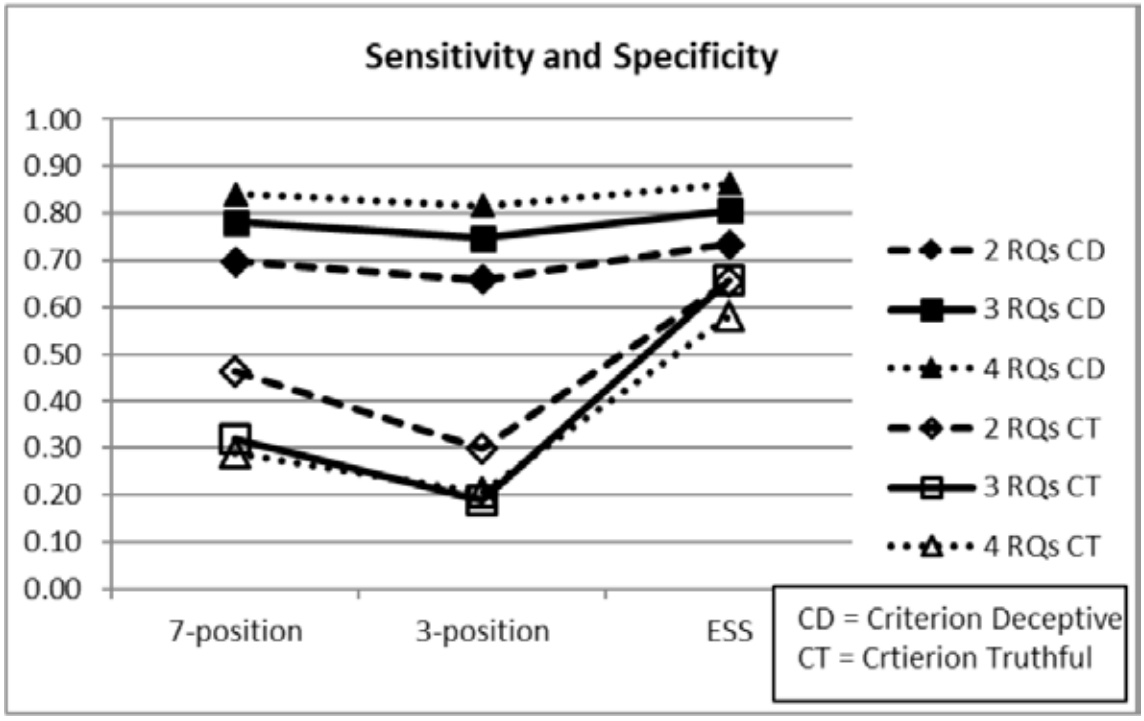


Figure 1 shows that mean test sensitivity to deception exceeded chance (.5) for all three scoring methods, while mean test specificity to truth-telling did not exceed chance for the seven-position and three-position scoring methods.

Because the 3-way ANOVA was significant, post-hoc 2x2 ANOVAs (TDA method x number of RQs) were completed separately for the deceptive and truthful case in the Monte Carlo model. The 2-way ANOVA, shown in Table 3, was statistically significant for the deceptive cases $F(1,441) = 4.848$, $p = .028$), indicating an interaction for TDA model and the number of RQs. One-way ANOVAs were not significant for the number of RQs ($p = .071$) or the scoring method ($p = .625$) with the deceptive cases.

Table 3. Two-way ANOVA summary for accuracy with deceptive cases (TDA model x number RQs).

Source	SS	df	MS	F	p	F crit .05
Model	0.277	2	0.002	0.942	.391	3.016
# RQs	1.564	2	0.010	5.327	.005	3.016
Interaction	0.009	1	0.009	4.848	.028	3.863
Error	0.863	441	0.002			
Total	1.850	446				

Results from a two-way ANOVA for the truthful cases are shown in Table 4. The interaction of TDA method x number of RQs was significant for the truthful cases $F(1,441) = 5.669$, $p = <.001$). One-way ANOVAs showed that main effects for the truthful cases were not significant for the number of RQs (p = .799) or for the scoring method (p = .056).

Table 4. Two-way ANOVA summary for accuracy with truthful cases (TDA model x number RQs).

Source	SS	df	MS	F	p	F crit .05
Model	12.593	2	0.084	5.428	.005	3.016
# RQs	1.044	2	0.007	0.450	.638	3.016
Interaction	0.364	1	0.364	23.541	<.001	3.863
Error	6.821	441	0.015			
Total	14.001	446				

These results suggest that the main source of variance for the three-way interaction can be attributed to differences in abilities of the three scoring methods to detect deception and truth-telling. To further understand the influence of scoring method on decision accuracy, a final 3 way contrast was calculated for the seven-position and three-position results, excluding the ESS results. The three-way interaction for number of RQs x scoring method x criterion state was not significant *[*$F(4,588) = 0.916$, $p = 0.454$] when ESS results were excluded. This suggests that the initial three-way interaction can be attributed to differences in decision accuracy for ESS results with truthful cases.

Inconclusive rates for USAF MGQT exams with two, three and four RQs.

A three-way ANOVA was conducted (criterion status x TDA model x number of RQs) for inconclusive results. The three-way ANOVA summary for inconclusive results is shown in Table 5.The three-way interaction for inconclusive results was significant $F(4,882) = 2.580$, $p = .036$ for TDA method x number of RQs x criterion state.

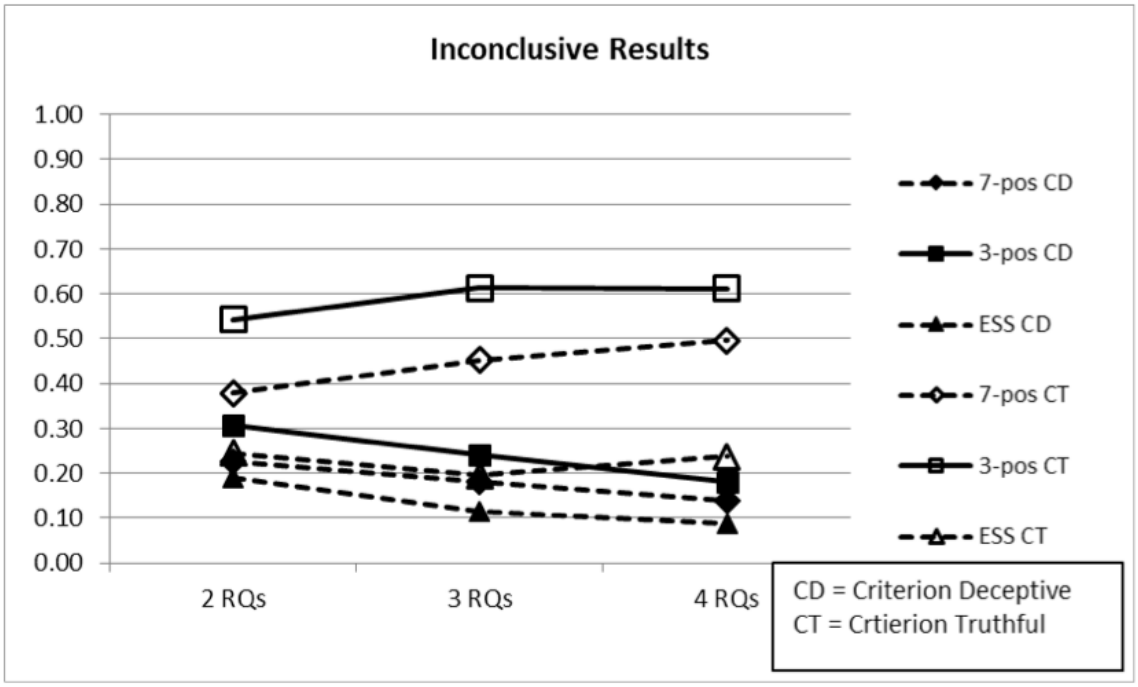
Table 5. Three-way ANOVA summary for inconclusive results (RQs x TDA method x criterion state)

Source	SS	df	MS	F	p	F crit .05
# RQs	0.076	2	0.038	3.103	.045	3.006
Status	1.786	1	1.786	145.371	<.001	3.852
Model	8.482	2	4.241	345.097	<.001	3.006
# RQs x Status	11.506	2	5.753	468.140	<.001	3.006
Status x Model	2.468	2	1.234	100.431	<.001	3.006
# RQs x Model	0.228	4	0.057	4.638	.001	2.382
# RQs x Status x Model	0.127	4	0.032	2.580	.036	2.382
Error	10.839	882	0.012			
Total	35.512	899				

Figure 2 shows the mean plot for inconclusive results for deceptive and truthful cases for the seven-position, three-position, and ESS methods with two, three, and four RQs. Mean inconclusive rates were generally higher for truthful than for deceptive cases, and this difference was more pronounced for the three-position and seven-position methods. Simple mean effects were not significant for differences in inconclusive results for the seven-position method (p = .156) or for the ESS (p = .415). The simple mean effect was significant for inconclusive results for the three-position scoring method with criterion deceptive and criterion truthful cases [$F(1,98) = 4.382$, (p = .039)].

Two-way ANOVAs for each scoring method showed a significant interaction for number of RQs x criterion status, including the seven-position [$F(1,294) = 31.435$, ($p < .001$)], three-position [$F(1,294) = 37.143$, ($p < .001$)] and ESS [$F(1,294) = 17.702$, ($p < .001$)]. Simple main effects for inconclusive results as function of RQs with the seven-position results were not significant for criterion deceptive cases (p = .316) or criterion truthful cases (p = .894). For

Figure 2. Mean plot for inconclusive results.



three-position results the simple main effects were also not significant for criterion deceptive cases (p = .157) or for criterion truthful cases (p = .936). Simple main effects for ESS scores also showed no significant difference between inconclusive results as function of the number of RQs for criterion deceptive (p = .161) or criterion truthful cases (p = .940).

A two way ANOVA for TDA method x number of RQs for *criterion truthful* cases was statistically significant F (1,441) = 14.183, (p < .001). Simple main effects for differences in scoring method were not significant for two RQs (p = .083), three RQs (p = .085) or four RQs (p = .428). After combining the cells for different scoring methods, the main effect for inconclusive rates as a function of the number of RQs cases was not significant (p = .962) with the criterion truthful cases. A post-hoc power analysis was completed using the power.anova.test() function in the R Language and Environment for Statistical Computing (R Core Team, 2019), indicating a power > .99 to detect a significant difference if one exists.

Simple main effects for the number of RQs were not significant for inconclusive results with criterion truthful cases for the seven-position scoring method (p = .866), the three-position method (p = .936) or the ESS (p = .940). After combining the cells for two, three and four RQs, the main effect for differences in inconclusive results as a function of scoring method was statistically significant F(2,447) = 1250.483, (p < .001) for the criterion truthful cases. This indicates that the observed interaction effects inconclusive results as a function of RQs x scoring method can be attributed to differences between the scoring methods with criterion truthful cases.

Another two-way ANOVA for TDA method x number of RQs showed a statistically significant interaction for the *criterion deceptive* cases F(1,441) = 17.789, p = <.001). Simple main effects were not significant for differences in inconclusive results among criterion deceptive cases as a function of different scoring methods with two RQs (p = .218), three RQs (p = .080) or four RQs (p = .218). After combining the cells for the different scoring methods, the main effect of RQs on inconclusive results was not statistically significant for the deceptive cases (p = .209). A post-hoc power analysis indicated a power > .99 to detect a significant effect for the number of RQs if one exists.

Simple main effects for the number of RQs were not significant for seven position (p = .316), three-position (p = .157) or ESS (p = .161) methods. After combining the cells for two, three and four RQs, the main effect for differences in inconclusive results as a function of scoring method was statistically significant F(2,447) = 3.424, (p = .033) for the criterion deceptive cases. This suggests that inconclusive rates for criterion deceptive cases varied more as a function of scoring method than the number of RQs.

Inspection of the plot in Figure 2 shows that mean inconclusive rates for criterion truthful cases with the ESS may to have a different slope compared to other results. To further understand the influence of scoring method on observed inconclusive rates a three-way ANOVA contrast was calculated for the seven-position and three-position scores, excluding the ESS scores. The three-way interaction for inconclusive results was not significant [F(4,588) = 0.051, (p = .995)] for the seven-position and three-position scoring methods when ESS results were excluded. These results suggest the three way interaction for inconclusive results can be attributed to the differences in results for criterion truthful cases with the ESS. The two-way interactions for each scoring method indicate that inconclusive rates can be expected to increase with the number of RQs for criterion truthful cases and decrease with the number of RQs for criterion deceptive cases.

False-negative and false-positive errors for USAF MGQT exams with two, three and four RQs.

Figure 3 shows the mean plot for false-positive and false-negative errors. A three-way ANOVA was completed (criterion status x TDA method x number of RQs) for decision errors. The ANOVA summary for decision errors is shown in Table 6. The three-way interaction was not statistically significant F(4,882) = 0.943, p = .438.

Because the three-way interaction was not significant, a two-way ANOVA was calculated for RQs x criterion state after combining the cells

Figure 3. Mean plot for false-positive and false-negative errors.

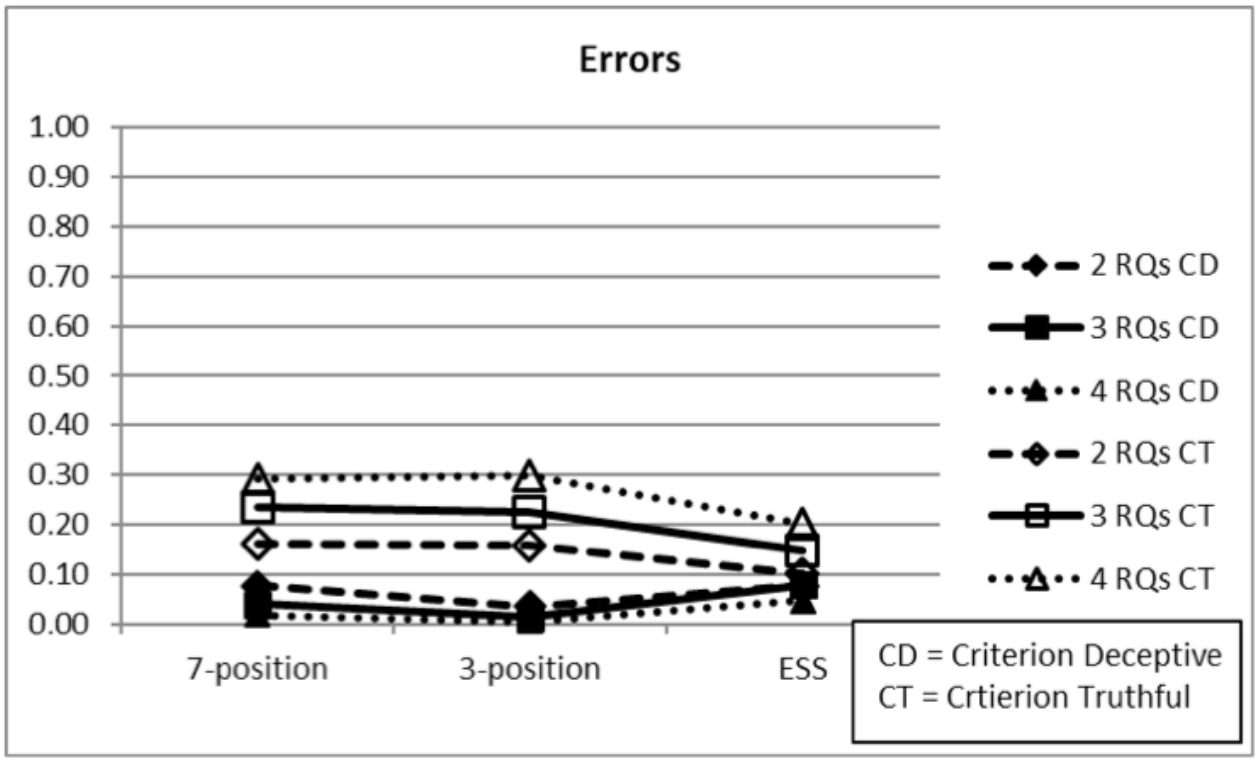


Table 6. Three-way ANOVA summary for errors (RQs x TDA method x criterion state)

Source	SS	df	MS	F	p	F crit .05
# RQs	0.273	2	0.137	14.086	<.001	3.006
Status	0.827	1	0.827	85.294	<.001	3.852
Model	0.123	2	0.062	6.358	.002	3.006
# RQs x Status	5.862	2	2.931	302.373	<.001	3.006
Status x Model	0.684	2	0.342	35.283	<.001	3.006
# RQs x Model	0.015	4	0.004	0.394	.813	2.382
# RQs x Status x Model	0.037	4	0.009	0.943	.438	2.382
Error	8.550	882	0.010			
Total	16.371	899				

for the three TDA methods. Figure 4 shows the mean plot. The two-way ANOVA summary shown in Table 7 indicates a significant interaction [$F(1,894) = 104.051$, ($p < .001$)] for decision errors as a function of the number of RQs and criterion state.

Although errors appear to increase with number of RQs for criterion truthful cases and decrease with the number of RQs for criterion deceptive cases, the simple main effects for the number of RQs were not statistically significant for criterion deceptive cases (p = .459) or for criterion truthful cases (p = .814).

Figure 4. Mean plot for decision errors with combined scoring methods.

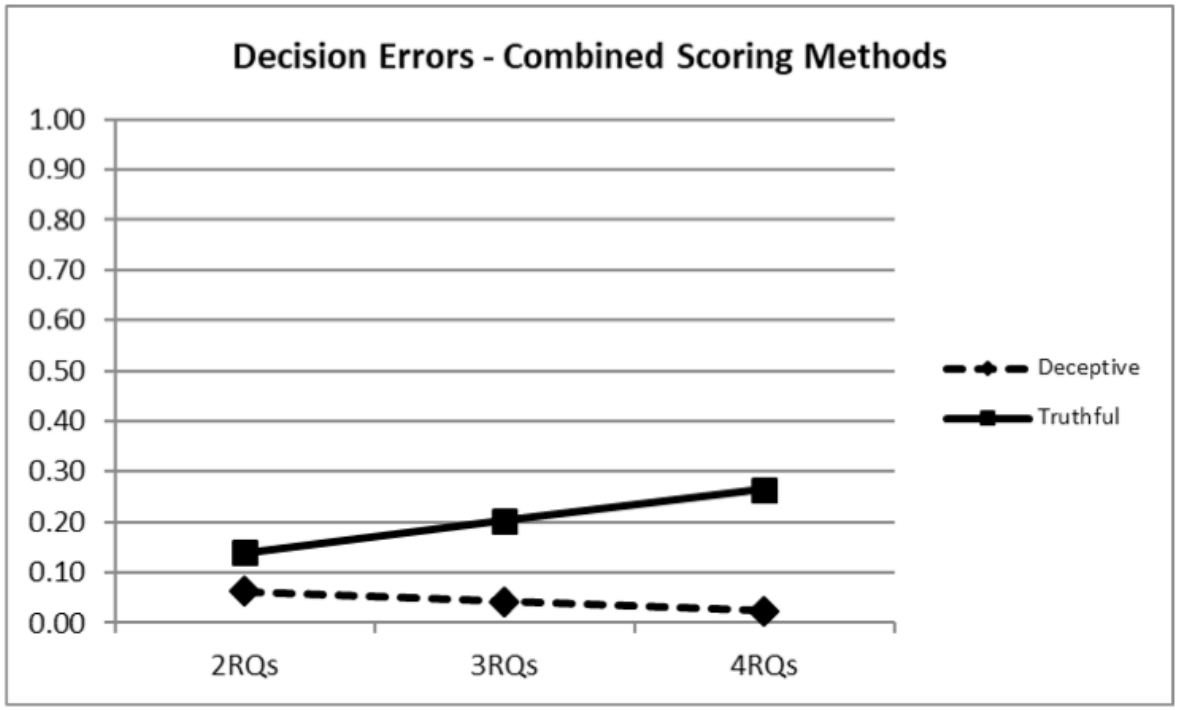


Table 7. Two-way ANOVA summary for decision errors with 7 position scores (RQs x criterion state).

Source	SS	df	MS	F	p	F crit .05
# RQs	0.273	1	0.001	0.094	.759	3.852
Status	5.680	1	0.013	1.302	.254	3.852
Interaction	1.009	1	1.009	104.051	<.001	3.852
Error	8.666	894	0.010			
Total	6.962	897				

A post-hoc power calculation for the one-way simple main effects, with n = 50 for each cell, had power > .99 to detect a significant effect if one actually existed. This suggests that the observed interaction can be attributed to the fact that, although the difference for two, three or four RQs are not significant within the truthful or deceptive cases, the likelihood of testing error for multiple issue polygraphs increases with the number of RQs for criterion truthful cases while decreasing for criterion deceptive cases.

Unweighted average accuracy.

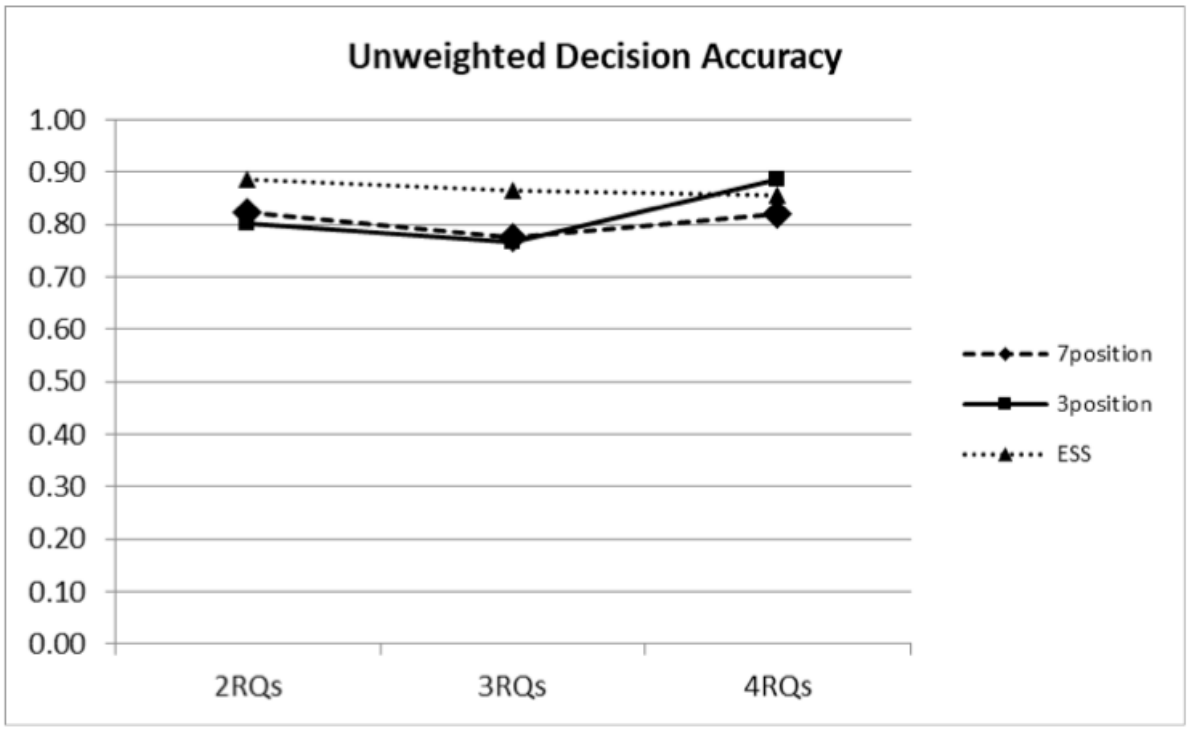
Unweighted decision accuracy excluding inconclusive results is shown in Table 2, and was significantly greater than chance (.5) for all three TDA methods with two, three, and four RQs ($p < .05$). Table 8 also shows that variation in test accuracy increases as a function of the number of RQs for all three scoring methods. A two-way .Similarly, as shown in the appendices, both false-negative and false-positive errors were reduced to statistically significantly less than chance for all TDA versions with two, three, and four RQs.

Table 8. Unweighted accuracy: mean (SD) {95% CI}.

	7-position	3-position	ESS
2RQs	.822 (.061) {.702 to .942}	.802 (.073) {.659 to .945}	.886 (.047) {.795 to .978}
3RQs	.775 (.104) {.571 to .979}	.766 (.128) {.515 to .999}	.866 (.067) {.734 to .998}
4RQs	.820 (.146) {.533 to .999}	.887 (.149) {.595 to .999}	.855 (.101) {.657 to .999}

Figure 5 shows the mean plot for unweighted average accuracy (i.e., unweighted average of decision accuracy with criterion deceptive and criterion truthful cases). A two-way interaction was significant for number of RQs x scoring method [F(1,891) = 51.009, (p < .001)]. However, simple main effects were not significant for the different scoring methods for two RQs (p = .711), three RQs (p = .824), or 4 RQs (p = .959). Simple main effects were also not significant for the seven-position method (p = .975), three-position method (p = .839), or the ESS (p = .871). Although the lines in Figure 1 exhibit different slope, none of the lines is itself significantly different from zero.

Figure 5. Mean plot for unweighted average accuracy.



After combining the cells for different scoring methods, a one-way ANOVA showed that differences in unweighted accuracy as a function of the number of RQs were not statistically significant [$F(2,897) = 0.046$, ($p = .955$)]. A post-hoc power analysis indicated the ANOVA had power > .99 to detect a significant effect. These results indicate there is no real difference in unweighted accuracy for PDD results with 2RQs, 3RQs or 4RQs, excluding inconclusive results.

Criterion accuracy for randomized two, three, or four questions.

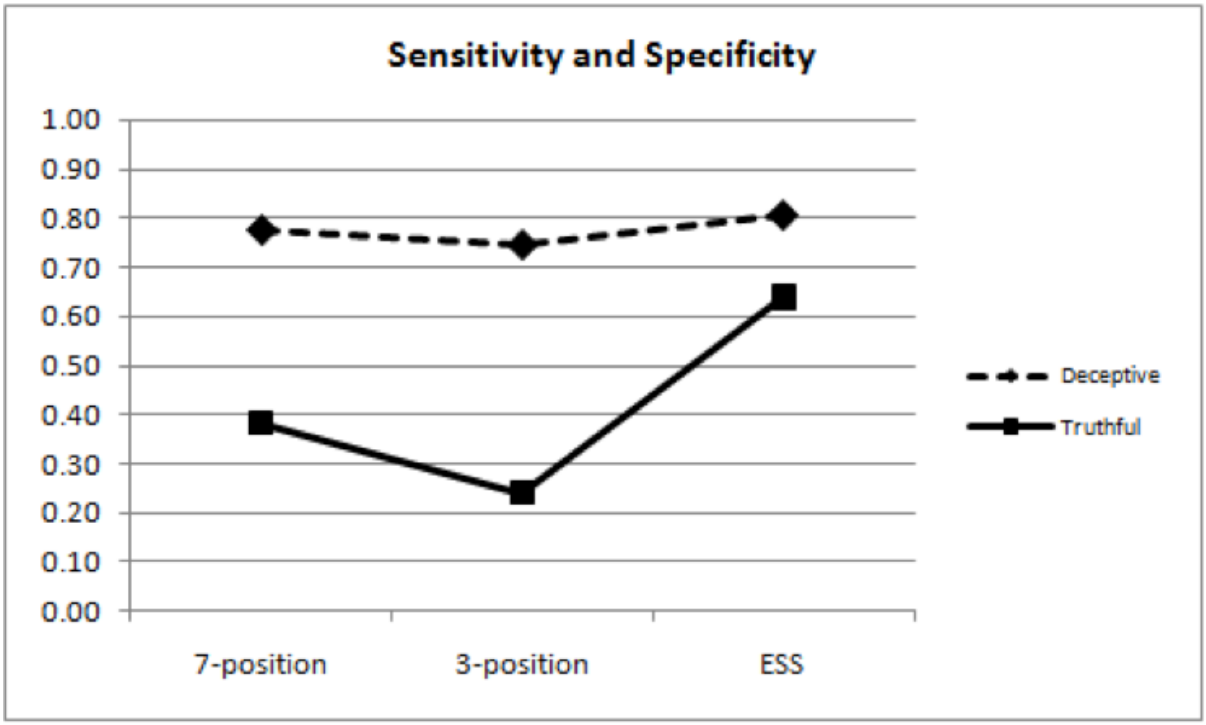
Three additional Monte Carlo models were used to further understand any differences between the seven-position, three-position, and ESS scoring methods while randomizing the number of RQs for each case in the Monte Carlo space. For each case, the number of RQs was varied randomly from two, three, or four by comparing a random number to the values .3333333 and .666666. The proportions of cases with two, three, and four RQs would vary for each iteration of the Monte Carlo space, and would converge to equal proportions in the Monte Carlo distribution of results that consisted of 10,000 iterations of the Monte Carlo space.

Base rates for the criterion state of individual questions were as follows; for cases with two RQs the base rate = .293, for cases with three RQs = .206, and four RQs = .159. For each RQ in each case a random uniform number was compared to the base-rate, and the criterion state was set to truthful if the base-rate was less than the random number. This ensured that although the proportion of criterion deceptive and criterion truthful cases would vary for each iteration of the Monte Carlo space, the base-rate for deception would converge to .5 for the Monte Carlo distribution of results while randomly setting the number of RQs for each exam and randomly setting the criterion state for each RQ. Each case was evaluated with the seven-position, three-position and ESS scoring methods using the SSR that was described earlier. Appendix D shows the means, standard deviations, and 95% confidence intervals for the Monte Carlo distribution of results while varying the number of RQs from two, three, or four.

Sensitivity and specificity for USAF MGQT exams with randomized two, three, or four RQs.

A two-way ANOVA for decision accuracy showed a significant interaction between scoring method and criterion status $F(1,294) = 177.039$, $p < .001$. Figure 6 shows a plot of the means for test sensitivity and specificity. The simple main effects were not statistically significant for test sensitivity to deception ($p = .659$) or for specificity to truth-telling ($p = .064$). A post-hoc power analysis indicated a likelihood of power > .99 for detecting a significant difference if one existed.

Figure 6. Monte Carlo mean estimates for test sensitivity and specificity.

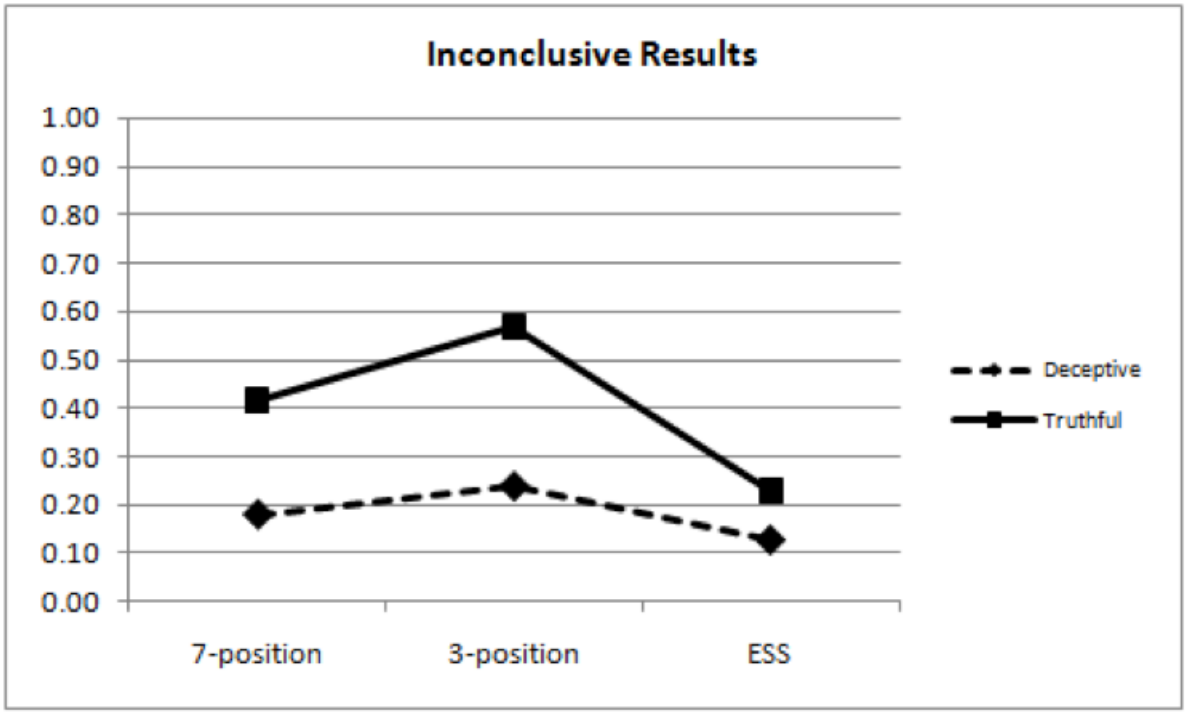


Evaluation of the simple main effects for scoring method showed that the difference in detection of deception differed significantly from detection of truth-telling for the seven-position scoring method [$F(1,98) = 8.307$, ($p = .005$)] and for the three-position scoring method [$F(1,98) = 19.438$, ($p < .001$)]. The simple main effect for criterion deceptive and criterion truthful cases was not significant for the ESS (p = .222). These results indicate the two-way interaction can be attributed to differences test sensitivity and test specificity for the ESS scoring method compared to the seven-position and three-position methods. As shown in Appendix D, although test sensitivity to deception was significantly greater than chance (.5) for all three scoring methods, test specificity to truth-telling did not exceed chance for the seven-position or three-position methods.

Inconclusive results for USAF MGQT exams with randomized two, three, or four RQs.

A two-way ANOVA for inconclusive results (scoring method x criterion status) showed significant differences in inconclusive results for the three TDA methods $F(1,294) = 71.927$, $p < .001$. Figure 7 shows the Monte Carlo mean for inconclusive rates for the three TDA methods. Simple main effects for inconclusive results were not significant for the deceptive cases ($p = .185$) or truthful cases ($p = .177$).

Figure 7. Monte Carlo mean estimates for inconclusive rates.



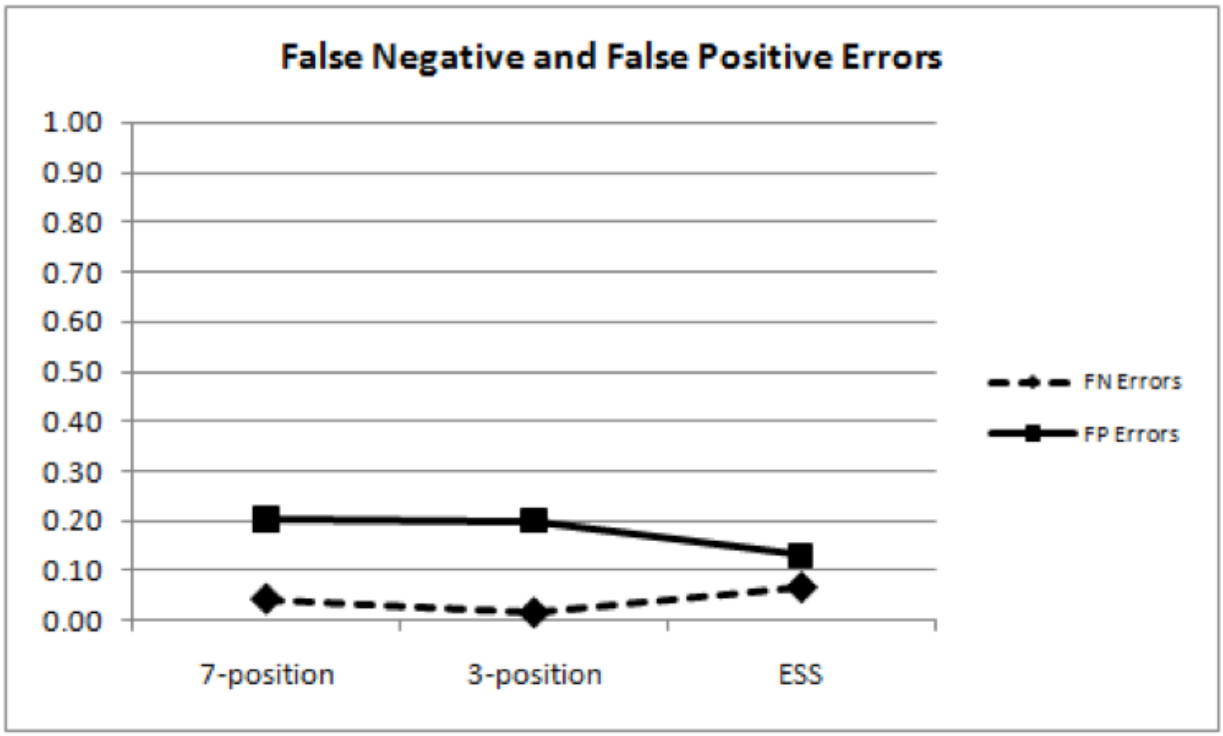
The simple main effect, for differences in inconclusive rates with criterion deceptive and criterion truthful cases, was significant for the three-position scores [$F(1,98) = 5.147$, ($p = .025$)], but not for seven-position scores (p = .084) or the ESS (p = .413). These results indicate that the observed two-way interaction (TDA method x criterion state) for inconclusive results can be attributed to the significant difference between the inconclusive rates for criterion deceptive and criterion truthful cases with the three-position scoring method. Mean inconclusive rates were elevated for three-position results compared to the seven-position and ESS results, and were greater for criterion truthful cases.

Decision errors for USAF MGQT exams with randomized two, three, or four RQs.

A two-way ANOVA for decision errors by criterion status showed a significant interaction

between TDA method and criterion status $F(1,294) = 31.456$, $p < .001$. Figure 8 shows the Monte Carlo means for error rates for the three TDA methods.

Figure 8. Monte Carlo mean estimates for inconclusive rates.



Simple main effects for were not significant for false-negative errors (p = .229) or for false-positive errors (p = .874). Additionally, none of the simple main effects were statistically significant for the seven-position scoring method (p = .223), three-position scoring method (p = .097) or the ESS (p = .510). Post-hoc power analysis using showed that the experiment had power > .99 to detect a significant effect if one existed. The observed interaction of decision errors can be thought of as indicating that the two lines in Figure 8 have significantly different slope though neither of the lines is itself significantly different from zero slope, meaning observed differences are within the range of expected uncontrolled/unexplained variation. These results indicate no real difference exists between the false-negative rates and no real difference exists in false-positive rates for the seven-position, three-position and ESS methods.

Discussion

This project is a Monte Carlo study of criterion accuracy effects of multiple-issue polygraphs with two, three, and four RQs, such as the USAF MGQT. Although some differences in criterion accuracy are expected as a function of the number of RQs, previous studies have not investigated these differences. Multiple issue polygraphs are commonly used in polygraph screening programs – in the absence of any known allegation or incident.

A defining characteristic of multiple issue screening polygraphs is that the questions are interpreted with an assumption of independent criterion variance. The overall test results for multiple issue polygraphs is inherited from the question results. In practical terms, test results of multiple-issue exams are inherited from the lowest question score. This differs from event-specific polygraphs for which the test result is determined at the level of the test as a whole, and where the question results are inherited from the overall test result. Some known difficulties exist in studying multiple-issue polygraphs. One difficulty is in acquiring knowledge about the criterion state for each of the individual test questions.

Another difficulty will be the management of multiplicity effects – the aggregation of statistical error as a function of making conclusions based on multiple probability events. Finally, there is the difficulty of acquiring a sample data, ideally a balanced sample with an equal number of cases in each different testing condition, of suitable size for study and analysis.

An advantage of the Monte Carlo approach to this project is the reduction of expense, in terms of human activity and other resources, in the acquisition of data for which the criterion state of each RQ can be known with certainty. Another advantage of the Monte Carlo approach to this project was the ability to more easily compare the effectiveness of different scoring methods – the seven-position, three-position and the ESS.

Results from this study indicate that some differences exist in the effectiveness of different scoring methods for criterion deceptive and criterion truthful cases with, two, three, or four RQs. However, these differences are not observed in terms of unweighted decision accuracy – the unweighted average decision accuracy with criterion deceptive and criterion truthful cases, excluding inconclusive results. No real differences were found in unweighted accuracy as a function of the number of RQs. Unweighted average decision accuracy for multiple-issue polygraphs with two, three or four RQs significantly exceeded chance (.5) for all three TDA methods.

Despite the fact that unweighted accuracy did not differ for multiple-issue polygraphs with two, three or four RQs, the results of study indicate that some differences do exist when considering the other dimensions of test accuracy. Mean test sensitivity to deception exceeded chance (.5) for all three scoring methods. However, mean test specificity to truth-telling did not exceed chance for the seven-position and three-position scoring methods, and test specificity was significantly greater than chance only for the two RQ model with the ESS.

Differences were observed in inconclusive rates as a function of the number of RQs and as a function of scoring method. Inconclusive rates can be expected to increase with the number of RQs for criterion truthful cases and decrease with the number of RQs for criterion deceptive cases. However, results with the ESS may produce a different pattern of inconclusive rates with criterion truthful cases compared to other scoring methods. One possible reason for this, not explored in this study, is the use of a statistical correction for multiplicity effects for the ESS cutscore for truthful classifications. It is possible that the use of ESS scores with traditional cutscores may result in inclusive rates that adhere more closely to the trend exhibited by the seven-position and three-position results in this study.

No significant differences were found in false-positive or false-negative error rates as a function of the number of RQs. Post-hoc power analyses suggest that this study had sufficient power to detect significant effects for testing if they exist. Although the differences for two, three or four RQs were not significant within the criterion truthful cases or criterion deceptive cases, the likelihood of testing error increased with the number of RQs for criterion truthful cases while decreasing for criterion deceptive cases.

In addition to the investigation of criterion accuracy differences that may exist as a function of the number of RQs in multiple-issue polygraphs, Monte Carlo methods were used to compare results for the seven-position, three-position and ESS methods. Results from this analysis showed that all three methods achieved unweighted decision accuracy that significantly exceeded the chance level (.5). Test sensitivity to deception exceeded chance for all three scoring methods. However, test specificity to truth-telling did not exceed chance for the seven-position or three-position methods. Mean inconclusive rates were highest for the three-position scoring method, and this was loaded for criterion truthful cases. Despite these observed differences, results showed no significant difference in the false-negative rates and no significant difference in false-positive rates for the seven-position, three-position and ESS methods.

A limitation of this study is that no effort was made to evaluate difference in criterion accuracy for the three scoring methods as function of differences in numerical cutscores. Results for the seven-position and three-position scoring methods were obtained using traditional numerical cutscores (-3 or less at any subtotal

for deceptive classifications, +3 or greater at all subtotals for truthful classifications) with no statistical correction for multiplicity effects. Results for the ESS were obtained using statistically referenced cutscores for which a statistical correction was used to manage multiplicity effects for truthful outcomes. ESS cut-scores were -3 or less at any subtotal for deception and + 1 or greater at all subtotals for truth-telling. It is possible that some interactions and some effects may differ if all results were obtained using cutscores that are optimized through statistically optimized (or if all results were obtained using traditional) cutscores. It is also possible that different decision rules, involving some use of the grand total score, may achieve an improvement in test specificity and inconclusive results without undesired compromises in test sensitivity and f false-negative rates. This should be subject to future research.

Another limitation of this project is the overall design as a Monte Carlo simulation. Monte Carlo models, although insufficient to provide a final or definitive answer to hypothetical questions, are highly useful to study high-cost, and high-risk problems as well as complex and difficult problems. Results of Monte Carlo studies should be replicated evaluated together with the results of other laboratory and field studies. Use of subtotal seed parameters that were obtained from the subtotals of confirmed single issue examinations represents another limitation. However, seed parameters from the subtotal scores of single issue examinations, although imperfect in their ability to represent the subtotal scores of multi-issue exams, offer the advantage of a reasonably known criterion status for use as seed parameters for Monte Carlo simulation.

Another noteworthy limitation of the present study is that no attempt was made to investigate test sensitivity or test specificity at the level of the individual questions. Although decision rules were executed at the level of the subtotal scores for individual questions, classifications of deception and truth-telling were made at the level of the test as a whole. No attempt was made to determine truthfulness to some questions and deception to other questions within the Monte Carlo cases. These procedures are is consistent with field polygraph practices.

In summary, results of this study support the validity of the hypothesis that multiple-issue PDD exams with two, three, or four RQs, can differentiate deception from truth-telling at rates that are significantly greater than chance when scored with the seven-position, three-position, and ESS TDA models. Suggestions for future research include the further study of multiplicity effects, statistical optimization of decision cutscores and decision rules for multiple-issue polygraphs. Multiple-issue polygraph formats that can be used with two, three or four RQs, such as the USAF MGQT, offer the potential for great adaptability and usefulness in a variety of field practice settings, and continued interest in multiple-issue PDD formats is indicated.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics. Sage.*

Barland, G. H., Honts, C. R. & Barger, S.D. (1989). *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.*

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

Cohen, B. (2002). Calculating a factorial ANOVA from means and standard deviations. *Understanding Statistics 1(3):191-203.*

Department of Defense (2006). *Federal psychophysiological detection of deception examiner handbook.* Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Reprinted in Polygraph, 40(1), 2-66.

Department of Defense (2006). *Psychophysiological Detection of Deception Analysis II - Course #503.* Test data analysis: DoDPI numerical evaluation scoring system. Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007).

Efron, B. & Tibshirani R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science,* 1(1), 54-77.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

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.

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.

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

Marin, J. (2000). He said/She said: Polygraph evidence in court. *Polygraph, 29*, 299-304.

Marin, J. (2001). The ASTM exclusionary standard and the APA 'litigation certificate' program. *Polygraph, 30,* 288-293.

Nelson, R. (2017). Multinomial reference distributions for comparison question polygraphs. *Polygraph and Forensic Credibility Assessment, 46(2),* 81-115.

Nelson, R. & Blalock, B. (2016). Extended analysis of Senter, Waller and Krapohl's USAF MGQT examination data with the Empirical Scoring System and the Objective Scoring System, version 3. *Polygraph, 45(1*), 90-94.

Nelson, R., Blalock, B. & Handler, M. (2011). Criterion validity of the Empirical Scoring System and the Objective Scoring System, version 3 with the USAF Modified General Question Technique. *Polygraph, 40(11),* 172-179.

Nelson, R., Blalock, B. & Handler, M. (2019). Practical Polygraph: How to Parse Categorical Results for Test Questions of Diagnostic and Screening Polygraphs. *APA Magazine, 52(3)*, 60-65.

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

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

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

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

Nelson, R. & Krapohl, D. (2011). Criterion Validity of the Empirical Scoring System with Experienced Examiners: Comparison with the Seven-Position Evidentiary Model Using the Federal Zone Comparison Technique. *Polygraph,* (In press).

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. & Rider, J. (2018). Practical polygraph: ESS-M made simple. *APA Magazine, 51(6),* 55-62.

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.

R Core Team (2019). R: *A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/.

Research Division Staff (1995). *A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope Polygraph and the test for espionage and sabotage question formats.* DTIC AD Number A319333. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 26(2), 79-106.

Research Division Staff (1995). *Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage.* DTIC AD Number A330774. Department of Defense Polygraph Institute. Fort Jackson, SC. Reprinted in Polygraph, 27, (3), 171-180.

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

Robertson, B. (2012). The Use of an Enhanced Polygraph Scoring Technique in Homeland Security: The Empirical Scoring System—Making a Difference. . Naval Postgraduate School, Dudley Knox Library: Retrieved from: https://www.hsdl.org/?abstract&did=710340.

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

Šidàk, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

Summers, W. G. (1939). Science can get the confession. *Fordham Law Review, 8,* 334-354.

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. T*he Newsletter of the Canadian Association of Police Polygraphists, 7,* 28-47. Reprinted in Polygraph, 20(2), 70-79.

Appendix A.

Criterion Accuracy of Multiple-issue Polygraphs with Two RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.822 (.061) {.702 to .942}	.802 (.073) {.659 to .945}	.886 (.047) {.795 to .978}
Unweighted INC	.302 (.055) {.195 to .409}	.424 (.054) {.319 to .529}	.217 (.050) {.119 to .316}
D INC	.226 (.050) {.128 to .324}	.306 (.054) {.201 to .412}	.190 (.043) {.105 to .275}
T INC	.378 (.097) {.188 to .567}	.542 (.095) {.355 to .729}	.245 (.088) {.072 to .417}
Sensitivity	.697 (.053) {.593 to .800}	.659 (.055) {.550 to .767}	.734 (.049) {.637 to .831}
Specificity	.462 (.101) {.265 to .659}	.300 (.090) {.123 to .476}	.655 (.076) {.506 to .804}
FN	.077 (.032) {.015 to .140}	.035 (.021) {.001 to .076}	.076 (.030) {.018 to .135}
FP	.160 (.077) {.010 to .310}	.158 (.071) {.018 to .298}	.100 (.060) {.001 to .217}
PPV	.929 (.035) {.861 to .998}	.925 (.036) {.854 to .996}	.957 (.027) {.905 to .999}
NPV	.666 (.116) {.439 to .893}	.743 (.142) {.465 to .999}	.737 (.098) {.545 to .929}
D Correct	.900 (.040) {.821 to .979}	.950 (.030) {.891 to 1.009}	.906 (.037) {.834 to .977}
T Correct	.743 (.118) {.513 to .974}	.654 (.145) {.369 to .940}	.867 (.080) {.710 to .999}

Appendix B.

Criterion Accuracy of Multiple-issue Polygraphs with Three RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.775 (.104) {.571 to .979}	.766 (.128) {.515 to .999}	.866 (.067) {.734 to .998}
Unweighted INC	.317 (.074) {.171 to .462}	.427 (.071) {.288 to .567}	.156 (.063) {.032 to .279}
D INC	.180 (.038) {.106 to .254}	.242 (.046) {.152 to .331}	.116 (.035) {.048 to .184}
T INC	.453 (.142) {.175 to .732}	.613 (.136) {.346 to .880}	.195 (.121) {.001 to .432}
Sensitivity	.781 (.041) {.701 to .862}	.747 (.046) {.656 to .837}	.806 (.042) {.724 to .889}
Specificity	.320 (.130) {.066 to .574}	.188 (.094) {.004 to .372}	.659 (.141) {.383 to .934}
FN	.039 (.021) {.001 to .08}	.012 (.012) {.001 to .035}	.078 (.028) {.023 to .133}
FP	.235 (.131) {.001 to .493}	.226 (.114) {.002 to .450}	.146 (.108) {.001 to .359}
PPV	.960 (.024) {.914 to .999}	.959 (.022) {.915 to .999}	.975 (.019) {.938 to .999}
NPV	.545 (.190) {.173 to .917}	.728 (.244) {.250 to .999}	.549 (.128) {.298 to .800}
D Correct	.953 (.025) {.903 to .999}	.984 (.016) {.954 to .999}	.912 (.032) {.850 to .974}
T Correct	.589 (.203) {.190 to .987}	.475 (.198) {.086 to .864}	.819 (.131) {.563 to .999}

Appendix C.

Criterion Accuracy of Multiple-issue Polygraphs with Four RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Accuracy	.820 (.146) {.533 to .999}	.887 (.149) {.595 to .999}	.855 (.101) {.657 to .999}
Unweighted INC	.318 (.108) {.107 to .528}	.396 (.112) {.177 to .615}	.163 (.096) {.001 to .351}
D INC	.140 (.035) {.071 to .208}	.180 (.039) {.103 to .257}	.089 (.031) {.028 to .150}
T INC	.496 (.211) {.082 to .91}	.612 (.220) {.181 to .999}	.237 (.191) {.001 to .611}
Sensitivity	.842 (.037) {.771 to .914}	.816 (.039) {.738 to .893}	.864 (.036) {.793 to .934}
Specificity	.289 (.148) {.001 to .580}	.205 (.109) {.001 to .419}	.581 (.200) {.190 to .972}
FN	.018 (.014) {.001 to .046}	.005 (.007) {.001 to .018}	.047 (.022) {.003 to .091}
FP	.292 (.198) {.001 to .680}	.298 (.205) {.001 to .700}	.202 (.180) {.001 to .555}
PPV	.976 (.018) {.940 to .999}	.976 (.017) {.942 to .999}	.985 (.013) {.959 to .999}
NPV	.581 (.262) {.067 to .999}	.815 (.25) {.324 to .999}	.454 (.185) {.092 to .816}
D Correct	.979 (.017) {.946 to .999}	.995 (.008) {.978 to .999}	.948 (.024) {.900 to .996}
T Correct	.546 (.257) {.042 to .999}	.505 (.259) {.001 to .999}	.754 (.201) {.359 to .999}

Appendix D.

Criterion Accuracy with Combined/Randomized (2, 3, or 4) RQs

	7-position Mean (SE) {95% CI}	3-position Mean (SE) {95% CI}	ESS Mean (SE) {95% CI}
Unweighted Average Accuracy	.799 (.088) {.627 to .971}	.775 (.107) {.565 to .984}	.878 (.060) {.760 to .996}
Unweighted Inconclusives	.294 (.072) {.154 to .434}	.403 (.071) {.263 to .543}	.178 (.059) {.062 to .294}
D INC	.177 (.044) {.091 to .263}	.238 (.047) {.146 to .331}	.129 (.036) {.059 to .198}
T INC	.411 (.133) {.149 to .672}	.568 (.136) {.300 to .835}	.228 (.114) {.004 to .453}
Sensitivity	.780 (.047) {.689 to .871}	.746 (.048) {.651 to .841}	.805 (.043) {.722 to .889}
Specificity	.382 (.128) {.131 to .633}	.241 (.110) {.025 to .456}	.642 (.130) {.387 to .897}
FN	.043 (.022) {.001 to .085}	.016 (.013) {.001 to .042}	.066 (.027) {.014 to .118}
FP	.208 (.111) {.001 to .427}	.200 (.109) {.001 to .414}	.130 (.092) {.001 to .310}
PPV	.956 (.025) {.907 to .999}	.956 (.025) {.906 to .999}	.974 (.019) {.936 to .999}
NPV	.605 (.165) {.281 to .929}	.733 (.205) {.331 to .9994}	.622 (.127) {.373 to .870}
D Correct	.948 (.026) {.897 to .999}	.979 (.017) {.945 to .999}	.924 (.031) {.864 to .984}
T Correct	.649 (.174) {.309 to .989}	.555 (.203) {.157 to .954}	.832 (.117) {.603 to .999}

A Proposed Framework for Polygraph Test Questions

Donald J. Krapohl[1] and Donnie W. Dutton[2]

Introduction

Polygraph test question construction, review and presentation are essential skills for polygraph examiners. The American Polygraph Association (APA) recognizes its importance in its educational standards that mandate 32 instruction hours for polygraph students on these areas. All APA polygraph education programs teach specific rules for question development, how they are introduced to the examinee and how they are presented during testing. The rules evolved over generations of polygraph examiners and there is wide consensus among practitioners about what those rules are. What has yet to appear is a theoretical framework for polygraph test questions; There seems to be fairly good agreement on the *how* of polygraph questions, but not the *why*. Why should relevant questions be direct and use action verbs? Why are emotionally charged words, legalisms and specialized jargon avoided? Why should comparison questions be broad or ambiguous but relevant questions must be clear and as narrow as possible? Why do we review all the test questions with the examinee before running charts? Why don't we test anyone after they've been through an intense interrogation? We know the rules, but those rules are not organized within some larger concept. Rather than a theoretical framework, professional faith is invested in the list of rules, most of which have no known originator.

To understand why the polygraph works it is first necessary to abandon the assumption that the act of lying is what instigates polygraph reactions. The reliance upon the notion that lying causes reactions and truthtelling does not is a common misapprehension. It would be true if the polygraph were a "lie detector", but no such device exists. This is not to deny lying very likely plays a mediating role in the ultimate size of the physiological reactions due to associated emotions such as guilt or fear of detection (see Kahn, Nelson & Handler [2009] for an excellent review). Lying appears to augment the intensity of the reaction (Elaad & Ben-Shakhar, 1989), but something else is going on to trigger the reaction.

There is evidence that overt lying may not even be necessary for the polygraph to function. This evidence comes in three parts. The first is found in the directed-lie comparison (DLC) question. When an examinee answers "no" to a DLC she or he is not being deceptive as we generally define the term. The examinee has no intent to misrepresent the truth. The examinee is not trying to fool the examiner. The act of answering incorrectly on DLCs is not even the examinee's idea – the examinee is merely following instructions. Nevertheless,

[1]APA Past President, School Director of the Behavioural Methods – UK Polygraph Training Center (BMUK) and regular contributor to this publication. Questions and comments regarding this article can be directed to APAkrapohl@gmail.com.

[2]APA Past President, current APA Director and regular contributor to this publication.

This paper is part of the series titled Best Practices.

Disclaimer

Both authors are with the Capital Center for Credibility Assessment (C3A) and provide instructions for the Behavioural Methods – UK Polygraph Training Center. The views expressed herein do not necessarily reflect those of C3A or BMUK.

Acknowledgments

We are very grateful to Pamela Shaw, Matthew Andrews and Mark Handler for their insightful comments and suggestions to an earlier draft of this paper.

the DLC seems to function well as a comparison question (Honts & Raskin, 1988; Horowitz, Kircher, Honts & Raskin, 1997). Second, polygraph examiners who have conducted the Silent Answer Test (Horvath & Reid, 1972) can attest that an examinee will react to the same test questions irrespective of whether the examinee answered the questions out loud. Lying is neither sufficient nor even necessary to explain those reactions. Without an examinee answer, the source of the reaction must be the question itself. Finally, most examiners will have noticed that the examinee normally begins to react before the question is completely asked and before the examinee answers. If the act of lying were the cause of the reaction, its onset would be tied to the lie rather than to the question.

Polygraph examiners who assume the polygraph is a "lie detector" may be vulnerable to certain errors as they develop polygraph test questions. Many of those errors are the focus of this paper. If deception is not what causes reactions, how can the polygraph detect deception? As we discuss in the next section, it doesn't – exactly.

The Basics

Polygraph testing is a straightforward Stimulus-Response paradigm. As in all such paradigms, a stimulus is presented, a response is recorded, and inferences are based upon the relationship between the two. In polygraphy the stimuli are the test questions (not the examinee's answer) and physiological arousals are the responses. These test questions vary from one another in some essential characteristic, and differences in arousal intensity are believed to covary with that characteristic.

Distilled to their essence there are three types of polygraph questions. There are neutral questions, which as the name suggests, are those that are not evocative in any way. There are relevant questions, which are evocative, and address the behaviors or actions of interest. The third category of question is called a comparison question, also evocative, the reactions to which polygraph examiners use as a benchmark to gauge the significance of the examinee's reactivity to the relevant question. During polygraph testing these three types of questions are interspersed in a sequence, repeated several times, and the resulting arousals are tracked with any of several available scoring systems. Some polygraph techniques include other kinds of questions. Cushman and Krapohl (2010) summarized the available evidence for these other kinds of questions, and reported that they have either been found to be invalid or have no published research.

Generally speaking, physiological arousals can be either spontaneous or associated with a stimulus. Those that are spontaneous are not informative, at least not in polygraph testing, and we will not consider them further. Arousals that correspond with the test stimuli may be informative if they are elicited by a certain stimulus characteristic. There are three characteristics of a stimulus that will evoke physiological responding. They are novelty, intensity, and salience (Dawson, Schell, & Filion, 2007). Please note that deception is not included in this list. Presence or absence of the three characteristics will determine whether there are non-random arousals to the stimuli. When present, increases and decreases in the degree of these characteristics are reflected in corresponding increases and decreases in the amplitude of the subsequent arousals (Barry, 1975; Hovland & Riesen, 1940; Katkin, 2003; Lole, Gonsalvez, Blaszcznski & Clarke, 2012).

The novelty of a stimulus regards whether it is unexpected, new, different, or surprising in a given context. In polygraph testing novelty can be a contaminant because it does not contribute to the goal of veracity testing. Reactions to an unexpected or surprising stimulus can only reveal that it was unexpected or surprising. Novelty is therefore strenuously avoided in the examination protocol. Steps to avoid novelty include reviewing all questions before the test, using a neutral or irrelevant question as the first question in the test sequence, and by first performing a practice test with the examinee. Examiners strive to ensure extraneous sights and sounds do not intrude upon the testing. Test question development considers whether there are dramatic differences in question length, and during testing every effort is expended to avoid attention-grabbing differences in test questions due to the pitch, speed, emphasis and hesitations in the examiner's voice.

Stimulus intensity refers to that which is painful or very aversive. A real-life example of stimulus intensity is the sound of fire and smoke alarms. They are exceptionally loud, harsh, and cause a physiological response. They are intended to capture attention, be arousing and motivate people to move to safety. As with stimulus novelty, stimulus intensity is a contaminant in polygraph testing. It elicits arousals but those arousals are not meaningful toward the aim of deception detection. The polygraph testing protocol, therefore, calls for presentation of the test questions in a normal volume and the testing procedure that carefully attends to the comfort of the examinee.

The third and final characteristic of a stimulus that elicits a physiological response is salience, a term used here to refer to the personal meaningfulness, significance or importance of the stimulus. The term "personal" is central to this definition, as it can vary from individual to individual, and in degree between guilty and innocent examinees. Salience is the stimulus characteristic that governs test question construction and presentation. In polygraph testing, the meaningfulness of a question can be assessed by observing the frequency and magnitude of physiological arousals that accompany the presentation of the question.

Polygraph relevant test questions can be especially meaningful if an examinee intends to lie to them. Similarly, relevant questions are probably meaningful to truthful examinees inasmuch as they relate to the reason the examinee is taking the test. It is logical to conclude that all examinees consider relevant questions important and therefore arousing.

What support is there, beyond the reasonableness of the assumption, that relevant questions are judged to be important by all examinees? There are two converging lines of evidence. One is the finding that, on average, differential arousal between relevant and comparison questions is smaller for truthful examinees than it is for deceptive examinees (Franz, 1989; Krapohl, Gordon & Lombardi, 2008; Krapohl & McManus, 1999; Patrick & Iacono, 1989; Raskin, Kircher, Honts, & Horowitz, 1988). Consequently, decision accuracy is higher for deceptive examinees than it is for truthful examinees across virtually every polygraph technique (Nelson, 2015). This is consistent with an expectation that truthtellers share some concern for relevant questions.

A second finding comes from the research with the Relevant-Irrelevant Test (RIT), which contains no comparison questions. In the RIT, the consistency of reactions to the relevant questions is used as an indication of deception whereas the opposite provides the basis for a decision of truthfulness. If truthtellers did not find relevant questions salient there could be an expectation that they would pass the RIT in equal proportions to liars who failed this test. Research on the RIT find that it produces high rates of false positives and low rates of false negatives (Horowitz, Kircher, Honts & Raskin, 1997; Krapohl & Goodson, 2015; Krapohl & Rosales, 2014) indicating that both truthtellers and liars find relevant questions personally significant. When added to the evidence from the comparison question test regarding response asymmetry, the trend in the findings support a conclusion that relevant questions can be expected to be important to truthtellers and liars.

Past explanations of the comparison question test have relied most heavily on the fear of detection (FoD) model, that is, the physiological arousals on relevant questions by liars were due to their worry about being revealed as the guilty party whereas innocent examinees were fearful of comparison questions because of a belief those questions would reveal their involvement in non-relevant socially proscribed behaviors. Related perspectives include psychological set, fear of punishment, conditioned reactions and conflict theories. Insufficiencies in the FoD model became clear with the advent of the directed-lie version of the comparison question. Fear could not explain reactivity to the directed lies, and these questions appeared to perform as well as traditional probable-lie comparison questions (Honts & Raskin, 1988; Horowitz, Kircher, Honts & Raskin, 1997). Additionally, the FoD failed to explain why high polygraph accuracy was found in low-motivation laboratory studies of the polygraph in which the participants had little reason to fear detection (Bradley & Ainsworth, 1984; Honts, Raskin & Kircher, 1987; Horvath, 1988; Horvath & Palmatier, 2008). It also provided no explanation for tentative evidence of a high false positive rate among truthful examinees

tested by law enforcement to support claims of having been victims of a violent crime (Barland, 1982; Raskin, Kircher, Honts, & Horowitz, 1988). Finally, data from a lab study by Offe and Offe (2007) suggested that differential reactivity arises from the salience of the relevant questions, not from the comparison questions. For these reasons an alternative to the FoD model was necessary.

A new theory was proposed by Ginton (2009) that appears to address the shortcomings of the FoD model. Called Relevant-Issue Gravity (RIG), the theory is premised on differences in the binding power of relevant questions upon the attention of truthtellers and liars. It takes as given that all examinees find relevant questions salient. The difference is that for the guilty examinee there is a memory of the behavior, called an episodic memory, that drives the salience, a memory the innocent examinee does not have. RIG theory predicts that the episodic memory resident in the guilty examinee compels the examinee's attention to the relevant question, which will give rise to a physiological response that is expected from salient stimuli. Truthtellers, in contrast, with no episodic memory for the relevant topic, find a larger portion of their attention shifted to distractor items, which polygraph examiners call comparison questions. From this vantage, the true purpose of comparison questions is not to impose fear to compete with the fear of the relevant question. Rather, comparison questions are placed in the question sequence to test the degree to which an examinee's attention can be shifted from the relevant questions. The expectation is that the presence of an episodic memory pertaining to the relevant issue can be revealed by the persistence of reactivity to relevant questions. Not only does the RIG theory overcome the incompatibilities between FoD prediction and the evidence, it is consistent with Offe and Offe's conclusion that a large contributor to decision accuracy is how much reactivity takes place on the relevant question when it is juxtaposed in a sequence with potential distractors.

The central role memory plays in the process of deception can easily be demonstrated with a simple thought experiment. Here it is. We propose the reader prepare to tell a lie for the following question: In what month were you born? Regardless of what month the reader has chosen to offer as the lie, the first answer to come to mind was the correct month, not the lie. The false answer never comes to mind first. This demonstrates a core function of the mind, to seek accurate information stored in memory before deciding on the answer. If the memory exists, asking the right question of an attentive person will trigger the memory. This is true for both liars and truthtellers, and can be exploited in polygraph testing by carefully choosing test questions.

To summarize the basics:

1. Polygraph testing is a Stimulus-Response paradigm. Questions are the stimuli and the physiological arousals are the responses.

2. The three characteristics of any stimulus that evokes a physiological response are novelty, intensity, and salience.

3. In general, the more of any of these three characteristics a stimulus has, the greater the response.

4. Polygraph is a test of salience. Deception is inferred from physiological arousals that signal the degree of salience. Fear or other emotions, as well as cognitive load, may follow the internal appraisal of salience and mediate the response intensity but they are not the initiator of polygraph reactions.

5. Relevant questions can be salient to both truthtellers and liars, though they generally differ in the degree of that salience.

6. Questions cause people automatically to seek accurate information in the form of memories to answer the question, irrespective of how they answer.

7. The use of comparison questions (distractor items) in a test is to help as-

sess whether the examinee has an episodic memory in the scope of the relevant question inasmuch as examinees who have such memories will produce larger relative reactions to relevant questions than examinees who do not. Examinee answers to comparison questions do not necessarily need to be lies for these questions to serve their function as distractors but should trigger the person's search for accurate information or memories should trigger the person's search for accurate information or memories.

Implications

If the previous assumptions are correct they give rise to guiding principles regarding examinee suitability and test question construction, introduction and presentation. The principles also reveal shortcomings of the "lie detector" and FoD perspectives in polygraph testing. We address these in the next four sections.

Suitability

Effects of Priming or Conditioning

It is generally accepted among polygraph practitioners that exams should be rescheduled if the examinee had recently been interrogated extensively. The reason is self-evident. If polygraph is a test of salience, and salience has been artificially imposed upon the relevant topic by an intense accusatory period shortly before polygraph testing, reactivity would be expected to appear on the relevant questions irrespective of whether the examinee was deceptive. For this reason, examiners typically allow a cooling-off period between an interrogation and a polygraph examination.

Testing Possible Victims of Trauma

It may be recalled from an earlier discussion that being asked a question will cause an individual to search his or her memory for accurate information, regardless of the answer chosen. Good polygraph questions are intended to target the recollections of deceptive examinees while simultaneously allowing truthful examinees to know such recollections do not exist in their memories. This process works well for most cases. An exception might be when the recollection is associated with trauma, such as having been a victim of a sexual assault. Test questions that trigger that type of memory can be very personally meaningful as well as emotionally disturbing. For these individuals physiological responding to relevant questions may have an explanation other than deception. One expected effect of an association between trauma and relevant questions that bring to mind that trauma is that polygraph examinees with those memories could be inclined toward false positive results. There has been one study that provides a glimpse of this possibility. In a reanalysis of a 1977 blind scoring study by Frank Horvath, Barland (1982) and Raskin et al. (1988) reported that almost all false positive errors in that study of field cases were from tests of victims. While more work is needed to confirm that finding, it is consistent with the notion that test questions that force the recall of a harrowing experience will always be highly significant to the examinee, and consequently more likely to elicit physiological responses that may be indistinguishable from those associated with deception.

Test Question Construction

Taboos

Overly intrusive test questions about very personal conduct can present special challenges in polygraph testing. This is most true on the topic of sex, an area that until recent years was routinely found in the polygraph examinations of police officer candidates. For the average individual, questions over personal sexual practices can be unsettling. This is because in most cultures queries about such topics, especially from strangers, are a substantial breach of etiquette. The topic of sex is always salient in such circumstances. For this reason, relevant questions about sexual practices may elicit physiological arousals among truthful examinees due to embarrassment, anger, defensiveness, or the concealment of tangentially related behaviors the examinee prefers not to discuss. Collectively, the polygraph profession found long ago it needed to be very careful of sexually based comparison questions, a recognition that the very strong reactions they often cause could risk a false negative result except under certain conditions.

There are two common circumstances where sex-based relevant questions may not be so problematic. One is when the examination is focused on a criminal sexual offense for which the examinee is a suspect. In these kinds of cases the relevant questions regard a particular criminal act, not the examinee's sexual behavior in general. Relevant questions are structured to prompt a memory of a very specific event in the mind of the guilty examinee, one for which the truthful examinee is certain he did not commit. For comparison questions one of the recommended practices for tests about sexual crimes is to cover broad areas of sexual activities (Abrams, 1989; Krapohl & Shaw, 2015; Matte, 1996; Reid & Inbau, 1977) where the concern addressed in the previous paragraph may be exploited for the benefit of the exam. Because of the power of sex-based comparison questions is so potentially great, they are restricted to examinations in which the relevant questions are about a sexual crime: Sexual activities are never covered in comparison questions in tests where the relevant question does not also cover sexual activities.

A second circumstance where sex-based relevant questions may be less troublesome are in the testing of convicted sex offenders in treatment. One of the pertinent factors that distinguishes these examinees from others is that they will have had multiple lengthy and in-depth conversations with therapists and parole/probation officers regarding their sexual interests and activities well before the offenders are subject to polygraph testing. As such the taboo normally associated with sexual inquiries may be less upsetting, even expected, than it would be for the average citizen in the community.

Evocative Terms

In work unrelated to polygraph testing Bradley and Lang (1999) investigated the affective power of single words. They asked undergraduate college students for subjective ratings on common English words along the dimensions of happy vs unhappy, excited vs calm and controlled vs in-control. From this they developed affective norms for more than 1000 words. The Bradley and Lang data indicated certain terms had a combination of significant arousal and negative valence. This subset of words was judged as strongly stimulating in a negative emotional direction. Said simply, they were significantly more disturbing than are most words. Among the strongest were rape, mutilate, murderer, assault, violent, crucify and slaughter. Relatedly, many polygraph students are taught to avoid these kinds of words because they are suspected of causing reactions in themselves. For practical reasons Bradley and Lang could not assess the affective power of all words, but the concordance between their findings and common polygraph instruction would indicate the indiscriminate use of potentially evocative words may come with a risk of adding a confounding source of salience to polygraph questions. Similar evidence is found in the work of Dindo and Fowles (2007).

Doubt

Polygraph test questions can garner importance to the examinee if he is uncertain of the truthfulness of his answer. Ambiguous test questions complicate the examinee's mental search for information, increasing cognitive demands and sometimes anxiety or other emotions. Questions that induce doubt will be salient by their nature. Doubt may be harmful or helpful to polygraph testing, depending on whether test questions evoking doubt are relevant or comparison questions.

If relevant questions are insufficiently clear to the examinee due to unfamiliar terminology, ambiguous expressions, poorly chosen words, excessive complexity or scoping that exceeds an examinee's capacity to completely recall or effortlessly process what is being asked, reactivity can occur to the question that is unassociated with deception. It is one of the reasons polygraph students are instructed to use clear and concise relevant questions and to stay away from legal terminology[3], scientific names, or words unfamiliar to the examinee. Unless the examinee is a specialist in a given field, memories associated with the behavior of interest will not be associated with specialized terms. Examiners are instructed in school to use the examinee's vocabulary in relevant test questions to minimize examinee doubt. Similarly, compound relevant questions or "shopping list" relevant questions may introduce complexity that also induces reactivity due to processing demands, a problem that can be

avoided by merely testing separate behaviors in separate test questions.

In contrast to doubt's negative effect on relevant questions, uncertainty can be very useful with probable-lie comparison (PLC) questions and is regularly used by examiners. The more doubt the question can induce the more effective it can be. Polygraph students are taught to make PLCs as broad as possible. The rationale for this approach is usually based on increasing the likelihood that the examinee will be lying to the PLC. Since lying is not necessary for comparison questions to function (remember DLCs?) the more plausible mechanism is that the significance of the question has been enhanced by the examinee's uncertainty. The effect of doubt can easily be demonstrated using a volunteer attached to a polygraph who is asked a series of unrehearsed and increasingly difficult trivia questions that require strictly *yes* or *no* answers. Reactivity will generally covary with the level of uncertainty the volunteer experiences.

One of the longstanding debates among polygraph examiners is which type of PLC is superior, broad questions proposed by John Reid which can incidentally encompass the relevant topic (Reid & Inbau, 1977) or a later version introduced by Cleve Backster that is specifically designed to exclude the relevant topic (Matte, 1996). They are often referred to as "non-exclusionary" and "exclusionary" PLCs, respectively. Backster argued that exclusionary PLCs reduced the chance that a guilty examinee would confuse a PLC with a relevant question, the consequence of such confusion leading to an inconclusive result rather than a correct deceptive one. Those in the Reid camp advocate for the non-exclusionary approach, arguing it functions better than does the exclusionary PLC. History proved Backster to be the more persuasive one, and exclusionary PLCs came to be the dominant method in the field.

Backster's rationale would be correct if one looked at the polygraph strictly as a "lie-detector". As a "lie-detector" the guilty examinee would be lying to the relevant topic with both the relevant question and the non-exclusionary PLC, potentially producing equivalent reactions to both categories of question and an inconclusive outcome. If "salience detector" is a better description of the polygraph, it is reasonable to anticipate Reid's more expansive non-exclusionary PLC would perform better than Backster's narrower PLC. Reid's broader question would be expected to engender more uncertainty for the examinee. Large-sample research suggests that Reid was correct (Amsel, 1999; Horvath & Palmatier, 2008). The non-exclusionary PLC produced higher polygraph classification accuracy for both truthful and deceptive examinees in those studies, though the evidence is mixed as to which group benefited more. Increased salience imposed via greater doubt may be a significant contributor to this difference.

Task Demands

As discussed earlier, DLC questions appear to function well as distractors even though they do not require deception as the PLC does[4]. The question is why this may be true. Returning again to the assumption that polygraph is a test of salience, why should DLCs evoke reactions?

The answer may be that DLCs are the only questions in the series of questions that include special instructions. That is, on DLCs and only on DLCs does the examinee need to remember to answer differently from the other questions. This is called a task demand. In psychology task demands are processes that

[3]One of the test questions in polygraph screening of individuals seeking a US Government security clearance uses the word "espionage," a term so specialized that examiners often must read or recite the legal definition to the examinee. It is a test question that speaks to a specific law rather than to a behavior (e.g., giving classified information to a foreign government without authorization), likely making it a less effective polygraph question. By way of example, consider the following two questions regarding income tax evasion. Question 1: Did you commit tax fraud in 2019? Question 2: Did you submit a falsified tax form to the IRS in 2019? The first question addresses the law, the latter the behavior. Most examiners are likely to agree that the second question is better because the language is probably how the examinee would have encoded the act in his memory. Because polygraph examiners are trained in the selection of test questions, it seems reasonable that the "espionage" question is the product of government policy drafted by those not familiar with the current understanding of polygraph.

are required to perform a task. The more difficult is the task the greater is the task demand. Because answering a DLC entails a different task from answering other questions, it gains significance for the examinee. To make DLCs more salient in practice, they are always linked to a memory. They are also examinee-referential, that is, they ask about the examinee's behavior (e.g., Have you ever broken a minor traffic law?) versus non-self-referential and trivial DLCs (e.g., Is Hawaii an island?) because the latter has been shown to be significantly less effective (Horowitz, Kircher, Honts & Raskin, 1997). This would be predicted using the salience model. The combination of linking the DLC with an episodic memory along with the added task demands associated with answering DLCs provide a better explanation for their evocative power than does the FoD model.

Test Question Introduction

Excessive Emphasis

A subtle influence could be caused by examiner behavior during the pretest interview. Examples might include where the examiner states he or she does not believe the examinee on the relevant topic(s), and that the examination is merely pro forma for an inevitable failed outcome. In the same vein, if an examiner relates to the examinee that the only questions on the examination that matter are the relevant questions, the instruction may diminish the distracting power of comparison questions and shift reactivity in the direction expected of deceivers. In both cases the examiner has imposed personal importance upon the relevant questions, increasing the likelihood of greater reactivity to them irrespective of whether the examinee is innocent.

Conversely, an overemphasis on comparison questions can also influence detection accuracy. If accusations can affect relevant questions, so too can they affect comparison questions. Similarly, a disproportionate amount of time invested on one question or category of question can signal to the examinee that they are more important than other questions. In this way these questions may inadvertently acquire additional salience above that which they would have otherwise.

Test Question Presentation

Uniformity

As discussed previously, novelty or surprise can evoke physiological responding. For this reason, the presentation of test questions during physiological recording should not vary in ways that are novel or surprising. Examples of the kinds of differences that may be novel include changes in the pitch, speed, fluidity, or volume in which some questions are presented but not others. Examiners with latent expectation bias may inadvertently change how some questions are asked, and thereby load novelty, a contaminant, onto those questions.

To avoid this possibility examiners can opt for the digitized voice for reading the questions during the test. Digitized voices are not encumbered with the degree of variability that human voices have. They present every question the same way every time. Reducing variability in the presentation of the test questions is expected to reduce the variability in the subsequent responses. Automated presentation of test questions may incrementally increase polygraph accuracy (Honts & Amato,

[4]Users of the DLC methodology were likely taught to have the examinee think of a transgression but not to tell the examiner what it is. The real purpose for discouraging the examinee from discussing the transgression is not found in the literature, but because the present authors present when it took place, they can now reveal that the prohibition on soliciting info¬rmation on DLCs was based on politics rather than science. The DLC entered the mainstream in the early 1990s as a replacement for the PLC in US government security screening. The shift was due to examinees complaining in large numbers to leadership and legislators about the intrusiveness of some of the questions, meaning PLCs. Moreover, some examinees lost their security clearances after making admissions on PLCs. Because not all examinees got the same PLCs, one's disqualifying admissions may have come about because of the PLC the examiner chose rather than the standardized coverage of the relevant questions everyone received. To address this problem, US government examiners were taught to explicitly tell examinees not to disclose their transgressions regarding DLCs, only to think of one or two peccadillos in the area the DLC covered. This became doctrinal even to those outside of government who adopted the DLC. A problem recognized by many examiners who use DLCs is that some examinees can habituate to these questions quickly. One reason for the habituation may be that the examinees begin to answer the DLCs by rote without accessing the related transgression in memory. Perhaps a better approach would be to tell the examinee that he will be asked after the exam what the transgression was and so he will need to recall it, though never actually pursuing the matter after testing. This manipulation would make for a partial test of the salience model proposed here.

1999). Almost all computer polygraphs offer a digitized voice option for reading the test questions. One obvious precaution is to ensure that an unfamiliar accent or poor quality of voice synthesis does not create a distraction in itself. There is at least one type of uniformity that should be avoided during testing. That is the practice of beginning several different relevant questions with the same phrase, thereby preventing the examinee knowing which question it is until late into the question presentation. If an examinee intends to deceive on one of the questions, the introductory phrase may prompt an initial reaction to each test question that begins with the same phrase, introducing noise into the physiological data. The problem could be more severe if relevant and comparison questions both share the same phrase.

Extraneous Stimuli

Unexpected sights and sounds during testing may also induce phasic responses. If examiners clear their throats, shift in their chair, or do other things just before asking a test question, these sounds may be novel enough to induce a reaction. Likewise, examinees can usually hear when an examiner types on the keyboard and may interpret the typing as an indication something occurred on the chart at the question when typing took place. This may cause reactions.

Examinees normally have a field of view of 210 degrees, and if the examiner is in that visual range his or her movements can catch the attention of the examinee. If found to be novel or interesting to the examinee, the movements may be responsible for reactivity. Examiners must be mindful that novel sights and sounds in the polygraph suite interfere with the testing process.

To control the potential effect of extraneous sounds on the polygraph data, including those caused by outside noises not under the control of the examiner, sound-abating headphones can be placed on the examinee. If the digitized voice is also used through the headphones, reactions are more likely to be associated with the test questions than the outside distractions.

Conclusion

We propose here an alternate to the FoD model to better explain the underlying causes of reactions during polygraph testing. The FoD model, a longstanding hypothesis in the polygraph field, only predicts effects when fear is the dominant factor during polygraph testing. Others have previously identified this model's insufficiency (Khan, Nelson & Handler, 2009). Complete reliance on the FoD model can lead to errors in question development and the interpretation of physiological responses that accompany those questions. In its place we encourage consideration of the effects of novelty, intensity and salience when developing, reviewing and presenting polygraph questions. We also invite investigation of Ginton's (2009) Relevant-Issue Gravity framework, a perspective not yet widely appreciated in the field. It expands upon the basis that the polygraph is a test of salience, provides a more generalizable foundation for how comparison questions function, and invites the possibility of an entirely new category of comparison question. It is one of the stronger candidates to replace the FoD model.

As a closing comment we observe that the development of theoretical frameworks is sometimes viewed by some as merely academic exercises. If an existing model is "good enough," there is no need for a new one, they might argue. No model will change how we go about our practice. Perhaps they have a point.

Except:

In the world of polygraph, where there are often substantial consequences that can accompany polygraph results, we would submit that any framework that can reduce errors of process should be of interest to all practitioners. In this paper we offer the possibility of just such a framework, constructed from the work of many before us, and encourage further tests that may prove, disprove or improve it.

References

Abrams, S. (1989). *The Complete Polygraph Handbook.* Lexington Books: Lexington, MA

Amsel, T.T. (1999). Exclusive or non-exclusive comparison questions: A comparative field study. *Polygraph, 28*(4). 273 – 283.

Barland, G.H. (1982). On the accuracy of the polygraph: An evaluative review of Lykken's *Tremor in the Blood. Polygraph, 11(*3), 258 – 272.

Berry, R.J. (1975). Low-intensity auditory stimulation and the GSR orienting response. P*hysiological Psychology, 3*(1), 98 – 100.

Bradley, M.M., and Lang, P.J. (1999). A*ffective norms for English words (ANEW): Instruction manual and affective ratings.* Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Bradley, M.T., and Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. *Psychophysiology, 21*(1), 63 – 71.

Cushman, B., and Krapohl, D. (2010). *The Evidence for Technical Questions in Polygraph Techniques.* Presentation to the American Polygraph Association Annual Seminar, Mytle Beach, SC.

Dawson, M.E., Schell, A.M., and Filion, D.L. (2007). The electrodermal system. In J.T Cacioppo, L.G. Tassinary and G.G. Berntson (Eds.) *Handbook of Psychophysiology.* Cambridge University Press: New York.

Dindo, L. and Fowles, D.C. (2008). The skin conductance orienting response to semantic stimuli: Significance can be independent of arousal. *Psychophysiology, 45*(1), 111 – 118.

Elaad, E. & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection in the guilty knowledge test. *Psychophysiology, 28,* 163-171.

Ginton, A., (2009). Relevant Issue Gravity (RIG) strength—a new concept in PDD that reframes the notion of psychological set and the role of attention in CQT polygraph. *Polygraph, 38*(3), 204–217.

Honts, C.R., & Amato, S.L. (1999). *The Automated Polygraph Examination:* Final report of U.S. Government Contract No. 110224-1998-MO. Boise State University.

Honts, C. R., & Raskin, D. C. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration*,16, 56-61.

Honts, C.R., Raskin, D.C. and Kircher, J.C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology, 1,* 241 – 247.

Horowitz, S.W., Kircher, J.C., Honts, C.R., and Raskin, D.C. (1997). The role of comparison questions in physiological detection of deception. Psychophysiology, 34, 108 – 115.

Horvath, F. (1977). The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology, 62*(2), 127 – 136.

Horvath, F. (1988). The utility of control questions and the effect of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16* (3), 198 – 209.

Horvath, F., and Palmatier, J.J. (2008). Effect of two types of control questions and two question formats on the outcomes of polygraph examinations. *Journal of Forensic Science, 53*(4), 889 – 899.

Horvath, F.S., and Reid, J.E. (1972). The polygraph silent answer test. *Journal of Criminal Law and Criminology and Police Science, 63*(2), 285 – 293.

Hovland, C.I. and Riesen, A.H. (1940). Magnitude of galvanic and vasomotore response as a function of stimulus intensity. *Journal of General Psychology, 23*, 103 – 121.

Katkin, E.S. (2003, Feb). *Final Report on Project on an Examination of Response Parameters of Electrodermal Responding to Standard Stimuli.* Report No. DoDPI03-R-0005. State University of New York at Stony Brook.

Khan, J., Nelson, R., and Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph, 38*(3), 184 – 197.

Krapohl, D.J., and Goodson, W. (2015). Decision accuracy for the Relevant-Irrelevant Screening Test: Influence of an algorithm on human decision-making. *European Polygraph, 9*(4), 189 – 208.

Krapohl, D.J., Gordon, N., and Lombardi, C. (2008). Accuracy demonstration of the Horizontal System using field cases conducted with the Federal Zone Comparison Technique. *Polygraph, 37*(4), 263 – 268.

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

Krapohl D., Rosales T. (2014): Decision accuracy for the Relevant-Irrelevant Screening Test: A partial replication. *Polygraph, 41*(1), 20–29.

Krapohl, D.J., and Shaw, P. (2015). *Fundamentals of Polygraph Practice.* Academic Press: San Diego, CA.

Lole, L, Gonsalvez, C.J., Blaszczynski, A., and Clarke, A.R. (2017). Electrodermal activity reliably captures physiological differences between wins and losses during gambling on electronic machines. *Psychophysiology, 49*, 154 – 163.

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

Nelson, R. (2015). Appendix B: Meta-analytic survey of validated polygraph techniques. In D. Krapohl and P. Shaw Fundamentals of Polygraph Practice. Academic Press: San Diego, CA.

Patrick, C.J., and Iacono, W.G. (1989). Psychopathy, threat, and polygraph test accuracy. *Journal of Applied Psychophysiology,* 74(2), 347-355.

Raskin, D.C., Kircher, J.C., Honts, C.R. and Horowitz, M.S. (1988, May). *A Study of the Validity of Polygraph Examinations in Criminal Investigation.* Final Report to the National Institute of Justice. Grant No. 85-IJ-CX-0040. University of Utah, Salt Lake City, UT.

Reid, J.E., and Inbau, F.E. (1977). *Truth and Deception: The Polygraph ("Lie-Detector") Technique, 2nd Ed.* Williams & Wilkins: Baltimore, MD.

A Discussion of PLC and DLC Question Procedure and Ironic Process Theory*

Raymond Nelson, Mark Handler, Rodolfo Prado and Ben Blalock

Abstract

Probable Lie and Directed Lie Comparison Questions are discussed within the framework of how they are used to generate appropriate levels of salience under the analytic theory of the polygraph. The authors provide a framework for the process of introducing both types of comparison question and discuss the concept of Ironic Process Theory as it applies to the Directed Lie Comparison Question. The authors discuss the Probable Lie process as a goal-oriented endeavor that requires manipulation of the examinee into denying commonplace transgressions and believing they must pass each question to pass the test. Finally, we offer examples of the introduction of both Directed and Probable Lie questions.

Try to pose for yourself this task: not to think of a polar bear, and you will see that the cursed thing will come to mind every minute.[Fyodor Dostoevsky, Winter Notes on Summer Impressions, 1863]

Introduction

Comparison questions are used in psychophysiological detection of deception (PDD) testing to provide a basis of recorded information to increase the objectivity, reliability, and reproducibility of analytic conclusion about deception or truth-telling in response to relevant test questions. Summers (1939) first described the use of what today is known as the comparison question test (CQT) format, consisting of a question sequence of three relevant question interspersed with three comparison question and three neutral questions, repeated three times – though some terminology differed from current usage. [Refer to Krapohl (1996) for a discussion about the evolution of terminology applied to these questions.] Use of the comparison question and comparison question test was promoted and popularized within the polygraph profession by Reid (1947), and others including Backster (1963) researchers at the University of Utah (Raskin & Hare, 1978; Kircher & Raskin, 1988; Bell, Raskin, Honts & Kircher, 1999), the U.S. Department of Defense (2006a), and the American Polygraph Association (2011).

Kubis (1962) first described the use of a Likert (1932) type integer scale to transform CQT data into numerical values. Use of numerical scoring and Likert type numerical transformations using a 7-position scale was promoted and popularized within the polygraph profession by Backster (1963), along with researchers at the University of Utah (Bell, Raskin, Honts & Kircher, 1999; Kircher & Raskin, 1988; Raskin & Hare, 1978). The 7-position scoring methods was later modified to become a more objective 3-position ordinal rank scoring method by Van Herk (1990) and the U.S. Department of Defense (2006b), and subsequently became the basis for the Empirical Scoring System (Nelson, 2017; Nelson, Krapohl & Handler, 2008; Nelson et. al., 2011).

The CQT differs from earlier PDD test formats in the use of a comparison question against which responses to relevant questions can be juxtaposed for analysis. A comparison question is a polygraph test question intended to provide innocent or truthful persons an opportunity to answer a greater response-inducing question in relation to the investigation target stimulus or relevant question. The analytic theory of PDD testing is that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception or truth-telling in response to relevant target stimuli. Although discussed here with regard to the CQT, this analytic theory can also be applied to the concealed information test (CIT), for which the different types of stimuli are the key-question and non-key questions, and also to the relevant-irrelevant test (RIT), for which the different types of stimuli are

*A portion of this content appeared in the APA Magazine 53.1 and is reprinted here with permission.

those to which a person may be deceptive or truthful. [See Nelson (2016) for a discussion of the analytic theory of the polygraph test.] Reactions to the test target stimuli can be compared with responses to comparison stimuli to calculate a statistical classifier for deception or truth-telling.

An important advantage of the CQT is that, in contrast to earlier test formats, it more readily accommodates some of the basic principles of scientific decision-making. One of those principles is the notion that all conclusions about the meaning of data from a scientific test or experiment are made with regard to other possible conclusions. Another important principle is that transformation of recorded data to numerical values, whether linear or non-parametric, can increase the objectivity and reliability of analytic conclusions compared to unstructured putative expert judgment. The purpose of any comparison question is to provide a basis of comparison that can support a more objective and reproducible numerical transformation and analysis of responses to relevant questions that describe the investigation target issues of a PDD examination. The CQT remains the most commonly used form of polygraph technique for both diagnostic exams – conducted in the context of a known allegation or incident – and screening exams – conducted in the absence of a known allegation or incident.

Two basic types of comparison questions are in use today: probable lie comparison (PLC) questions, and directed lie comparison (DLC) questions. For PLC questions the examinee is manipulated into answering *NO* where it is assumed that this verbal response is most probably incorrect. For DLC questions the examinee is instructed to answer *NO* though it is established and known that this answer is incorrect. Like all PDD questions, both PLC and DLC questions must be carefully reviewed during the PDD pretest interview. For reasons, both ethical and scientific, there are no un-reviewed questions during PDD testing. Of these two types, the PLC is subject to greater controversy due to their inherently manipulative use and presentation. Some examiners mistakenly believe that PLCs will be ineffective unless adapted or customized to the individual and case circumstances. This has led to expressions of concern and criticism among scientists about standardization and reliability (NRC, 2003). DLC questions are more easily standardized and may offer some advantages because their effective use is less reliant upon psychosocial manipulation and subjectivity. However, no published information suggests any significant difference between the effect sizes for the two types of comparison question[1]. [See Blalock, Nelson, Hander & Shaw (2011; 2012) for a discussion of the published literature on DLC questions.]

Neither DLC nor PLC questions should be misinterpreted as premised on an assumption that the polygraph measures lies *per se.* While past discussions about PLC questions have tended to emphasize emotion as a source of response, more recent discussion has centered on a plurality of factors that may contribute to responses to PDD test question – both relevant and comparison. These include emotion, cognition or mental activity, and behavioral conditioning. [Refer to Khan, Nelson Handler (2009) along Handler, Shaw and Gougler (2010) and Handler, Deichman, Kuczek, Hoffman and Nelson (2013) for further discussion about emotion and cognition in PDD testing.] It is neither possible nor necessary to know the exact emotion or exact cause of any emotion. In the same way, it is neither possible nor necessary to know the exact details of all mental activity and the various cognitive factors – which may include memory, attention, decision, novelty, and other factors – related to PDD test stimuli. Although we may want to know the details of an examinee's involvement in behaviors described by the relevant questions (RQs), it is similarly not necessary nor possible to know the exact details of a person's behavioral experience related to the comparison questions (CQs).

[1] There have been earlier discussions regarding "exclusive" or "non-exclusive" types of PLC questions. However, scientific studies have not supported the assumption of any real difference in effect-sizes from these two (Amsel, 1999; Honts & Reavy, 2009; Horvath & Palmatier, 2008; Horvath, 1988; Palmatier, 1991). Consequently, the discussion is moot for the exclusive CQ hypothesis, as it is not supported by evidence. Field practices have evolved to include both exclusive and non-exclusive CQs as indicated by individual circumstances. Discussion of PLC question herein includes both exclusive and non-exclusive types.

Polygraph Theory

To appreciate how polygraphs are intended to work, it is important to understand (1) the theory of the polygraph test, (2) relevant questions and their construction, and (3) the two types of comparison questions. The theory of PDD testing is premised on the fact that humans generate recordable physiological reactions to test stimuli – the RQs as well as to the CQs. The *analytic theory* of PDD testing is that greater changes in physiological activity are loaded at different types of test stimuli (i.e., RQs and CQs) as a function of deception or truth-telling in response to relevant target stimuli (Nelson, 2016). RQs refer to the investigation target issue, or topic of the polygraph examination, and are reviewed during the pretest interview. CQs – whether DLC or PLC – are questions regarding integrity and deception in general and must also be carefully reviewed and correctly introduced in order to produce the desired effect. Many decades of study have confirmed the practical value and validity of this analytical theory – having shown that responses are loaded sufficiently to permit probabilistic inferences about deception that exceed chance expectations. The *analytic theory* of the PDD and CQT is the same whether comparison questions are of the PLC or DLC variety, and the same despite differences in their introduction and usage.

It is assumed that all test subjects want to pass the polygraph test – that is their goal. The test questions function as a challenge to the test subject's goal of passing the test. The amount of mental effort required to answer a test question truthfully or correctly versus deceptively (or incorrectly) has been discussed as an underlying mechanism for the physiological reactions to test questions. In this model, cognitive activity is associated with the challenge to the examinee's goal of passing the test, and gives rise to the changes in physiological activity that are observed and recorded during PDD testing. Questions that require more mental activity (because of a deceptive or incorrect answer) will, in general, produce the larger physiological responses (Barland & Raskin, 1973; Craig, 1998; Day & Rourke, 1974; Kircher, 1983; Waid, Orne, Cook & Orne, 1978).

Relevant Questions

Effective use or selection of investigation targets and formulation of RQs is fundamental to the effectiveness of the CQT. RQs will describe the examinee's involvement in the specific behaviors under investigation. RQs should be clear, concise and behaviorally descriptive so that a truthful person is sure they are answering truthfully. Answering truthfully to RQs should require little mental effort. If an RQ is overly broad, a truthful person may engage in excessive or unwanted mental activity due to confusion or ambiguity, or due to uncertainty about whether their answer is correct or true. It is possible that ineffectively formulated RQs can induce physiological responses from truthful persons that may be quantitatively similar to those of deceptive persons.

This cognitive effort hypothesis assumes that persons who answer truthfully to the RQs will be required to engage in mental activity that differs quantitatively from persons who are engaging in deception. Although it is difficult to evaluate directly the qualitative or quantitative content of responses to RQs, quantitative differences can be observed more easily by comparing responses to RQs with responses to CQs. This model also includes the possibility that quantitative differences in CQs may occur as a function of deception or truth-telling in response to RQs – where the cognitive demands of the RQs draw mental resources away from the CQs. It is also likely that emotional and behavioral conditioning factors play a role in differential responses to PDD test stimuli. Regardless of the exact mechanism of difference, studies have supported the analytic theory of the polygraph for the greater part of a century, and the differential salience of the RQs and CQs can be inferred, coded numerically, and used to calculate a statistical classifier for deception or truth-telling. The process of introducing the CQs can be considered of equal importance with the selection and introduction of RQs.

PLC Questions

CQs are intended to generate mental activity. As with RQs, CQs may also invoke emotional content and responses due to behavioral experience. Correct use of CQs is fundamental to the effectiveness of the CQT. As mentioned earlier, there are currently two approaches to the preparation and presentation of the com-

parison question that are generally accepted. Comparison questions can be presented as either a PLC or as a DLC question (Raskin & Honts, 2002). As shown in Table 1, the preparation and presentation will differ for DLC and PLC questions.

In the PLC approach, the examinee is manipulated by the examiner during the pre-test interview into denying transgressions that are often topically similar to that addressed by the RQs. For example, if the relevant questions address a reported theft of object of value, a PLC comparison question might be, "Have you ever stolen anything from another person?" It is also common to use the general topics of *lying and dishonesty* as information or topical content for CQs. For example: "Have you ever told a serious lie to get out of trouble with people in authority?" Introduction of PLC questions is an area of great variability in field practice, because individual personality attributes can play a role in the ambiguous communication and manipulation that are central to these questions. The following is a short example of PLC introduction dialog.

> *Now these next questions I'm going to ask you are just as important as the things we have been discussing, if you want to pass this polygraph test. Because you cannot pass this test if you are not telling me the truth. And that means the complete truth. One-hundred percent truth. This polygraph cannot tell the difference between a white lie and a serious lie, and cannot tell the difference between a half-truth and a lie. Just as there is no such thing as sort-of pregnant or partially pregnant, you are either an honest and truthful person – in which case you must already have told me the complete truth – or you may be the kind of person who says you are honest and truthful even though you are not. So, my question is just this: are you an honest and truthful person? Not just kind of honest, like when it is convenient. But really honest and truthful?*
>
> *Now many people are not perfect in their honesty and integrity, and maybe they tell some lies. Maybe big lies. Serious lies. Maybe they do it a lot. Or, maybe they only lie sometimes – and they try to tell themselves they are an honest and truthful person, most of the time. But that is not honest and truthful and they know it. We all know it. Maybe they only tell small lies. Only to strangers, people that don't matter to them. But that is not honesty or truthfulness or integrity or trustworthiness. Maybe they only tell white lies, maybe only to people close to them. But that also is not honest or truthful. Anyone who would lie to the people close to them – to people that love and trust them – would lie to anyone. Of course, they would lie to get out of trouble, such as this situation. But you already told me you are not that kind of a person. So, unless you have been lying to me all along, these next questions should be super easy for you, because you already told me you don't lie to people who are trying to trust you, or people in authority, or to cover up something you have done wrong.*

PLC questions are reviewed with a demeanor of subtle or overt judgment towards others who have engaged in the activities described by the PLCs. Examiners will attempt to give a plausible reason for the inclusion of the CQs in the PDD testing procedure – though without discussion of the actual purpose of the PLC questions. Most examinees will want to convey a positive impression of themselves to the examiner – or they will at least cooperate superficially. The social dynamics of the PDD testing situation is used to discourage examinees from making admissions, and to maneuver them into an answer that is most likely untrue – a probable lie. Examinees can also be maneuvered to agree that a test should be developed so as to provide them an opportunity to show that they are a trust-worthy person for whom the alleged theft or crime is uncharacteristic or unlikely.

In the review of PLC questions, the examiner will emphasize to the examinee that *they must pass every question to pass the test* – often stating to the examinee that they will fail the test if they lie to any question. This is intended to create the dilemma for the truthful person – and a perceived barrier to passing the test – in that they have falsely denied the behavior described by the PLC question. For example: "Have you ever lied to anyone who was trying to trust you?" RQs are expected to be the greatest barrier to passing the test for persons engaging in deception – for reasons that may involve both cognition, emotion, and behavioral experience – whereas the PLC questions are merely a procedural aspect of the test. Done effectively, truthful examinees will be cognitively and emotionally uncertain about their answers to PLC questions. They may be aware of having somehow been tricked into lying to the comparison questions during the pretest interview – and will ideally believe they may fail the test because of their responses to the PLCs. The examiner will admonish the examinee about the PLC topic, and in doing so will admonish the examinee against making any admission. If the examinee has already admitted anything or assented to the PLC topic, the examiner will develop the admonishment to prevent any further admissions.

> *So, at the end of this polygraph test I think you and I are both hoping that we can tell them you are an honest and truthful person, and not a liar or cheat or criminal of any kind. Because if you are innocent, if you did not do this [investigation target issue]. Then the only thing standing in your way is if you are honest and truthful enough to pass a polygraph test. If you are then that is great. But if you are not – if you are the kind of person that might do this – well then that is one of the things they will want to know. So if you think there is a chance that you are not able to be completely honest with me about things like your own personal honesty and integrity – if you have done things that would indicate that you are the type of person that could lie or cheat or betray the trust of people close to you then we are going to need to talk and they are going to want to know what that is all about. What kind of person you really are. So, take a moment and think about this question and make sure you are not hiding something from me, not trying to hide something from yourself. Can you answer this question? Truthfully? Completely truthful. About whether you [besides that what you told me], have ever told a lie, a serious lie, to someone who loved or trusted you?*

These example PLC dialogs are intended to be ambiguous and judgmental and psychologically over-bearing. They are intended to constrain the examinee into a desired answer, *NO*, while overtly appearing to give information and request information. In most interviewing contexts – including counseling, coaching, information gathering, consultation, leadership, investigation, and even interrogation – this manner of communication would be regarded as disrespectful and also ineffective at facilitating the exchange of accurate or real information. However, in the PLC context, accurate and real information is not the objective. Instead, the objective is to obtain only superficial compliance from the examinee – and to pretend to interpret this for rapport. The examinee will ideally be aware of the fact they are being inauthentic, aware that the polygraph examiner has been arbitrary and judgmental and insincere in their manner of questioning, uncertain about whether their answer to the PLC question will be satisfactory to pass the polygraph test, and conflicted emotionally and cognitively as to their choice of solution to the dilemma, and resigned to a posture of superficial cooperation instead of attempted discourse or other resolution. Examiners will accept the examinee's superficial compliance as a form of acceptable rapport, and at no point in the process will the examiner use the word comparison or indicate the use of a PLC question.

Truthful persons should, ideally, know they are telling the truth in response to the narrowly constructed and behaviorally descriptive RQs. As before, the mental activity or cognitive

loading hypothesis holds that truthful persons will be more focused on, and will engage in increased mental activity, when responding to CQs than when answering truthfully to the RQs. Deceptive persons are expected to focus greater attention on, and exert more mental effort in response to, the RQs and the need to appear truthful than to the CQs. This is because – having engaged in the investigation target behavior as described – the RQs present the most substantial barrier to passing the test.

A traditional explanation among polygraph field practitioners in the past was the "psychological set" hypothesis (Matte & Grove, 2001) – citing Ruch (1948) – which attributed differences in physiological reactions to RQs and CQs to anxiety, apprehension and perceived threat to survival and well-being, subsumed as a fear of detection and consequences. A number of problems surround this hypothesis, beginning with the fact that the concept does not appear in the psychological literature as it was used within the polygraph profession. Handler and Nelson (2007) and Senter, Weather, Krapohl and Horvath (2010) pointed out that the narrow definition precluded this hypothesis from integrating the variety of hypotheses that have been proposed by psychologists and psychophysiologists, and suggested the more general term *salience* as a more inclusive alternative.

Another burden of liability for the psychological set hypothesis is the troublesome citation of Ruch by Matte and Grove, for which Handler and Nelson found that neither the term nor a description of the hypothesis appear in the textbook. It is therefore not surprising that scientists have referred to some polygraph terminology as "Alice-in-wonderland-vocabulary" (Furedy, 1991). Even more concerning is that use of well-defined psychiatric terms such as anxiety, along with highly subjective concepts such as apprehension create a non-trivial, and possibly insurmountable, barrier in satisfying the generally held requirement that a hypothesis must be falsifiable to be held as scientific (Popper, 1959). Finally, there may be plausible reasons as to why an innocent person may experience greater fear of RQs than CQs. Some effort has gone towards the development of metaphors other than fear to help explain differences to RQs and CQs[2].

Despite the shortcomings of traditional PLC theory, persons who are engaging in deception in response to RQs are expected to produce greater changes in physiological activity to the relevant than to comparison questions (Offe & Offe, 2007), while truthful persons are expected to produce greater changes in physiological activity in response to PLC questions. Due to the complex social dynamics – which can involve a combination of education, training,

[2] An example of this is Ginton (2009) who reframed the fear of consequences metaphor as the "relative issue gravity" metaphor, for which attempts to explain the differences in response to RQs and CQs as a function of the gravity of the RQs, in a manner that may be consistent with results described by Offe & Offe (2007). In this usage, gravity can be regarded as an attraction force acting on the attention of the examinee, and can also be thought of as the seriousness of the relevant target issue. Although potentially useful as a discussion metaphor, this reframe of the "psychological set: hypothesis may not resolve the problem that RQs may present more gravity to innocent persons than CQs, and may still fall short of illuminating the underlying psychological processes. It is likely that a variety of mechanisms will continue to play a role in the discussion of responses to PDD test stimuli.

[3] Concern has also been expressed toward the traditional fear-hypothesis for PLC questions, which emphasized emotion and fear of detection and consequences as the basis of response. Problems with this older hypothesis are numerous, and included the fact there are plausible reasons why a truthful or innocent person may experience greater fear of RQs than CQs. More importantly, PDD sensors and signal processing methods cannot differentiate the emotion of fear from anger, disgust or other strong emotions. Although emotions of different types may be a factor in responses, PDD technology cannot determine the reasons for an emotional experience or response. The traditional fear-hypotheses is also problematic because it ignores the role of cognition and behavioral conditioning, focuses solely on emotion as the basis of response, relies on problematic use of the psychiatric term anxiety, relies heavily on the subjective experience of apprehension, and conveys an impression that PDD questions and PDD testing equate to a threat to an examinee's survival. Also, the fear hypothesis cannot adequately account for the similar effect sizes for PLC and DLC questions – prompting a need to update the working theory in lieu of rejecting or ignoring empirical evidence. Although perhaps useful at the time it was introduced, this older hypothesis no longer provides a satisfactory understanding or explanation for PDD testing. Contemporary PDD theory emphasizes a plurality of factors, including emotion, cognition and behavioral experience, emphasizes a testable and falsifiable statement of the effects that are expected to be observable in recorded test data, and attempts a discussion that can accommodate empirical evidence suggesting similar effect sizes for PLC and DLC questions.

and mission priorities for professionals, in addition to the potential for mental health, level of functioning, and developmental considerations for examinees – and the ethics of manipulation when using PLC questions, some scientists and credibility assessment experts have expressed concern about PLC questions[3]. Raskin and Honts (2002) suggested that DLC questions were developed, in part, as an effective alternative to the PLC for these reasons. Despite these concerns, there is a substantial body of laboratory and field research that supports the validity of the CQT with both DLC and PLC questions[4].

DLC Questions

In the DLC approach, the examiner instructs the person to answer *NO* to the CQs [5]. The subject is told that it is important for the examiner to observe the normally expected physiological responses to the DLCs, otherwise the test will be inconclusive (Raskin & Honts, 2002), which will mean that they won't pass the test. A note here: many people will simplistically accept that answering *NO* to a DLC is a lie. Although the *NO*, answer is incorrect, it is not actually a lie. A lie is an attempt to deceive another, to convince another to believe some statement or information that is factually inconsistent with reality. Examinees answer *NO* to DLC questions because they are instructed to do so. The DLC question is simply a procedure – for which the name of the procedure is *directed-lie* – used to elicit physiological responses for comparison with physiological responses to the RQs. Notwithstanding this philosophical and epistemological nuance, examiners will, as part of the DLC procedure, explain to the examinee that, in fact, DLC questions are equally as important to the test results as are the questions about the investigation target issue. Examinees will increase their risk of not passing the test if they do not attend to and respond properly to the DLC questions.

One of the advantages of DLC questions is that the PDD test can be conducted in a factual and straightforward manner. Another advantage of the DLC approach is the potential for greater standardization – they are less reliant upon psychological manipulation and individual personality than PLC questions. However, examiner skill in understanding and using DLC questions will remain an important factor. DLC questions must be introduced correctly, or truthful and innocent examinees may easily fail to appreciate how important these questions are to their test results.

From a practical perspective, and for standardization, it is sometimes useful to organize the use of DLCs into a coherent process. Without a coherent and organized understanding of the DLC process there may be an increased risk that ineffective adaptations of the DLC

Table 1. Outline of the DLC and PLC processes.

	DLC	PLC
1.	Introduce and normalize the DLC topic	Introduce and stigmatize the PLC topic
2.	Obtain the examinee's endorsement and assent to some experience with the DLC topic	Tell the examinee they must pass every question to pass the test.
3.	Advise the examinee about the topic	Admonish the examinee about the topic
4.	Instruct the examinee to answer NO	Manipulate the examinee into denial
5.	Review the question and practice the answer	Review the question and practice the answer
6.	Repeat steps 1-5 for each DLC question	Repeat steps 1-5 for each PLC question
7.	Further explain the need for the DLC questions	Admonish the examinee further about honesty and integrity

[4] Refer to the meta-analytic survey of validated polygraph techniques (American Polygraph Association, 2011) for more information. Also, see Offe and Offe (2007) for more information and the results of an experimental test of the analytic theory of the CQT.

[5] Refer to Menges (2004) for a brief history on DLC questions.

question may contribute to problematic CQT outcomes. Nevertheless, while structure and organization are important, effective use of both DLC and PLC questions may require that the process be executed in a fluid and natural dialogue and not as a robotic or mechanized step-by-step procedure. Tavle 1 shows an outline of the basic processes for introducing DLC questions, in parallel with a process for PLC questions. Again, published studies have failed to show significant differences in effect sizes for PLC and DLC questions (Blalock Nelson, Handler and Shaw, 2011; 2012; Honts & Reavy, 2009).

Both DLC and PLC questions begin with an introduction of a question topic about a category of behavior related to honesty and integrity, or related in a general way to the investigation target issue. Both procedures review the question topic with a defined objective – to normalize (DLC) or stigmatize (PLC) the topic. DLC questions seek assent or endorsement of the topic, while PLC questions seek denial or avoidance of the topic. DLC questions rely on simple and clear instructions to answer incorrectly, whereas PLC questions rely on psychological manipulation to solicit a response that is assumed to be incorrect (i.e., a probable lie). Procedures for both the DLC and PLC question include a review of the exact language of the question that will be asked during PDD test data collection, and both procedures include a review, prior to the onset of the data collection, of the examinee's intended verbal response. Both DLC and PLC questions are formulated as closed questions – soliciting a *NO* answer – and require no other talking or discussion during data collection.

DLC questions produce physiological reactions and effect sizes similar to PLC questions, and can be understood as subject to some of the same psychological factors as PLC questions. The mental effort hypothesis holds that persons who are innocent and truthful in response to the RQs and investigation target issues will devote attention and mental activity to the DLC questions because these may present the greatest barrier or challenge to their goal of passing the test. As with PLC questions, it is likely that emotion, behavioral experience, memory, and possibly that a variety of other psychological factors including the orienting response (Sokolov, Spinks. Naatanen, & Lyytinen, 2002), conditioned responses (Davis, 1961), arousal theory (Ben-Shakur, Lieblich & Kugelmass, 1970) and cognitive dissonance[6] and other hypothesis may all play some role in psychophysiological response to DLC questions.

Regardless of whether field practitioners use DLC or PLC questions, we can expect that any unidimensional theory or hypothesis may be inadequate to explain the complexities of human psychology and physiology. Instead, it is more likely that an integration of various psychological theories and perspectives may be useful to more completely and more adequately understand the correlation between observable and recordable physiological reactions and PDD test questions.

Ironic process theory and DLC questions

Ironic process theory (IPT) refers to a psychological phenomenon wherein deliberate attempts to suppress or avoid certain thoughts (Wegner, 1989; 1994; 2009; Wegner & Schneider, 2003; Wegner, Schneider, Carter & White, 1987; Wenzlaff & Wegner; 2000) or emotions (Gross & Levenson, 1993; Geraerts, Merckelbach, Jelicic & Smeets, 2006) can induce the paradoxical or ironic effect. This results in increasing their occurrence or causing a person to become more immediately or acutely aware of those thoughts or emotions. In short, researchers have shown that although it may be possible to suppress emotion, as recorded through physiological activity associated with autonomic activity, under some conditions, the effects of suppression are reduced as a function of cognitive loading or mental activity. In the PDD testing context mental activity is induced by the need to attend to and respond deceptively to relevant test questions.

IPT may provide interesting insight on the results of a study conducted in the middle east wherein DLC questions worked well with examinees who were also polygraph examiners (Nelson, Handler, Blalock & Hernandez, 2012). IPT is potentially useful to PDD examiners in

[6] See Handler and Nelson (2012) for an introduction to cognitive dissonance and its application to the polygraph context.

that it can help to understand and formulate an approach to DLC question formulation that requires neither overt psychological manipulation nor intrusion beyond the scope of the investigation target or referral issue.

The following is an example dialogue for each stage of the DLC process. An outside observer might be struck by the transparent nature of this procedure.

1. Introduce and normalize a DLC topic.

> *Now it's important to listen carefully, because this is something that is quite normal. People are only human, and that means all people are imperfect. People make mistakes. People make errors. Most normal people have made a mistake or error, and most of the time they take responsibility and fix it. But most normal people have also had some situation in which they may have made a mistake or an error and then kept it secret, or maybe they even told a lie about it. It's unfortunately common. People do the best they are capable of, and it's sometimes not perfect. If you are like most people, then you may have had this kind of situation. Most people, including those who are honest and truthful, have had such an experience where they had committed some mistake or error, and then, instead of accepting responsibility for the situation, they may have kept it a secret, or maybe even told a lie about it. Now perhaps this was as a young person, or, quite often, even as an adult.*

The first objective, when introducing a DLC question, is to *normalize* the DLC topic for the examinee. It is important to note that to normalize a DLC topic is not to *trivialize* its importance on the exam. It is also important that examiners convey the notion that DLC question are *equal in importance* with RQs. Most importantly, each DLC topic must be introduced without criticism or stigma. A good strategy is to use common and comfortable language while introducing the topic, while being careful to avoid expression of disapprobation or reproach.

One way to increase the social comfort of an examinee while introducing a topic of potential discomfort is to make careful use of platitudes. Platitudes are superficial statements, often meaningless and factually unnecessary, that convey no actual message or information. Social platitudes can have the effect of increasing interpersonal comfort by stopping and replacing other, potentially more authentic, thoughts and communication. For example: the phrase *"Hi, how are you"* can be used as a friendly greeting, for which a common response is *"fine, and how are you,"* with little actual interest in the details of each other's recent experiences. Platitudes allow people to interact in socially comfortable ways. Social platitudes can be useful because they allow people to greet each other and make contact in a friendly manner that also allows people to anticipate the quality and context of the ensuing interaction. In the context of introducing a DLC question, in the sample dialog above, phrases such as *"like most people"* are intended to reduce stigma, normalize the topic, and increase the comfort of the examinee. Done effectively, the topic will be introduced in a manner that does not prompt the examinee to assume a posture of denial or avoidance toward the DLC topic. Done effectively, the examinee will assent and endorse the DLC topic.

DLC topics are often related to integrity and deception. Following are some examples of DLC topics related to integrity and honesty.

- Secrecy or dishonesty to hide or avoid responsibility for a mistake or error.
- Secrecy or dishonesty to avoid responsibility for violating rules or regulations.
- Secrecy or dishonesty to avoid shame or embarrassment.
- Secrecy or dishonesty to impress others or make yourself look better.

The use of other DLC topics is also possible and DLC topics can sometimes be quite similar to PLC topics.

- Telling lies to people who loved or trusted you.
- Telling lies to a family member or friend.
- Telling lies to anyone in a position of authority.
- Telling lies to avoid consequences.

2. Obtain the examinee's assent and endorsement.

> *What I want you to do is just to think carefully about your past, your entire lifetime. You, like all people must have a lot of experiences. Some great, some not so great. So now, just tell me if you have ever had that type of experience. Have you ever done that or had that type of situation?*

If the DLC topic is introduced and normalized in the correct way – using comfortable words, comfortable language, and comfortable platitudes – the examinee will answer *YES,* or will indicate their assent or endorsement in some manner. If an examinee does not endorse a DLC topic, it is often best to simply discard the planned DLC and select an alternative topic. It is sometimes useful for field practitioners to have a short list of planned DLC topics along with another short list of alternative DLC topics to use in the event that an examinee will not endorse one or more of the intended DLC topics. The following is an example of a list of alternative DLC topics.

- Ever making errors or mistakes.
- Violation of traffic laws (may not be useful in some cities or locales).
- Being disloyal to anyone.
- Engaging in lies or deception.
- Disappointing anyone.

Alternate DLC topics may tend to be even more commonplace and simple than planned DLC topics. When an examinee does not endorse a second, alternate, DLC topic it may be an indication of other problems – possibly indicating an examinee who is deceptive or intends on nothing more than superficial compliance with the PDD testing process, and also possibly indicating that an examiner has been ineffective at comfortably normalizing a DLC topic. Regardless, in this case, selection of a PLC testing strategy may be more effective than continuing to attempt to work with DLC topics.

3. Advise the examinee about the DLC topic.

> *Now listen, whatever that was that happened, whatever you did, I do not need you tell me <u>exactly</u> what it was, or <u>exactly</u> who was involved. I don't need you to tell me <u>exactly</u> what your reasons were, or even <u>exactly</u> what the situation was. All of that is not what this test is all about.*

The objective at this point is to bring some information about that memory or past behavior more prominently into the examinee's attention and awareness. However, it is neither necessary nor desirable to attempt to verify some memory of past behavior by soliciting the details. Having endorsed the DLC topic, examinees have already acknowledged some memory, whether vague or explicit, of some past behavior or incident that is consistent with the DLC topic.

In this example DLC dialog, you can observe the use of IPT in that subtle emphasis is given to the word *exactly*. The overt content of the communication conveys that the examinee should not provide the details regarding the behavior and statements at the time, other persons involved, context or situation, and motivation. However, IPT makes use of the fact that people generally know what serious faults, transgressions, and shameful or embarrassing details they would prefer never to reveal to others – especially strangers, colleagues and professionals who may exercise some form of judgment. For example, telling a person *"it is not necessary for you to tell me your most personally embarrassing and shameful secret"* can produce the paradoxical or ironic effect of alerting or prompting their working memory

to some awareness of details that were previously compartmentalized out of conscious awareness. Discussion of this type, using the principles of IPT, can prompt an examinee to recall important details, can do so in a manner that maintains the personal privacy and dignity of the examinee, and is not intrusive into personal issues that are outside the scope of a required investigation. Moreover, IPT allows us to begin to rely on unstated information as a basis of response to DLC questions.

4. Instruct the examinee to answer *NO*.

> *It's important that you to listen carefully to this question, and make sure you answer NO. Do you understand? OK, let's practice this question…*

A strategy that is sometimes useful is to instruct the examinee to *listen carefully or think carefully* at each stage of the process. This must be done skillfully to avoid adopting an authoritarian demeanor prior to PDD test data acquisition. Done effectively, it can convey importance and increase an examinee's awareness of the need to listen carefully, and may also increase the attention and conscious awareness of examinee's who may have intended on not listening carefully as a form of strategic faking.

As discussed earlier, answering *NO* to a DLC question may be incorrect, but is not, in an epistemological sense, an act of deception. In other words, the examinee is not attempting to deceive the examiner when answering *NO* as instructed. Said differently, a DLC question, and the examinee's responses to a DLC question is a *procedure*. Answering *NO* to a DLC question is *incorrect*, but is not actually a *lie*. This is in no way problematic, because the polygraph does not measure or detect lies *per se*. Polygraph, like other scientific tests, measures and quantifies responses to test stimuli and enable us to make categorical classifications based on probabilistic inferences and correlations. Regardless of this nuance, if the examinee refers to a *NO* answer to a DLC question as a lie it will be a convenience to accept the examinee's usage, - perhaps by responding *"exactly like that"* - and proceed without admonishing or correcting this detail.

5. Review the question and practice the answer.

> *As an adult, did you ever make a mistake and then keep a secret or tell a lie about it? (NO)*

Taking the time to carefully introduce each DLC question will ensure that examinees who are truthful or innocent, and who wish to cooperate, are prepared to understand and participate correctly in PDD testing. Although it may be possible to simply read a DLC question and instruct the examinee to answer NO in a matter of a few seconds, short-cutting the introduction of DLC questions may increase the likelihood of problematic testing outcomes, including an unknown increase in potential for inconclusive results as well as for false-positive or false-negative error. A carefully developed understanding of DLC questions and DLC procedure is among the most important ways to maximize the effectiveness of the polygraph test. Additionally, beginning each DLC similarly (and different from the RQ) may help the examinee more quickly recognize the question as a one to which they must respond incorrectly. This can be accomplished using time bars on DLC questions that differ from those of the RQs, and can also be done by strategically using phrases such as "did you…" or "did you ever…" for RQs and "have you ever…" for CQs.

6. Repeat steps 1-5 separately for each DLC question.

> *Pay careful attention now, because we are going to review another question similar to this one, but just slightly different in focus.*

It may be tempting to review and practice all DLC questions together, thereby relieving the burden of making a separate introduction, soliciting a separate endorsement and providing a separate instruction for each DLC. *However, this is not advisable.* The process of introducing DLC questions should be considered equally as important as the topic and content of the DLC questions and verbal answers to DLC questions. DLC questions themselves are unlikely to replace the role and importance of the polygraph examiner in assuring that each examinee correctly understands the content

and the importance of the DLC questions. Taking the time to carefully (lather, rinse and repeat) introduce each DLC topic and question will provide an opportunity for an examiner to convey the importance of these questions, in the same way that carefully reviewing the RQs will ensure that examinees with understand and respond correctly to the topic or target of the investigation.

7. Further explain the need for these questions.

> *The reason I will ask you these questions is this: I want to see what your body does when you answer these questions. I want to know that your body is capable of reacting correctly when you lie to those other questions. I want to know that you will react. If you your body doesn't react to those questions, you could possibly lie to the questions about the (relevant issue) and remain un-noticed, and that would be a problem. Now, if for some reason your body cannot react correctly to these questions, then that could be a problem for you because you are going to have an inconclusive test. If you are telling the truth today then you do not want an inconclusive test, because that is not a passed test. If you are telling the truth about (relevant issue) then I want you to have the best results possible. So, I want to observe and record what happens and how you react when you answer these questions incorrectly. So, listen carefully to each of these questions, and answer just the way we have discussed and practiced. It is not necessary to make your body do anything. Just make sure that you listen carefully to every question and answer 'NO' just the way we have discussed. Whether you are telling the truth or lying your body will do what it is supposed to do. Do you understand?*

PDD professionals who fully understand this process will execute it in a natural and fluid dialog – without explicitly emphasizing or conveying the structure and organization. This will be the most effective way to encourage truthful and innocent persons to cooperate authentically during the test. An overly mechanized or step-by-step execution can become problematic in that it may encourage a similar step-by-step form of participation and response during testing, and this may contribute to the appearance of unnatural, inauthentic, or feigned behavioral responses during testing. Although persons engaging in deception can often be expected to participate in ways that are superficially cooperative, the goal of the examiner will be to engage the examinee in a natural, though planned, dialog that will enable truthful and innocent persons to participate in a cooperative and natural manner during the recording and acquisition of PDD test data.

An easily avoidable failure mode can occur when an examiner has become bored with the process (as if polygraph work could ever become boring) and introduces the DLC questions while conveying the notion that they are not important and therefore deserve little time and attention. This is not limited to DLC questions, and can occur for all types of PDD questions. *The caution against professional boredom cannot be overstated.*

Some truthful and innocent examinees, because of the stress and acuity of the PDD examination, can have a heightened sense of social vigilance and may take notice of professional boredom and rote behavior. This may inadvertently lead them to conclude that DLC questions are not important. Competent expert PDD professionals will remain interested in their work, including each exam, each examinee, each step in the process and each question. For this reason, as discussed above, we consider it best practice to introduce and review each DLC question individually, carefully attending to the objectives of the DLC process at each stage. The end goal of the examiner will be to engage the examinee in attending to each of the DLC questions.

An important aspect of effective interviewing is the rapport or connection between persons. It is important to keep the dialogue and discus-

sion natural and fluid throughout the introduction of the DLC questions – and throughout the pretest interview. Examiners should strive to avoid a rigidly scripted presentation that can easily telegraph the fact that one is *interviewing-on-autopilot.* Interviewers who *talk past, to, at or above* others will be at risk for misinterpreting superficial compliance for rapport and will inevitably be less effective at than those who talk and listen *with* others.

Summary and discussion

We have provided a discussion of both PLC and DLC questions, along with a parallel process outline for the two types of CQs. We have also briefly discussed the analytic theory of the PDD test, and a number of psychological discussions that can be applicable to both PLC and DLC questions. In addition, we introduce Ironic Process Theory as it applies to the PDD test and the introduction of DLC questions. Both PLC and DLC questions require skill, training and some experience.

It is likely that both PLC and DLC questions will continue to be used in polygraph field practice, long into the future. However, DLC questions are easier to learn and easier to standardize because they are less dependent on individual personality or examiner subjectivity as variables that may influence their effective use. DLCs have been used in field polygraph examinations for almost 60 years. Numerous studies have shown the effectiveness of DLC questions in different languages and cultures involving almost two dozen researchers and scientists, including both laboratory and field studies, and involving everything from multiple issue screening topics to the most serious crimes in society. PLC questions, though well established in both polygraph research and field practice, will ultimately have greater vulnerability to potential problems such as *stomping-the-CQs* wherein a test may be loaded for truthful outcomes by over-discussion or over-emphasis on PLC content. Additionally, they may be vulnerable to criticisms of *soft-selling* PLCs wherein a test may be loaded to produce deceptive results, especially when an examiner uses the polygraph only as a *pro forma* for interrogation and only briefly discusses the PLC questions. DLC questions, because they are less dependent upon the influence of personality and persona on the tactical aspects of psychosocial manipulation, appear to offer greater potential for standardization – and there are favorable reports from polygraph field practitioners who make use of automation in both the pretest introduction and in-test presentation of DLC questions.

Although not discussed at great length, IPT may also be helpful to understanding the psychological basis of PLC questions, as well as to understanding other known phenomena during PDD interviewing and testing. Examples include why the discussion of breathing activity may contribute to increased problems with respiration data (Goodson et al., 2014) and why innocent persons who attempt countermeasure may increase their chances of producing test data that is interpreted as indicative of deception (Handler, Honts & Goodson, 2015; Nelson, 2015; NRC, 2003). IPT may also provide a mechanism for increased understanding of the responses of deceptive persons when answering relevant questions during PDD testing, but that is beyond the scope of this manuscript and may be a topic for another publication.

IPT is a simple theory with potentially simple mechanisms for usage in the PDD context, with practical application to DLC questions. We showed a series of sample dialogues for the process of introducing DLC questions using IPT, and provided explanations and rationale for effective usage along with points of caution about potential misunderstanding and misuse. These examples, and the related discussion and information are not intended to be taken as dogma, and should not be interpreted as intended to convey the only correct way to make effective use of DLC questions. There are, without doubt, other ways to introduce both DLC and PLC questions. Also, although it is sometimes necessary and helpful to discuss flaws in our applied theories and hypotheses as these become known, we do not, at this time, suggest that IPT or any theory is superior to other viable theories as applied to the PDD testing context. There are, without doubt, a variety of psychological theories that can be applied to PDD testing. It is our view that IPT is compatible with other operational theories such as mental-effort and goal-attainment, and that it can be useful to field

practitioners to add another layer of interesting discussion to our present understanding of both DLC and PLC questions.

The CQT remains the most commonly used form of polygraph technique for both diagnostic and screening polygraphs. Part of the reason for the prevalence of this is that CQT formats are easily amenable to numerical and statistical analysis methods that have served to make the polygraph test more objective and make polygraph test results more reliable and reproducible. Despite decades of use in both research and field practice, some confusion persists around both PLC and DLC field practices. The most obvious point of confusion is whether the polygraph records, measures, or detects lies *per se*; it does not. Another point of confusion and discussion has been the psychological basis of responses to RQs and CQs. It is our hope that discussion of DLC question, and this introduction to IPT, will be of some value.

All scientific tests are intended to quantify some phenomena of interest that cannot be subject to perfect deterministic observation or direct physical measurement. All scientific tests make use of proxy information that is correlated with those phenomena of interest, though not of itself the phenomena. Scientific tests are not expected to be infallible, and are only expected to quantify the level of confidence, margin of uncertainty, or strength of information in support of a conclusion or test result. Amenability of the polygraph test to reliable forms of analysis is, in large part, a function of the PLC and DLC questions. It is our hope that this manuscript may help to fill a gap in published information that reflects contemporary knowledge and contemporary PDD field practice with the CQT and DLC questions.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph, 40(4*), 196-305.

Amsel, T. T. (1999). Exclusive or nonexclusive comparison questions: A comparative field study. *Polygraph, 28*, 273-283.

Barland, G. H., & Raskin, D. C. (1973). Detection of deception. In W. F. Prokasy & D. C. Raskin (Eds.), *Electrodermal activity in psychological research* (pp. 417-477). New York: Academic Press.

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

Ben-Shakur, G., Lieblich, I. & Kugelmass, S. (1970). Guilty Knowledge Technique: application of signal detection measures. *Journal of Applied Psychology. 54(5)*, 409-413.

Blalock, B., Nelson, R., Handler, M. & Shaw, P. (2011). A position paper on the use of directed lie comparison questions in diagnostic and screening polygraphs. *Police Polygraph Digest, (2011)*, 2-5.

Blalock, B., Nelson, R., Handler, M. & Shaw, P. (2012). The empirical basis for the use of directed lie comparison questions in diagnostic and screening polygraphs. *APA Magazine, 45(1)*, 36- 39.

Craig, R.A. (1998). *The use of physiological measures to detect deception in juveniles: Possible cognitive developmental influences*. Dissertation Abstracts International: Section B: The Sciences and Engineering. 58(10-B).

Davis, R. C. (1961). Physiological responses as a means of evaluating information. In, D. Bidermman & H. Zimmer (Eds.) *Manipulation of Human Behavior.* Wiley.

Day, D. A., & Rourke, B. P. (1974). The role of attention in "lie-detection." *Canadian Journal of Behavioral Science, 6, 270-276.*

Department of Defense (2006a). *Federal psychophysiological detection of deception examiner handbook.* Reprinted in Polygraph, 40(1), 2-66.

Department of Defense (2006b). Test data analysis: *DoDPI numerical evaluation scoring system.* [Retrieved from http://www.antipolygraph.org on 3-31-2007].

Dostoyevsky, F. (1955). [1863]. Winter *notes on summer impressions.* New York: Criterion Books.

Furedy, J. (1991). Alice in Wonderland terminological usage in, and communicational concerns about, that peculiarly American technological flight of fancy. The CQT polygraph. *Integrative Physiological and Behavioral Science. 26(3)*, 241-247.

Gross, J. & Levenson, R. W. (1993). Emotional suppression: physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology, 64(6),* 970-986.

Geraerts, E., Merckelbach, H., Jelicic, M. & Smeets, E. (2006), Long term consequences of suppression of intrusive anxious thoughts and repressive coping. *Behaviour Research and Therapy, 44(10)*, 1451–1460.

Ginton, A. (2009). Relevant Issue Gravity (RIG) strength – a new concept in PDD that reframes the notion of psychological set and the role of attention in CQT polygraph. *Polygraph, 38(3)*, 204–217.

Goodson, W., Honts, C. R., Handler, M., Nelson, R. Hicks, M., & Westerman, D. (2014). Pre-test breathing instructions increase perceptions of respiratory countermeasures. *Polygraph, 43(4)*, 114-122.

Handler, M., Deitchman, G., Kuczek, T., Hoffman, S. & Nelson, R. (2013). Bridging emotion and cognition: a role for the prefrontal cortex in polygraph testing. *Polygraph, 42(1)*, 1-17.

Handler, M., Honts, C. R. & Goodson, W. (2015). A literature review of polygraph countermeasures and the comparison question technique. *Polygraph, 44(2)*, 129-139.

Handler, M. & Nelson, R. (2007). Polygraph terms for the 21st century. *Polygraph, 36(3)*, 157-164.

Handler, M. & Nelson, R. (2012). A Primer on cognitive dissonance and its application to polygraph testing. *Polygraph, 41(3)*, 170-175.

Handler, M., Shaw, P. & Gougler, M. (2010). Some thoughts about feelings: A study of the role ofcognition and emotion in polygraph testing. *Polygraph, 39*, 139-154.

Honts, C. R. & Reavy, R. (2015). The comparison question polygraph test: a contrast of methods and scoring. *Physiology and Behavior, 143*, 15-26.

Horvath, F. & Palmatier, J. (2008). Effect of two types of control questions and two question formats on the outcomes of polygraph examinations. *Journal of Forensic Sciences, 53(4),* 1- 11.

Kahn, J., Nelson, R. & Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph, 38,* 184-197.

Kircher, J.C. (1983). *Computerized decision-making and patterns of activation in the detection of deception.* Unpublished dissertation submitted to the faculty of the University of Utah in partial fulfillment of the requirements for the degree of doctor of philosophy.

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

Krapohl, D. (1996). Coming to terms with terms: control questions. *Polygraph, 25(3*), 243-245.

Krapohl, D. (2001). A brief rejoinder to Matte and Grove regarding "psychological set." *Polygraph, 30(3),* 203-205.

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations. RADC-TR 62-205, Contract AF 30(602)-2270.* Air Force Systems Command, U.S. Air Force, Griffiss Air Force Base. New York: Rome Air Development Center.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-55.

Matte, J. A. & Grove, R. N. (2001). Psychological set: its origin, theory, and application. *Polygraph, 30(3),* 196-202.

Menges, P. (2004). Directed lie comparison questions in polygraph examinations: history and methodology. *Polygraph, 33(3)*, 131-142.

National Research Council (2003). *The Polygraph and Lie Detection.* National Academies Press.

Nelson, R. (2015). Scientific basis for polygraph testing. *Polygraph 41(1)*, 21-61.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5)*, 69-82.

Nelson, R. (2017). Multinomial reference distributions for the Empirical Scoring System. Polygraph & Forensic Credibility Assessment, 46 (2). 81-115.

Nelson, R., Handler, M., Blalock, B. & Hernandez, N. (2012). Replication and extension study of Directed Lie Screening Tests: criterion validity with the seven and three position models and the Empirical Scoring System. *Polygraph, 41(3)*, 186-198.

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.

Offe, H. & Offe, S. (2007). The comparison question test: does it work and if so how? *Law and Human Behavior, 31,* 291-303.

Palmatier, J. J. (1991). *Analysis of two variations of control question polygraph testing utilizing exclusive and nonexclusive controls.* Unpublished doctoral dissertation.

Popper, K. R., (1959). T*he Logic of Scientific Discovery.* Routledge. [German version is currently in print by Mohr Siebeck.]

Prokasy, W. F. & Raskin, D. C. (1973). Electrodermal Activity in Psychological Research. Academic Press

Raskin, D. C. & Hare, R.D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15,* 126-136.

Raskin, D. C. & Honts, C. R. (2002). The comparison question test. In M. Kleiner (Ed.), *Handbook of Polygraph Testing.* San Diego: Academic Press, p.15-16.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology, 37,* 542-547. Reprinted in Polygraph 11, 17-21.

Ruch, F. L. (1948). *Psychology and Life.* Chicago: Scott Foresman.

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(2),* 109-117.

Sokolov, E. N., Spinks. J. A., Naatanen, R., & Lyytinen, H. (2002) *The Orienting Response in Information Processing.* Erlbaum Associates.

Summers, W. G. (1939). Science can get the confession. *Fordham Law Review, 8,* 334-354.

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.

Waid, W. M., Orne, E. C., Cook, M. R., & Orne, M. T. (1978). Effects of attention, as indexed by subsequent memory on electrodermal detection of information. Journal of Applied *Psychology, 63,* 728-733.

Wegner, D. M. (1989), White bears and other unwanted thoughts: Suppression, obsession, and the psychology of mental control. Viking/Penguin.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101(1)*, 34-52.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science 325(5936),* 48-50.

Wegner, D. M. & Schneider, D. J. (2003). The white bear story. *Psychological Inquiry, 14(3/4),* 326–329.

Wegner, D. M., Schneider, D. J., Carter, S. R. & White, Tl L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53(1),* 5–13.

Wenzlaff R. M., & Wegner, D. M. (2000) Thought suppression. *Annual Review of Psychology, 51,* 59-91.

Strategic Cognitive And Mobility Room (SCAMR)

Joseph R Stainback IV*

Edited by Stephen P. Ray, Hannah Mendel and Matthew Stevenson**

Paper Motivation and Context

In 2018 and 2019, the Department of Defense, under the Office of the Director of National Intelligence (ODNI), collaborated with their Intelligence Advanced Research Projects Activity (IARPA) program sought out novel methods to measure the performance of current and future credibility assessment techniques and technologies (eg Polygraph Examinations, Voice Stress Analysis, Ocular Methods, etc.). ODNI IARPA, in collaboration with Johns Hopkins University Applied Physics Laboratory, Booz Allen Hamilton and HEROx (a crowdsourcing firm) issued an open innovation in 2019 called Credibility Assessment Standardized Evaluation (CASE) Challenge. The CASE challenge was open to researchers and scientists around the world providing the opportunity for individuals and teams to earn prizes by creating and/or proposing methods that can further the research into the validity of credibility techniques. Subsequent to the close of the challenge, HEROx and IARPA announced twenty-seven (27) solutions were submitted from the invitation. After a competitive evaluation, three (3) champions were selected. This paper describes one of the selected techniques called Cognitive And Mobility Room (SCAMR). SCAMR is a theoretical concept and should be viewed within this paper in the context of a proposed solution.

Solution Abstract

The validity assessment methodologies are a problematic aspect of Credibility Assessment Techniques (CATs). The psychodynamics are challenging in analogue studies (those that employ a mock crime paradigm) as they may or may not duplicate emotions or autonomic arousal in field studies; or simply easily followed or replicated.1,5 The Solver (Dr. Joseph R Stainback IV) for this submission proposes a mental and physical adventure gaming approach within a well-constructed Escape Room concept that is deployed commercially.3,4 At the same time, elements of Red Teaming (good-bad guy scenarios), psychological clinical testing, and Synthetic Training Environments (STE) will be applied.2 The name of this proposed concept is Strategic Cognitive And Mobility Room (SCAMR). SCAMR is a comprehensive method of testing (MoT) CATs. The use of SCAMR will standardize and validate all current and proposed CATs by introducing physical adventure scenarios in which participants solve a series of puzzles, riddles, and/or clues. Participants employ teamwork, problem solving skills, and strategy to complete a predefined, real-life scenario. SCAMR will be advertised as an Escape Room concept with a win/lose component of play. A real-life theme is introduced that includes cognitive challenges such as solving problems, making decisions, etc., for a prescribed objective or set of objectives. Complex choices are introduced throughout the scenarios to encourage (or force) deceptive decisions. The credibility assessment component of the SCAMR concept will be introduced throughout the scenarios, but most likely at the end.

Index Terms – Credibility Assessments, Lie Detection, Polygraph Testing, Voice Stress Anal-

*Dr. Stainback is the sole proprietor of Applied Investigations, a DBA of Applied Systems Analysis, Inc. and can be reached at jrstainback@gmail.com

** University of Tennessee, Knoxville

ysis, Ocular Methods for Deception, IARPA, CASE Challenge, HeroX,

Solution Validity

The proposed SCAMR solution increases validity in testing CATs. The Solver theorizes that participants will be more willing to participate and possess motivation, jeopardy, and emotional arousal within an Escape Room scenario. This is especially important with the changing demographics of younger participants who are dedicated problem-solvers from video gaming and other computer-based challenges. Operational validity from this proposed solution will be confirmed by emulating Escape Room commercial game play which attracts willing participants who are mentally and physically capable of performing while removing error-prone factors of previous CAT studies (eg lack of jeopardy).[3,4,5,6]

Solution Background

The Solver of this proposed solution believes a CAT will only be achieved through a structured process to induce fear of telling a lie; e.g., "a person's fears, anxieties, and apprehensions are channeled toward the situation which holds greatest immediate threat to one's self-preservation or general well-being".[7,8] The basic assumption of the proposed SCAMR MoT is therefore based on the premise that all proposed, current, and future CATs work only if there is a process to induce fear (physiological response) of telling a lie. As such, a proposed MoT should preserve elements of "ground truth," jeopardy, motivation, and autonomic arousal from the testing of psychological responses to stress by questioning when there is a need to determine the ground truth of a situation. The Solver evaluated historical MoTs while reflecting on their own experience in polygraph training, Escape Room events, participation in military Red-Blue teaming scenarios, and scenarios based psychological testing to derive the SCAMR solution. It has been shown that decision making within intense gaming can induce autonomic responses.[9] Thus, if a CAT is administered within a gaming scenario, a well-designed solution like SCAMR can be used as a valid tool to test CATs.

Solution Design and Methodology

The steps for implementing this SCAMR MoT solution are explained in the following sections: Set-up and Alpha-Testing, Beta-Testing, Clinical Tests, Structured Experimental Design, and CAT Testing and SCAMR as a MoT Model Standard. The descriptions of these sections will be subject to change after an Institutional Review Board (IRB) review. The Solver of this submission is an independent research scientist. As a result, an IRB review will be required from an institution willing to review and approve this project.10 The Solver is affiliated with the University of Tennessee and Texas A&M University, both of which have IRBs. Furthermore, experimentation using COTs are subject to state laws (e.g., consent agreements) and general guidelines by the American Polygraph Association. These laws and guidelines will be factored into the IRB process. This process could take up to 3 months to complete.

Set-up and Alpha-Testing: SCAMR will require a small, dedicated facility or a set of offices. These offices will be configured to have video cameras (open and unobtrusive) fed to an adjacent "control-room" to capture the events (e.g., ground truth). Once a facility is acquired, an intense physical adventure scenario, emulating Escape Room concepts, STEs, and Red Teaming will be developed around the constraints of the acquired space. Initially, only one scenario developed as the quality of the scenario is critical to the success of the SCAMR MoT solution. A team of Subject Matter Experts (SME) will be recruited as consultants to develop the scenario and CAT testing points. These SMEs will come from the fields of Red Teaming (military), Polygraph or other CAT (e.g., Converus Eye-Detect) Testing, STEs, Experimental Psychology, and Escape Room Commercial Business and leadership training. The consultants will be paid a fixed fee stipend. White Team monitors, assistants, project coordinator (Solver), and Red Team participants will be selected among the SMEs during the Alpha-Testing only. White Team monitors are necessary to monitor videos, referee engagements, ensure participant safety, judge activities, resolve issues, and handle participant requests. Once the scenario is drafted, the SMEs will Alpha-Test and "act-

out" the scenario making note of all necessary facility modifications to make the scenario realistic. CATs will be simulated during the Alpha-Test. The work of the SMEs ends after the Alpha-Test of the scenario. Upon completion of the Alpha-Test, minor facility modifications of décor, furniture, and specialized fixtures may be necessary. Minor modifications to the offices may include reconfiguration of the doors, rooms, and entryways. The cost of this phase should be minor to the overall budget. Modifications and associated costs of the technologies can be throttled back by the injection of equipment associated with STEs. (e.g., Augmented Reality, 3D Simulators, etc.). However, SCAMR is intended to include mostly human-to-human interactions to invoke jeopardy, motivation, and autonomic arousal. Therefore, the use of STE technologies will be considered but not used as the primary driver of this solution.

Beta-Testing: Once the facility modifications are complete, a Beta-Test with a targeted set of participants (human subjects) will be required. One or two paid assistants will be necessary to assist the Solver in the Beta-Testing phase (coordination and White Team). The assistants can be acquired through a temporary employment agency. The participants can be recruited through advertising on a college campus or an internet-based advertising platform such as Craigslist. Participants will be persuaded by advertising as an Escape Room like experience. Upon recommendations from the SME team and the Alpha-Test results, CATs will become part of the Beta-Test. For the Beta-Test, standard Polygraph testing CATs will be utilized first with at least two (2) examiners available to perform a Polygraph test. The polygraph examiners will be paid at their normal examination rate. The Beta-Test participants will be provided with a typical clinical testing remuneration. After the Beta-Test. the participants will be interviewed to determine future changes to the scenarios and/or the facility. The scenarios will be repeated as necessary with the same or different participants to refine scenario and facility modifications.

Clinical Testing: After Beta-Testing and facility modifications, the SCAMR MoT will be ready for clinical testing. Similar to Beta-Testing, the set of participants for the clinical testing can be recruited through advertising within an internet-based advertising platform such as Craigslist. To avoid population bias, the Solver does not recommend recruitment on a college campus for the clinical testing. Participants will be persuaded again by advertising as an Escape Room-like experience. Depending on the IRB conditions relative to informed consent and the details of the scenario developed by the SMEs, each participant will be minimally informed to protect ground truth and/or the development of ground truth during the scenarios. For the first clinical test, like Beta-Testing, standard Polygraph testing of CATs will be utilized with at least two (2) examiners available to perform a Polygraph test. The Polygraph Examiners will be paid at their normal examination rate. The Beta-Test participants will be provided a typical clinical testing remuneration. The clinical test will be repeated as necessary with different participants depending on the structured experimental design.

Strawman Initial Scenario: The following scenario is an example that helps explain SCAMR MoT. It, therefore, does not reflect inputs from SMEs. It is discussed here to improve the explanation of the SCAMR process model. Emulating an Escape Room theme, "Espionage," is chosen for this strawman. Under 'Espionage', four (4) participants (human subjects) will be recruited from the public by advertising as described in the Alpha- and Beta-Testing phases. As shown in FIGURE 1, all participants will be briefed on the scenario and given confidential information relative to their goals in Room A, the "Scenario Briefing Room." Their goals are to solve problems/riddles/puzzles/codes (enigmas) to gain access (escape) from room to room to an ultimate prize. As with Escape Room commercial concepts, doors open between rooms when the enigmas are solved. There will be two monetary prizes, one small and one large. To arouse greed, only one participant (Red Participant) is told to go after both prizes and can choose to collaborate with anyone but must share the prize if they collaborate. Knowledge is power in the scenario and is intended to be used by the participants to deceive each other out of greed. The participants will be divided into two (2) groups of two (2) and instructed to go to their respective rooms (Team A to Room B and Team B to Room C). Each group will compete for the prize, but they

must work together in Room D after solving enigmas within their respective rooms. Each team holds separate information from their respective rooms in order to escape from Room D to Room E. The Red Participant may create a way to go to their competitor team's room to acquire information (espionage) for their own need to win the entire prize. Room F holds enigma-protected encasements for both prizes whereby the most knowledgeable person can 'open' the encasement to obtain one or both monetary prizes. The result depends on the knowledge and or deception used during the scenario. CATs deployed in Rooms F and G will be used to test the deceptive choices of the winners and losers. Scenarios are expected to be three (3) to four (4) hours in length.

Structured Experimental Design: Likely, the experimentation will follow a factorial (e.g., 2 x 3) mixed design. The between-subjects independent variables will be of two levels: known ground truth (verified by video) and deception 'acts.' These variables will be incremental through the scenarios depending on the levels and number of CAT interviews. Importantly, the SMEs should define and 'evoke' a specific 'deception' act during the scenario.11 The deception act will need psychological elements of regret, fear, guilt, remorse, etc. The participants will be polygraph tested on these specific acts of deception. Although ground truth will be absolute and verified through video or scenario design, testing of innocent participants may be necessary in order to prevent suspicion by the deceptive participant. The number of recommended scenarios can be throttled by the number of CATs required during a given scenario, however the number of scenarios should be sufficient for the experimental design statistical evaluation. The Solver expects an intense literature review of previous clinical testing of CATs in parallel with the budget constraints to ultimately determine the number of clinical tests.

CAT Testing and SCAMR as a MoT Model Standard: The initial testing of the proposed MoT will use the polygraph as the CAT. The Solver recommends this CAT initially because the Polygraph test is the current standard for measuring deception. After steady state clinical testing using this SCAMR MoT, subsequent MoTs can use any current CAT (e.g., Converus Eye-Detect, Voice Stress, etc.). More data collection (scenario testing) in a standard model results in a decrease of data variance. The Solver recommends involvement with the APA and associated experts in the field of CAT to "certify" the MoT model. Comparing CATs using SCAMR as a 'routine' MoT will be extremely interesting and beneficial to the credibility assessment community.

Replicability

SCAMR parallels Escape Room commercial concepts that have already been proven to be replicable. Using commercial Escape Room concepts, business proprietors have demonstrated replicability through the training of their employees and the standardization of their physical set-ups (the Solver acquired this information by participating in an Escape Room event). The SCAMR MoT model will be designed in the same manner with physical set-ups and White Team training. It will, however, also include written procedures, written scripts, and checklists. The Solver has experience in Red Teaming whereby replication was achieved by these methods.

Generalization

SCAMR can be generalized to any CAT as it is the outcome activity of the proposed solution. Under this solution, the act of testing participants using a CAT is independent of the mechanics of the scenario. The connection between the scenario and the CAT ensures activation of the body's sympathetic branch of the autonomic nervous system. This result occurs through competition to achieve a monetary prize while elements of greed, deception, and withholding of information evokes jeopardy, motivation, and autonomic arousal during questioning.

Ground Truth

Ground truth within the SCAMR scenarios is 'preserved' by scripting, the White Team, and

hidden video (unobtrusive) recordings. Participants will be tested on their greed, deception, and withholding of information to win the monetary prize of the scenario. In the 'Espionage' strawman example, one or more participants will be encouraged to deceive their 'peers' for individual achievement. The method of deception will be recorded on video within the SCAMR (Escape Room) environment. In essence, this solution is a controlled and closed analogue study whereby ground truth and deception to ground truth is observed continuously.

Psychological Realism

The psychological realism of SCAMR capitalizes on current attitudes toward game play and the desire to win, mirroring the popularity of Escape Room concepts. The Solver theorizes that participants will be more willing to participate and possess motivation, jeopardy, and emotional arousal within an Escape Room scenario. These responses are especially pertinent with the changing demographics of younger participants (personally meaningful) who are dedicated problem-solvers from video gaming and other computer-based challenges. The incentive of money as a "prize" differentiates SCAMR from ER, thus increasing the motivation to perform deceptive acts (consequences) to win.

Practicality

SCAMR is practical in that it parallels an existing commercial gaming concept, Escape Rooms. These parallels are scenario development, moderation (White Team), solving of enigmas, entering of rooms based on solved enigmas, and winning. Additional features will increase interest, making it more practical for implementation. Such features are STE (Augmented Reality), Red Teaming, elements of-deception, team play (competition), monetary incentives, and enhanced jeopardy

Procedure

Written scripts will be developed for the scenarios to proceduralize each event. The scenario designs will have flexibility to encourage creativity by the participants (eg to be deceptive during the scenarios) while keeping a general standardization. Proceduralized ground rules will be necessary for the safety of all participants. The credible and non-creditable participants will be distinguished between each other within the closed and controlled analogue room environment.

Motivation

By design, the researcher is hands-off within the SCAMR scenario. The Scenario Briefing (description, rules, individual instructions, and incentives) is provided at the beginning of 'play.' The participants choose their path as guided by solving enigmas as well as the desire to win money and defeat their competitors. The assumption is that SCAMR will work like Escape Room concepts and that participants will increase their chances of winning by deceiving others through the suggestions and scripting from the moderators. This assumption is sound given the success of Escape Room concepts and the theory of changing demographics of younger participants.

Enhanced Realism

The enhanced realism of SCAMR centers around the theory that participants are more willing to participate and possess motivation, jeopardy, and emotional arousal within an Escape Room scenario given its commercial success and especially with the changing demographics of younger participants who are dedicated problem-solvers from video gaming and other computer-based challenges.[3,4,5,6]

Technology

Experts in Synthetic Training Environment (STE) used in the military will be integral in making suggestions to technologically enhance SCAMR as a state-of-the-art MoT. However, increased technology is not central to the SCAMR solution. SCAMR scenarios are intended to encourage human-to-human interactions to invoke jeopardy, motivation, and autonomic arousal. Therefore, the use of STE technologies will be considered but not made into the primary driver of the solution.

Objective Measurement

This solution is a controlled and closed analogue study whereby ground truth and decep-

tion to ground truth is observed continuously by unobtrusive video. The testing rooms will be configured to have open-live video camera feeds in an adjacent "control-room" to capture the events. Ground truth is also preserved by scripting and White Team monitoring.

Beneficence

The experience and lessons learned from existing Escape Room enterprises will be utilized to ensure SCAMR staff and participant safety. Upon completion of the Alpha- and Beta-Testing, the rooms/facility will be reviewed for occupational safety and health considerations. During the Scenario Briefing (beginning) stage of the scenario, safety factors will be reviewed with the participants. In addition, White Team monitoring during the scenarios is intended to ensure that safety is always maintained. The SCAMR concept will adhere to the philosophy of "do no harm" while completing the scenarios.[12]

Respect for persons

The premise of SCAMR is to emulate Escape Room methodologies in that autonomy, courtesy, and respect during scenario play are central to motivating customers. After the Scenario Briefing, participants will be on their own to 'play' the game and will have the power to make their own decisions (Figure 1). Continuous monitoring by the White Team will ensure safety, respect for others, and that the integrity of the scenario is maintained. The Institutional Review Board (IRB) and state laws will define the appropriate informed consent for each participant.

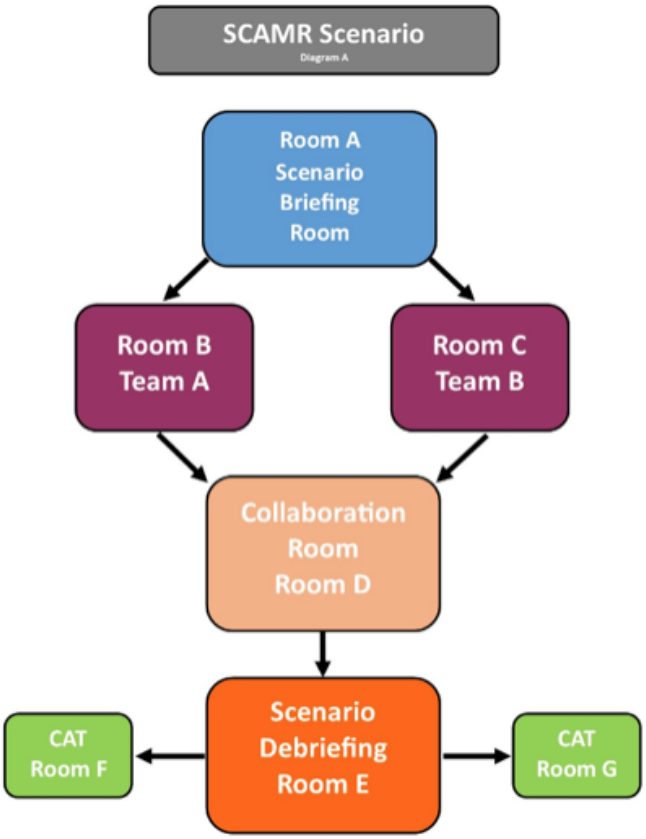
Justice

The participant selection for SCAMR will follow Ethical Principles and Guidelines for the Protection of Human Subjects of Research.[10,12] Thus, they will be unbiased and neutral. Thorough explanations of the scenario(s) including elements of CAT methodologies, targeted deception, solving of enigmas, and physical activities (moving from room to room) will be made clear to the participants. IRB approval conditions and state law will govern these factors of justice as appropriate.

Investment

SCAMR will require an investment into both the initial set-up (Alpha-, Beta-, and initial clinical testing) and the sustaining of testing additional CATs. Investments identified in this solution include, but are not limited to, facility leasing, facility modifications, minor facility modifications, décor, fixtures, furniture, specialized fixtures, administrative staffing including the White Team, CAT examiners to support CAT testing, nominal payments to participants, and monetary prizes for the winners These costs should not be insurmountable as observed through participation in commercial Escape Room activities.

Figure 1 SCAMR Process



Acknowledgments

Final Editing: Dr. Russ Hirst and his students in Global Communication for Science & Technology at UT Knoxville University of Tennessee
AUTHOR INFORMATION

Dr. Joseph R Stainback IV- Dr. Stainback is a Research Scientist and Polygraph Examiner performing domestic and international security education, training, and research with emphasis in human reliability, credibility and trustworthiness in the workplace; preventing aberrant behavior(s) through enhanced qualitative behavioral observations, data collection architectures, mobile / mHealth technologies including wearables and biosensors; and performing research in the causes and predictability of certain unusual / abnormal behaviors, weaknesses and disruptors within the physical, mental, and social well-being of the human state within the organizational construct.

Dr. Stainback has acquired over 30 years of direct operations experience including significant industrial security related programs and projects for the Department of Energy. Dr. Stainback also served as a Research Professor at the University of Tennessee focused on domestic and international nuclear security matters. Dr. Stainback works closely with US National Laboratories and universities within his research capacity. Dr. Stainback holds a B.S. degree in mechanical engineering technology from Old Dominion University, an M.S. degree in engineering administration from George Washington University, and a Ph.D. degree in industrial engineering from the University of Tennessee.

Copyright and Permission

Notes

[1] HEROx CASE Challenge Overview, 2019. https://www.herox.com/CASEchallenge

[2] Matt Leonard, *"Synthetic Training Environment, built to fuse the real world with the virtual world current live, virtual, constructive and gaming training environments"*; Accessed on: April 10, 2019. [Online]. Available: https://defensesystems.com/articles/2018/05/08/army-virtual-training-architecture.aspx

[3] Sally French, *"The unbelievably lucrative business of escape rooms"*; MarketWatch.com. Accessed on: April 12, 2019. [Online]. Available: https://www.marketwatch.com/story/the-weird-new-world-of-escape-room-businesses-2015-07-20

[4] Penttilä, Katriina. *"History of Escape Games : examined through real-life-and digital precursors and the production of Spygame"*, Acessed on: April 12, 2019. [Online]. Available: https://www.utupub.fi/handle/10024/145879

[5] Dean A. Pollina, Andrew B. Dollins, Stuart M. Senter, Donald J. Krapohl, and Andrew H. Ryan, *"Comparison of Polygraph Data Obtained From Individuals Involved in Mock Crimes and Actual Criminal Investigations"*, Department of Defense Polygraph Institute, Journal of Applied Psychology, Vol. 89, No. 6, 1099–1105, 2004

[6] Sarah Sladek, Josh Miller, *"Ready or Not Here Comes Z"*, XYZ University, January 2018.

[7] Stuart Senter, Dan Weatherman, Donald Krapohl, and Frank Horvath, *"Psychological Set or Differential Salience: A Proposal for Reconciling Theory and Terminology in Polygraph Testing"*, Polygraph, 39(2), 2010

[8] Mark Handler and Raymond Nelson, *"Polygraph Terms for the 21st Century"*, Polygraph, 36(3), 2007

[9] Eduardo Massad, Paulo Cesar Costa dos Santos, Armando Freitas da Rocha, Edward J. N. Stupple, The Monty Hall problem revisited: Autonomic arousal in an inverted version of the game, journals.plos.org, Accessed on: April 13, 2019. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192542

[10] Todd W Rice, How to Do Human-Subjects Research If You Do Not Have an Institutional Review Board, Respiratory Care, rc.rcjournal.com, Accessed on: April 13, 2019. [Online]. Available: http://www.rcjournal.com/contents/10.08/10.08.1362.pdf

[11] Alex W. Stedmon, Peter Eachus, Les Baillie, Huw Tallis, Richard Donkor, Robert Edlin-White, Robert Bracewell, "Scalable interrogation: Eliciting human pheromone responses to deception in a security interview setting", Applied Ergonomics, 47, 26e33, 2015.

[12] Department of Health, Education, and Welfare, Office of the Secretary, "The Belmont Report", 1979, Accessed on: April 13, 2019. [Online]. Available: https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf

Multinomial Cutscores for Bayesian Analysis with ESS and Three-Position Scores of Comparison Question Polygraph Tests

Raymond Nelson

Abstract

Multinomial reference distributions calculated under the analytic theory of the comparison question test and are available for both Empirical Scoring System and the Federal three-position scores. They are then used as a likelihood function for Bayesian analysis of the posterior strength of information for deception and truth-telling. Bayesian classification of comparison question test data is accomplished by using Bayes theorem, along with the test data, prior information and a statistical likelihood function, to calculate a posterior likelihood of deception or truth-telling and then quantifying the expected variability of the test data as a Bayesian credible interval. A classification of deception or truth-telling is supported when the strength of the 1-alpha lower-limit of a coverage interval has exceeded the strength of the prior information for deception or truth-telling. Field polygraph practitioners traditionally work with comparison question test data in the form of point scores and cutscores. Multinomial cutscores are the minimum scores for which strength of the posterior information exceeds the prior information with the uncertainty or expected variation reduced to the alpha tolerance level. However, until this time published multinomial cutscores have been available only the equal prior condition and only for the symmetrical alpha scheme of a = .05, .05 for deception and truth-telling. This project involved the tabular calculation of multinomial cutscores for the Empirical Scoring System and Federal three-position scoring methods for all permutations of alpha levels at .01, .05, and .10 for truth-telling and deception using a distribution of prior odds from one in 10 for truth-telling and deception. These cutscore tables permit polygraph field practitioners to make use of the advantages of Bayesian analysis while relying on the practical intuition of scores and cutscores, and without the need for the recalculation of Bayes theorem or Bayesian credible intervals. Multinomial cutscore tables are provided in appendices.

Introduction

Multinomial reference distributions were calculated for comparison question polygraphs (Nelson 2017; 2018), including event-specific diagnostic exams and multiple-issue screening polygraph with two, three and four, relevant questions using the Empirical Scoring System (ESS/ESS-M; Nelson, Krapohl & Handler 2007; Nelson et. al., 2011) and the U.S. Federal three-position scores (Department of Defense, 2006). The multinomial distributions were calculated under the null-hypothesis to the analytic theory of the comparison question test (CQT), which holds that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception or truth-telling in response to relevant target stimuli (Nelson, 2016).

This analytic theory is premised on a more foundational hypothesis that some predictable changes in physiology are correlated with deception and truth-telling and can be recorded and quantified for probabilistic inference and classification. As a practical matter, human

physiology and psychology and sufficiently noisy that data from a single sensor or signal, and single presentation of the test stimuli, provides weak and insufficient information. Instead, an array of sensors, each contributing unique diagnostic variation or information, and systematic repetition of test stimuli, are necessary achieve a satisfactory level of statistical power and signal discrimination.

Multinomial Distributions.

ESS and three-position scores of CQT data are multinomial because each score can take one of three values – indicating that a change in physiological activity in response to a relevant question (RQ) is either greater than, less than, or indiscernible from the change in physiological activity in response to a comparison question (CQ). Multinomial distributions can be calculated using combinatoric math (Abramowitz & Stegun, 1972; Chen & Koh, 1992). CQT formats consist of and array of three or four sensors, include two to four RQs, in a question sequence that is repeated three to five times. A single sensor can produce a large number of combinations of scores: from 3^6 = 729 for three iterations of two RQs, to 3^20 = 3,486,784,401 for five iterations of four RQs. For each CQT format there is a finite number of combinations of multinomial scores and a finite number of ways to achieve each possible sensor score.

A multinomial distribution can be determined for the sensor scores by dividing the number of ways to achieve each sensor score by the number of possible combinations. Calculation of the exact number of ways to – the most complicated part – can be calculated using combinatoric math and multinomial coefficients (Riordan, 2002/1958). It is also possible, and simpler, to simulate the multinomial distribution using Monte Carlo methods. Nelson (2017) shows the results of both methods.

Multinomial distributions for sensor scores can also be combined or permuted to calculate a multinomial distribution for subtotal and grand total CQT scores. Again, this can be accomplished through combinatoric math or via simulation. Nelson (2017, 2018a) provided exact calculations of the multinomial distributions for CQT scores. Regardless of whether obtain through exact combinatoric calculation or via simulation, multinomial distributions are useful as a likelihood function for Bayesian analysis of the change in the strength of posterior information in support of deception or truth-telling.

Bayesian Analysis

Bayesian analysis is the use of Bayes' theorem to analyze data and estimate an unknown parameter or unknown quantity of interest (Bayes & Price, 1756; Berger, 1985, 2006a; Bernardo & Smith, 1994; Box & Tiao, 1973; Casella, 1985; Downey, 2012, Efron, 1986; Gelman et al., 2014; Gill, 2007; Laplace, 1812; Lee, 2004; Rubin, Gelman, Carlin & Stern, 2003; Stone, 2013; Western & Simon, 1994; Winkler, 1972). In the context of the CQT the unknown quantity or parameter is the likelihood of deception or truth-telling. It is not possible to detect or quantify deception or truth *per se* because these are not physical quantities. However, Bayesian analysis permits the application of Bayesian probability – the degree of belief, based on analysis and objective information, in some knowledge or conclusion – to the constructs of deception and truth-telling. Bayesian analysis makes use of observed test data, along with prior probability information and a statistical likelihood function, to calculate a posterior probability. Bayesian analysis can also be used to calculate a Bayes Factor (Berger, 2006b, Jeffreys, 1939/1961; Kaas & Raftery, 1993; Morey & Rouder, 2011), which is the magnitude of change in the posterior strength of information. Bayes Factor is advantageous because it is a robust statistic – the magnitude of change in the strength of posterior information will be the same regardless of the prior value.

Bayesian Classifier for ESS-M and Three-position Scores.

Bayesian analysis of CQT data involves the use of Bayes' theorem, along with the observed test data, prior information and likelihood function, to calculate a posterior conditional likelihood, expressed as an odds, of decep-

tion or truth-telling. The posterior conditional odds can be thought of as a description of the strength of the test result or degree of belief that can be attributed. Use of the odds to express posterior probabilities is advantageous because it permits the discussion of probabilities using whole numbers and also explicates that all probabilities are a comparison of the strength of some possibility compared to the strength of some other possibility. The posterior value from Bayes' theorem can also be thought of as a Bayes Factor when the posterior odds are calculated under the equal prior.

After calculation of the posterior odds or Bayes Factor, the expected variation in test data – if it were possible to conduct the same examination repeatedly under the same circumstances – is quantified in the form of a Bayesian credible interval, analogous to a frequentist confidence interval, using the Clopper-Pearson method (Clopper & Pearson, 1934; Nelson, 2018b). This method is advantageous for the CQT because the resulting upper and lower probability boundaries never result in mathematically absurd values (i.e., never exceeding the 0 and 1 limits of the uniform probability distribution), and the resulting coverage area is known to always exceed the 1-alpha nominal value. A classification of deception or truth-telling is supported when the strength of the 1-alpha lower-limit of a Bayesian credible interval has exceeded the strength of the prior information for deception or truth-telling.

Point Scores and Cutscores.

Field polygraph practitioners have traditionally relied on numerical point scores and numerical cutscores as an expedient method of classifying and interpreting CQT data. Although traditionally little emphasis was placed on the relationship between point scores and probabilities, the availability of both empirical and multinomial reference tables has increased the accessibility and intuition for discussion of this information in field practice during recent years. When using the multinomial distributions for ESS and three-position scores, numerical cutscores can be selected as the minimum (absolute) score for which the strength of the lower limit of the 1-alpha Bayesian credible interval exceeds the strength of the prior information. Numerical scores that equal or exceed the numerical cutscore can be said to increase the strength of information indicative of deception or truth-telling, at the 1-alpha level, relative to the prior information.

Until this time, published multinomial cutscores have been available only for the equal prior condition and only for the symmetrical alpha scheme of α = .05/.05 for deception and truth-telling. This project involved the tabular calculation of multinomial cutscores for the ESS-M and three-position scoring methods for all permutations of alpha levels at .01, .05, and .10 for truth-telling and deception using a distribution of prior odds from one in 10 for truth-telling to one in 10 for deception. Appendices A, B, C and D show the cutscore tables. To reduce the number of tables to the minimum possible, all tables were calculated using the simplified ESS-M solution described by Nelson and Rider (2018) as shown by Nelson, Handler, Coffee, Prado and Blalock (2019). Appendix A shows the tabular calculation of multinomial cutscores for ESS scores of event-specific diagnostic exams. Appendix B shows the multinomial cutscores for ESS scores of multiple-issue screening polygraphs. Appendices C and D show the multinomial cutscores for three-position scores of event-specific diagnostic polygraphs and multiple-issue screening polygraphs, respectively.

Careful inspection of these appendices will shows cutscores that may be at first counter-intuitive; cutscores are selected so that posterior information is strengthened, relative to the prior information, at the 1-alpha level, with the result that under some strong prior conditions cutscores may increase for both deceptive and truthful classifications. Also, information contained in subtotal scores will be of insufficient statistical power to provide posterior information at the 1-alpha level under some strong prior conditions.

Summary and Conclusion.

Analysis of CQT test data is conceptually similar to the analysis of other scientific test data,

Table 1. shows the multinomial cutscores for ESS and three position scores under the equal prior with alpha is .05 for both truth-telling and deception. Notice that three-position multinomial cutscores for multiple issue exams are similar to to ESS-M cutscores as a result of blunted precision when using integer values.

Table 1. Multinomial Cutscores for equal prior and alpha = .05 and .05 for truth-telling and deception.

	Single-issue exams – two-stage rules	Multiple-issue exams – subtotal score rules
ESS-M cutscores	+3 / - 3 (-7)	(+1) / -3
Three-position multinomial cutscores	+2 / - 2 (-6)	(+1) / -3

Parenthesis indicate the use of a statistical correction for multiplicity effects.

and consists of four main functions or operations. These include feature extraction, numerical transformation and data reduction, calculation of a statistical classifier using of some form of likelihood function, and interpretation of the meaning of the numerical information.

These operations are often reduced to simple procedures that can be executed with little awareness of or attention to the underlying processes – and often with imprecise boundaries between the operations. For example, feature extraction can be accomplished simultaneously when the feature of interest is a measurement. In a narrower sense, feature extraction is the identification of useful or meaningful changes in physiological activity in response to test items. Numerical transformation, in practical terms, is the assignment of numerical point scores to responses observed in recorded CQT data. Data reduction can involve a variety of mathematical transformations. However, when working with point scores, data reduction can be a simple matter of addition of subtotal and grand total scores. The simplest form of likelihood function is a numerical cutscore for which we expect the rate of misclassification error or precision to achieve certain desired levels, based on empirical and theoretical evidence. Another simple form of likelihood function can be observed in the form of empirically derived test sensitivity, specificity and error rates. Statistical equations are another form of likelihood function. The purpose of any likelihood function is to calculate a coherent and reproducible likelihood value for the observed data. Interpretation, in its simplest form, is the parsing of analytic results into categorical conclusions such as statistically significant and not statistically significant, or positive and negative.

Interpretation of CQT data, in terms of deception and truth-telling, will involve a number of scientific, philosophical and ethical complexities. These can include: the need to understand the use of probabilistic inference where direct physical measurement is not possible; epistemological questions of precisely what precisely is truth and deception – and what does it mean to test, measure and quantify these; the need for professional accountability when making conclusions that may influence the human rights or future of other persons; and other concerns. A well-developed and satisfactory system of test data analysis will address and manage these concerns by enabling professionals to achieve reproducible analytic conclusions that are correctly anchored in scientific and probabilistic knowledge. Ideally, a test data analysis system will lead to discussions of analytic conclusions that are both scientifically coherent and practically useful.

Polygraph professionals have long ago transitioned away from the interpretation of CQT results using terms such as *deceptive,* as this can encourage unrealistic expectations for deterministic perfection or infallibility. In common usage among polygraph field practitioners today are the terms *deception indicated* and *significant reactions* which more reasonably convey that test results are, of themselves, neither a physical substance nor a physical action, but can be interpreted as a probabilistic indicator when they are statistically significant. With the understanding that all scientific test results are probabilistic, a common question for may will be this: what is the strength of the probabilistic infor-

mation for deception or truth-telling? Another version of the same question is this: what can be reasonably said about the strength of the analytic conclusion? It is here that Bayesian decision-making offers a practical and intuitive advantage over the practice of significance testing – referred to as null-hypothesis significance testing (NHST; Fisher, 1934; Neyman & Pearson, 1928; 1933; Pernet, 2015).

Statistical values using the NHST paradigm – p-values and alpha levels – refer only to strength of evidence for a null-hypothesis (which can be rejected in favor of the alternative hypothesis if sufficiently weak). This nuance important because there is often a problematic impulse to misuse the statistical values themselves as an indication of effect-size or strength of the analytic conclusion. Most importantly, a p-value – intended to reject a null hypothesis – is not an estimate of the strength of the effect size for either the hypothesis or null-hypothesis. Attempts to portray a p-value as an estimate of effect size are an example of a logical fallacy known as *argument from ignorance,* in which the absence of information is misinterpreted as a form of proof. Another important consideration is that, in the NHST paradigm, results are significant, and categorical conclusions are possible, only at the stated alpha level. It is possible that some conclusions that are significant at alpha = .05 may not be statistically significant at alpha = .01.

In contrast, Bayesian statistical values can be interpreted as referring to the strength of information in direct support of a hypothesis or conclusion. An important consideration here is that, although Bayesian probabilities can be interpreted as referring to the hypothesis or conclusion – in the CQT context this is a probability or odds of deception or truth-telling – Bayesian probabilities are *conditional probabilities*. That is, the Bayesian posterior probability can be thought of a test likelihood statistic conditioned on the prior information (or the prior information conditioned on the test likelihood statistic). It is possible that categorical conclusions may change if the posterior conditional probability were calculated with different prior information.

Both NHST and Bayesian analysis assume that available test data are an imperfect representation of an unknown parameter of interest and are subject to sampling variation or measurement error. NHST estimates the expected variability from the available data and the sample size using statistical *confidence intervals*. Bayesian analysis assumes that available data are all the information that is presently available to support a conclusion, and also employs procedures to estimate expected variation in test data. Bayesian analysis differentiates the nuanced meaning of these estimations from the frequentist paradigm by using the term *credible interval* to describe the 1-alpha coverage area for expected variation. In practical terms, this means that multinomial cutscores for Bayesian analysis of CQT data are function of both the prior information and the required alpha level for statistical significance.

Use of numerical cutscores serves as a practical convenience that relieves field practitioners of the odious burden of mathematical and statistical calculations. Multinomial cutscores will permit field practitioners to make classifications of deception and truth-telling with the knowledge that the lower limit of the 1-alpha credible interval will exceed the prior information all point scores that exceed a numerical cutscore. In practical terms this can be interpreted as the 1-alpha level at the data have strengthened the information in support of a deceptive or truthful conclusion. This can also be thought of as the 1-alpha level at which a test is indicative of deception or truth-telling. Also, the 1-alpha level that another test, under the same conditions, will give a similar result. Or, the 1-alpha proportion of repeated tests, under the same conditions, that would give similar results.

Determination of multinomial cutscores for ESS and three-position scores requires the calculation of both Bayes theorem and the Clopper-Pearson interval for the distribution of possible scores. Effectively, this results in the calculation of a unique reference table for every alpha and prior scheme. These calculations can be accomplished manually, though the process is tedious, and can also be accomplished quickly, easily and accurately using any desktop or laptop microcomputer. It is also possible to complete all calculations in a controlled environment and make the infor-

mation available in tabular reference format – which is the purpose of this project.

The multinomial cutscore tables, shown in Appendices A-D, can be a useful convenience to field practitioners and program managers who want visual and tactile access to cutscore information without the need for either manual calculations or the experience of a *black-box* calculations that may provide one solution at a time. Tabular information are of such great convenience that computer algorithms and digital calculators will sometimes make use of tables as an alternative to the repetition of complex mathematical and logical operations. Published cutscore tables also provide additional advantages; they can facilitate training in the use of manual analytic procedures that will strengthen understanding and intuition for the analytic process. Also, tables can be used in circumstances in which computers are not available to complete the required calculations. It is hoped that these multinomial cutscore tables can be useful to field practitioners and program managers who desire more visual and intuitive access to the distribution of numerical cutscores and their relationship to various alpha boundaries and prior information. Multinomial cutscores will permits field practitioners to make classifications of deception and truth-telling with the knowledge that the lower limit of the 1-alpha credible interval will exceed the prior information for all point scores that exceed a multinomial cutscore.

References

Abramowitz, M. & Stegun, I. A. (1972). Multinomial Coefficients. §24.1.2 [pp. 823-824] in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th printing.* National Bureau of Standards (now National Institute of Standards and Technology).

Bayes, T. & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370– 418.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second edition. New York: Springer Verlag.

Berger, J. (2006a). The Case for Objective Bayesian Analysis. *Bayesian Analysis, 1(3)*, 385–402.

Berger, J. O. (2006b). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* John Wiley.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *American Statistician, 39 (2),* 83–87.

Chen, C. C. & Koh, K. M. (1992). *Principles and Technique in Combinatorics.* World Scientific.

Laplace, S. P. (1812). *Théorie analytique des probabilités.* Paris: Courier.

Clopper, C.; Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika. 26,* 404–413.

Department of Defense (2006). *Psychophysiological Detection of Deception Analysis II -- Course #503. Test data analysis: DoDPI numerical evaluation scoring system.* Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007).

Downey, A. B. (2012). *Think Bayes. Bayesian Statistics Made Simple.* Green Tea Press.

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician. 40(1).* 1-5.

Fisher, R. A. (1934). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtary, A., & Rubin, D. B. (2014). *Bayesian Data Analysis.* CRC Press.

Gill, J. (2007). *Bayesian Methods: A Social and Behavioral Sciences Approach.* Second Edition. Chapman and Hall/CRC.

Jeffreys, H. (1939). *Theory of Probability. 3rd edition.* Clarendon Press. [Oxford: 1961].

Kaas, R. E. & Raftery, A. E. (1993). *Bayes Factors and Model Uncertainty. Technical Report No. 254.* Department of Statistics: University of Washington.

Lee, P. M. (2004). *Bayesian Statistics, an introduction (3rd ed.)*. Wiley.

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16(4)*, 406–419.

Nelson, R. (2016). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5)*, 69-82.

Nelson, R. (2017). Multinomial reference distributions for the Empirical Scoring System. *Polygraph & Forensic Credibility Assessment, 46(2)*, 81-115.

Nelson, R. (2018a). Multinomial reference distributions for three-position scores of comparison question polygraph examinations. *Polygraph & Forensic Credibility Assessment, 47(2),* 158-175.

Nelson, R. (2018b). Five-minute science lesson: Clopper-Pearson credibility intervals for Bayesian analysis of multinomial polygraph scores. *APA Magazine, 51(3)*, 61-70.

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., Handler, M., Coffey, T., Prado, R. & Blalock, B. (2019). How to: a step-by-step worksheet for the multinomial ESS. *Polygraph & Forensic Credibility Assessment, 48 (1)*, 60-75.

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., Rider, J. (2018). Practical Polygraph: ESS-M Made Simple. *APA Magazine, 51 (6)*, 55–62.

Neyman J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika, 20A(1/2)*, 175–240.

Neyman, J., and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Ser. A, 231,* 289–337.

Pernet, C. (2015). Null hypothesis significance testing: a short tutorial. *F1000Research, 4*, 621.

Riordan, J. (2002). *Introduction to Combinatorial Analysis.* Dover. [1958: Wiley].

Rubin, D. B., Gelman, A., Carlin, J. B. & Stern, H. (2003). *Bayesian Data Analysis (2nd ed.).* Boca Raton: Chapman & Hall/CRC.

Stone, J. (2013). Bayes' Rule: *A Tutorial Introduction to Bayesian Analysis.* Sebtel Press.

Western, B. & Simon J. (1994). Bayesian inference for comparative research. *American Political Science Review,* 88(2), 412-423.

Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision.* Holt McDougal.

Appendix A: Multinomial Cutscores for ESS Scores of Single Issue Exams

ESS-M Scores / Event-Specific Exam

		Alpha (truth/deception)								
Prior odds of deception	prior probability	.01/.01	.01/.05	.01/.10	.05/.01	.05/.05	.05/.10	.10/.01	.10/.05	.10/.10
9 to 1 (9 in 10)	.90	+14 / -9 (none)	+14 / -6 (none)	+14 / -4 (-21)	+13 / -9 (none)	+13 / -6 (none)	+13 / -4 (-21)	+13 / -9 (none)	+13 / -6 (none)	+13 / -4 (-21)
8 to 1 (8 in 9)	.89	+13 / -8 (none)	+13 / -5 (none)	+13 / -4 (-20)	+13 / -8 (none)	+13 / -5 (none)	+13 / -4 (-20)	+12 / -8 (none)	+12 / -5 (none)	+12 / -4 (-20)
7 to 1 (7 in 8)	.88	+13 / -7 (none)	+13 / -5 (-22)	+13 / -4 (-18)	+12 / -7 (none)	+12 / -5 (-22)	+13 / -4 (-18)	+12 / -7 (none)	+12 / -5 (-22)	+12 / -4 (-18)
6 to 1 (6 in 7)	.86	+12 / -7 (none)	+12 / -4 (-20)	+12 / -4 (-16)	+11 / -7 (none)	+11 / -4 (-20)	+11 / -4 (-16)	+11 / -7 (none)	+11 / -4 (-20)	+11 / -4 (-16)
5 to 1 (5 in 6)	.83	+11 / -6 (none)	+11 / -4 (-17)	+11 / -3 (-15)	+10 / -6 (none)	+10 / -4 (-17)	+10 / -3 (-15)	+10 / -6 (none)	+10 / -4 (-17)	+10 / -3 (-15)
4 to 1 (4 in 5)	.80	+10 / -5 (-21)	+10 / -4 (-15)	+10 / -3 (-14)	+10 / -5 (-21)	+10 / -4 (-15)	+10 / -3 (-14)	+9 / -5 (-21)	+9 / -4 (-15)	+9 / -3 (-14)
3 to 1 (3 in 4)	.75	+9 / -5 (-16)	+9 / -3 (-13)	+9 / -3 (-12)	+8 / -5 (-16)	+8 / -3 (-13)	+8 / -3 (-12)	+8 / -5 (-16)	+8 / -3 (-13)	+8 / -3 (-12)
2 to 1 (2 in 3)	.67	+7 / -4 (-13)	+7 / -3 (-11)	+7 / -2 (-10)	+6 / -4 (-13)	+6 / -3 (-11)	+6 / -2 (-10	+6 / -4 (-13)	+6 / -3 (-11)	+6 / -2 (-10)
1 to 1 (1 in 2)	**.50**	**+4 / -4 (-9)**	**+4 / -3 (-7)**	**+4 / -2 (-6)**	**+3 / -4 (-9)**	**+3 / - 3 (-7)**	**+3 / -2 (-6)**	**+2 / -4 (-9)**	**+2 / -3 (-7)**	**+2 / -2 (-6)**
1 to 2 (1 in 3)	.33	+4 / -7 (-11)	+4 / -6 (-9)	+4 / -6 (-8)	+3 / -7 (-11)	+3 / -6 (-9)	+3 / -6 (-8)	+2 / -7 (-11)	+2 / -6 (-9)	+2 / -6 (-8)
1 to 3 (1 in 4)	.25	+5 / -9 (-11)	+5 / -8 (-9)	+5 / -8 (-8)	+3 / -9 (-11)	+3 / -8 (-9)	+3 / -8 (-8)	+3 / -9 (-11)	+3 / -8 (-9)	+3 / -8 (-8)
1 to 4 (1 in 5)	.20	+5 / -10 (-12)	+5 / -10 (-10)	+5 / -9 (-9)	+4 / -10 (-12)	+4 / -10 (-10)	+4 / -9 (-9)	+3 / -10 (-12)	+3 / -10 (-10)	+3 / -9 (-9)
1 to 5 (1 in 6)	.17	+6 / -11 (-12)	+6 / -10 (-10)	+6 / -10 (-10)	+4 / -11 (-12)	+4 / -10 (-10)	+4 / -10 (-10)	+3 / -11 (-12)	+3 / -10 (-10)	+3 / -10 (-10)
1 to 6 (1 in 7)	.14	+7 / -12 (-13)	+7 / -11 (-11)	+7 / -11 (-10)	+4 / -12 (-13)	+4 / -11 (-11)	+4 / -11 (-10)	+4 / -12 (-13)	+4 / -11 (-11)	+4 / -11 (-10)
1 to 7 (1 in 8)	.13	+7 / -12 (-13)	+7 / -12 (-11)	+7 / -11 (-10)	+5 / -12 (-13)	+5 / -12 (-11)	+5 / -11 (-10)	+4 / -12 (-13)	+4 / -12 (-11)	+4 / -11 (-10)
1 to 8 (1 in 9)	.11	+8 / -13 (-13)	+7 / -13 (-11)	+8 / -12 (-10)	+5 / -13 (-13)	+5 / -13 (-11)	+5 / -12 (-10)	+4 / -13 (-13)	+4 / -13 (-11)	+4 / -12 (-10)
1 to 9 (1 in 10)	.10	+9 / -14 (-13)	+9 / -13 (-12)	+9 / -13 (-11)	+6 / -14 (-13)	+6 / -13 (-12)	+6 / -13 (-11)	+4 / -14 (-13)	+4 / -13 (-12)	+4 / -13 (-11)

Parenthesis indicate the use of a statistical correction for multiplicity effects.

Appendix B: Multinomial Cutscores for ESS Scores of Multiple Issue Exams

ESS-M Scores / Multiple Issue Exam

		Alpha (truth/deception)								
Prior odds of deception	prior probability	.01/.01	.01/.05	.01/.10	.05/.01	.05/.05	.05/.10	.10/.01	.10/.05	.10/.10
9 to 1 (9 in 10)	.90	(+8) / none	(+8) / none	(+8) / -9	(+7) / none	(+7) / none	(+7) / -9	(+7) / none	(+7) / none	(+7) / -9
8 to 1 (8 in 9)	.89	(+7) / none	(+7) / none	(+7) / -6	(+7) / none	(+7) / none	(+7) / -6	(+7) / none	(+7) / none	(+7) / -6
7 to 1 (7 in 8)	.88	(+7) / none	(+7) / -10	(+7) / -6	(+7) / none	(+7) / -10	(+7) / -6	(+7) / none	(+7) / -10	(+7) / -6
6 to 1 (6 in 7)	.86	(+7) / none	(+7) / -8	(+7) / -5	(+6) / none	(+6) / -8	(+6) / -4	(+6) / none	(+6) / -8	(+6) / -5
5 to 1 (5 in 6)	.83	(+6) / -14	(+6) / -5	(+6) / -4	(+6) / -14	(+6) / -5	(+6) / -4	(+6) / -14	(+6) / -5	(+6) / -4
4 to 1 (4 in 5)	.80	(+6) / -9	(+6) / -5	(+6) / -4	(+5) / -9	(+5) / -5	(+5) / -4	(+5) / -9	(+5) / -5	(+5) / -4
3 to 1 (3 in 4)	.75	(+5) / -6	(+5) / -4	(+5) / -3	(+5) / -6	(+5) / -4	(+5) / -3	(+4) / -6	(+4) / -4	(+4) / -3
2 to 1 (2 in 3)	.67	(+4) / -5	(+4) / -3	(+4) / -3	(+3) / -5	(+3) / -3	(+3) / -3	(+3) / -5	(+3) / -3	(+3) / -3
1 to 1 (1 in 2)	**.50**	**(+2) / -4**	**(+2) / -3**	**(+2) / -2**	**(+1) / -4**	**(+1) / -3**	**(+1) / -2**	**(+1) / -4**	**(+1) / -3**	**(+1) / -2**
1 to 2 (1 in 3)	.33	(+1) / -6	(+1) / -5	(+1) / -4	(0) / -6	(0) / -5	(0) / -4	(0) / -6	(0) / -5	(0) / -4
1 to 3 (1 in 4)	.25	(0) / -7	(0) / -6	(0) / -6	(0) / -7	(0) / -6	(0) / -6	(0) / -7	(0) / -6	(0) / -6
1 to 4 (1 in 5)	.20	(+1) / -7	(0) / -7	(0) / -6	(0) / -7	(0) / -7	(0) / -6	(0) / -7	(0) / -7	(0) / -6
1 to 5 (1 in 6)	.17	(+4) / -8	(+4) / -7	(+4) / -7	(0) / -8	(0) / -7	(0) / -7	(0) / -8	(0) / -7	(0) / -7
1 to 6 (1 in 7)	.14	(none) / -8	(none) / -8	(none) / -7	(0) / -8	(0) / -8	(0) / -7	(0) / -8	(0) / -8	(0) / -7
1 to 7 (1 in 8)	.13	(none) / -9	(none) / -8	(none) / -8	(0) / -9	(0) / -8	(0) / -8	(0) / -9	(0) / -8	(0) / -8
1 to 8 (1 in 9)	.11	(none) / -9	(none) / -8	(none) / -8	(none) / -9	(none) / -8	(none) / -8	(0) / -9	(0) / -8	(0) / -8
1 to 9 (1 in 10)	.10	(none) / -9	(none) / -9	(none) / -8	(none) / -9	(none) / -9	(none) / -8	(0) / -9	(0) / -9	(0) / -8

Parenthesis indicate the use of a statistical correction for multiplicity effects.

Appendix C: Multinomial Cutscores for Three-Position Scores of Single Issue Exams

3-Position Scores / Event-Specific Exam

		Alpha (truth/deception)								
Prior odds of deception	prior probability	.01/.01	.01/.05	.01/.10	.05/.01	.05/.05	.05/.10	.10/.01	.10/.05	.10/.10
9 to 1 (9 in 10)	.90	+11 / -8 (none)	+11 / -5 (none)	+11 / -4 (none)	+10 / -8 (none)	+10 / -5 (none)	+10 / -4 (none)	+10 / -8 (none)	+10 / -5 (none)	+10 / -4 (none)
8 to 1 (8 in 9)	.89	+10 / -8 (none)	+10 / -5 (none)	+10 / -4 (none)	+10 / -8 (none)	+10 / -5 (none)	+10 / -4 (none)	+9 / -8 (none)	+9 / -5 (none)	+9 / -4 (none)
7 to 1 (7 in 8)	.88	+10 / -7 (none)	+10 / -4 (none)	+10 / -4 (-17)	+9 / -7 (none)	+9 / -4 (none)	+9 / -4 (-17)	+9 / -7 (none)	+9 / -4 (none)	+9 / -4 (-17)
6 to 1 (6 in 7)	.86	+10 / -6 (none)	+10 / -4 (none)	+10 / -3 (-14)	+9 / -6 (none)	+9 / -4 (none)	+9 / -3 (-14)	+9 / -6 (none)	+9 / -4 (none)	+9 / -3 (-14)
5 to 1 (5 in 6)	.83	+9 / -5 (none)	+9 / -4 (-15)	+9 / -3 (-12)	+8 / -5 (none)	+8 / -4 (-15)	+8 / -3 (-12)	+8 / -5 (none)	+8 / -4 (-15)	+8 / -3 (-12)
4 to 1 (4 in 5)	.80	+8 / -5 (none)	+8 / -3 (-13)	+8 / -3 (-11)	+8 / -5 (none)	+8 / -3 (-13)	+8 / -3 (-11)	+7 / -5 (none)	+7 / -3 (-13)	+7 / -3 (-11)
3 to 1 (3 in 4)	.75	+7 / -4 (-14)	+7 / -3 (-11)	+7 / -2 (-10)	+7 / -4 (-14)	+7 / -3 (-11)	+7 / -2 (-10)	+6 / -4 (-14)	+6 / -3 (-11)	+6 / -2 (-10)
2 to 1 (2 in 3)	.67	+6 / -3 (-11)	+6 / -3 (-9)	+6 / -2 (-8)	+5 / -3 (-11)	+5 / -3 (-9)	+5 / -2 (-8)	+5 / -3 (-11)	+5 / -3 (-9)	+5 / -2 (-8)
1 to 1 (1 in 2)	**.50**	**+3 / -3 (-8)**	**+3 / -2 (-6)**	**+3 / -2 (-5)**	**+2 / -3 (-8)**	**+2 / -2 (-6)**	**+2 / -2 (-5)**	**+2 / -3 (-8)**	**+2 / -2 (-6)**	**+2 / -2 (-5)**
1 to 2 (1 in 3)	.33	+3 / -6 (-9)	+3 / -5 (-7)	+3 / -5 (-6)	+3 / -6 (-9)	+3/ -5 (-7)	+3 / -5 (-6)	+2 / -6 (-9)	+2 / -5 (-7)	+2 / -5 (-6)
1 to 3 (1 in 4)	.25	+4 / -7 (-9)	+4 / -7 (-8)	+4 / -6 (-7)	+3 / -7 (-9)	+3 / -7 (-8)	+3 / -6 (-7)	+2 / -7 (-9)	+2 / -7 (-8)	+2 / -6 (-7)
1 to 4 (1 in 5)	.20	+5 / -8 (-10)	+5 / -8 (-8)	+5 / -7 (-7)	+3 / -8 (-10)	+3 / -8 (-8)	+3 / -7 (-7)	+3 / -8 (-10)	+3 / -8 (-8)	+3 / -7 (-7)
1 to 5 (1 in 6)	.17	+5 / -9 (-10)	+5 / -8 (-9)	+5 / -8 (-8)	+4 / -9 (-10)	+4 / -8 (-9)	+4 / -8 (-8)	+3 / -9 (-10)	+3 / -8 (-9)	+3 / -8 (-8)
1 to 6 (1 in 7)	.14	+6 / -10 (-10)	+6 / -9 (-9)	+6 / -9 (-8)	+4 / -10 (-10)	+4 / -9 (-9)	+4 / -9 (-8)	+3 / -10 (-10)	+3 / -9 (-9)	+3 / -9 (-8)
1 to 7 (1 in 8)	.13	+6 / -10 (-10)	+6 / -9 (-9)	+6 / -9 (-8)	+4 / -10 (-10)	+4 / -9 (-9)	+4 / -9 (-8)	+3 / -10 (-10)	+3 / -9 (-9)	+3 / -9 (-8)
1 to 8 (1 in 9)	.11	+8 / -10 (-11)	+8 / -10 (-9)	+8 / -9 (-8)	+5 / -10 (-11)	+5 / -10 (-9)	+5 / -9 (-8)	+4 / -10 (-11)	+4 / -10 (-9)	+4 / -9 (-8)
1 to 9 (1 in 10)	.10	+8 / -11 (-11)	+8 / -10 (-9)	+8 / -10 (-9)	+5 / -11 (-11)	+5 / -10 (-9)	+5 / -10 (-9)	+4 / -11 (-11)	+4 / -10 (-9)	+4 / -10 (-9)

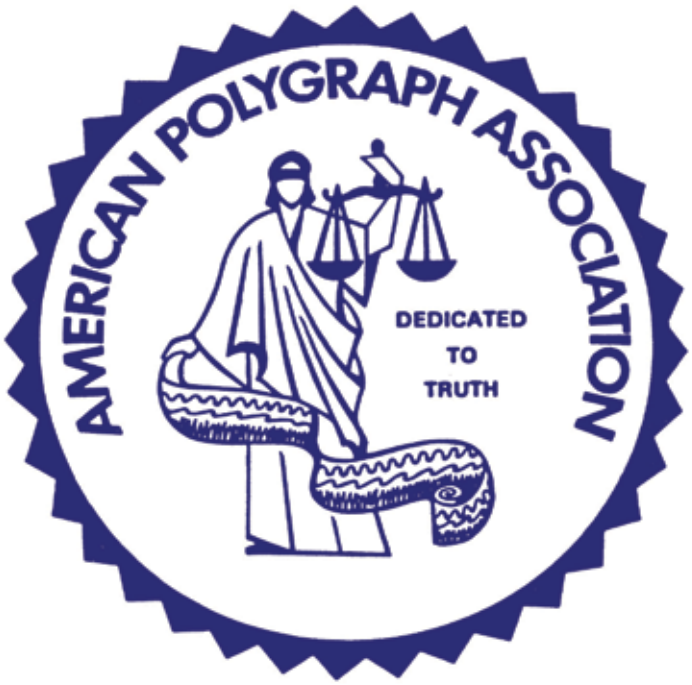
Parenthesis indicate the use of a statistical correction for multiplicity effects.

Appendix C: Multinomial Cutscores for Three-Position Scores of Multiple Issue Exams

3-Position Scores / Multiple Issue Exam

		Alpha (truth / deception)								
Prior odds of deception	prior probability	.01/.01	.01/.05	.01/.10	.05/.01	.05/.05	.05/.10	.10/.01	.10/.05	.10/.10
9 to 1 (9 in 10)	.90	(+6) / none	(+6) / none	(+6) / none	(+6) / none	(+6) / none	(+6) / none	(+6) / none	(+6) / none	(+6) / none
8 to 1 (8 in 9)	.89	(+6) / none	(+6) / none	(+6) / -10	(+6) / none	(+6) / none	(+6) / -10	(+5) / none	(+5) / none	(+5) / -10
7 to 1 (7 in 8)	.88	(+6) / none	(+6) / none	(+6) / -7	(+5) / none	(+5) / none	(+5) / -7	(+5) / none	(+5) / none	(+5) / -7
6 to 1 (6 in 7)	.86	(+5) / none	(+5) / -10	(+5) / -5	(+5) / none	(+5) / -10	(+5) / -5	(+5) / none	(+5) / -10	(+5) / -5
5 to 1 (5 in 6)	.83	(+5) / none	(+5) / -5	(+5) / -4	(+5) / none	(+5) / -5	(+5) / -4	(+5) / none	(+5) / -5	(+5) / -4
4 to 1 (4 in 5)	.80	(+5) / none	(+5) / -4	(+5) / -3	(+4) / none	(+4) / -4	(+4) / -3	(+4) / none	(+4) / -4	(+4) / -3
3 to 1 (3 in 4)	.75	(+4) / -7	(+4) / -4	(+4) / -3	(+4) / -7	(+4) / -4	(+4) / -3	(+4) / -7	(+4) / -4	(+4) / -3
2 to 1 (2 in 3)	.67	(+3) / -4	(+3) / -3	(+3) / -2	(+3) / -4	(+3) / -3	(+3) / -2	(+3) / -4	(+3) / -3	(+3) / -2
1 to 1 (1 in 2)	**.50**	**(+1) / -3**	**(+1) / -3**	**(+1) / -2**	**(+1) / -3**	**(+1) / -3**	**(+1) / -2**	**(+1) / -3**	**(+1) / -3**	**(+1) / -2**
1 to 2 (1 in 3)	.33	(+1) / -5	(+1) / -4	(+1) / -4	(0) / -5	(0) / -4	(0) / -4	(0) / -5	(0) / -4	(0) / -4
1 to 3 (1 in 4)	.25	(+1) / -6	(+1) / -5	(+1) / -4	(0) / -6	(0) / -5	(0) / -4	(0) / -6	(0) / -5	(0) / -4
1 to 4 (1 in 5)	.20	(none) / -6	(none) / -5	(none) / -5	(0) / -6	(0) / -5	(0) / -5	(0) / -6	(0) / -5	(0) / -5
1 to 5 (1 in 6)	.17	(none) / -6	(none) / -6	(none) / -5	(0) / -6	(0) / -6	(0) / -5	(0) / -6	(0) / -6	(0) / -5
1 to 6 (1 in 7)	.14	(none) / -7	(none) / -6	(none) / -6	(+2) / -7	(+2) / -6	(+2) / -6	(0) / -7	(0) / -6	(0) / -6
1 to 7 (1 in 8)	.13	(none) / -7	(none) / -6	(none) / -6	(none) / -7	(none) / -6	(none) / -6	(0) / -7	(0) / -6	(0) / -6
1 to 8 (1 in 9)	.11	(none) / -7	(none) / -7	(none) / -6	(none) / -7	(none) / -7	(none) / -6	(+1) / -7	(+1) / -7	(+1) / -6
1 to 9 (1 in 10)	.10	(none) / -7	(none) / -7	(none) / -6	(none) / -7	(none) / -7	(none) / -6	(none) / -7	(none) / -7	(none) / -6

Parenthesis indicate the use of a statistical correction for multiplicity effects.


AMERICAN POLYGRAPH ASSOCIATION
DEDICATED
TO
TRUTH