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 orfalse-negative rates for seven-position, three-position or ESS scores with two, threeor 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 criteriondeceptive 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 cutscoresand 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 comparisonquestion 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 canachieve 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 testingerrors: wherein the hypothesis is that multiple-issue polygraphs with two, three and fourRQs can achieve false-positive and false-negative error rates that are significantly less thanchance.

Method

Monte Carlo methods were used to calculate confidence intervals for criterion accuracy of multiple-issue polygraph examinations withtwo, 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 thesenine 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.

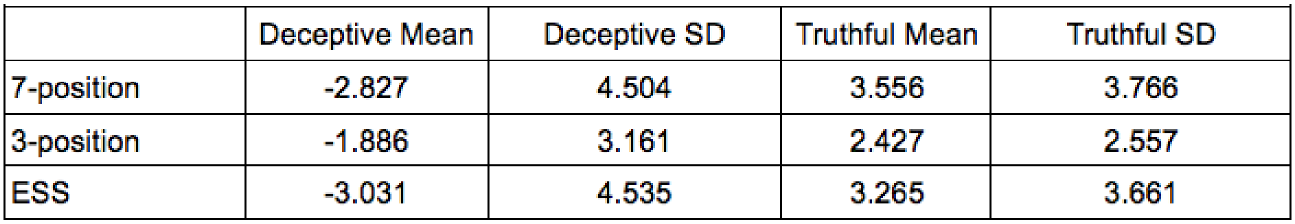
Subtotal scores were simulated by standardizing random numbers to seeding parametersthat were the means and standard deviations of the subtotal scores provided by the participants in the Krapohl and Cushman (2006) study after transforming the seven-position subtotal scores of the guilty and innocent 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, andfour RQs = .159. For each RQ in each casea 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 foreach Monte Carlo distribution that converged at .5 while randomly setting the criterion statefor 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 eachof the Monte Carlo spaces the criterion state ofeach case in the Monte Carlo space was set todeceptive if the criterion state of one or more of the RQs was deceptive. The criterion statesof the cases were set to truthful if the criterionstatus of all RQs was truthful.

Traditional cutscores were used for the for theseven-position and three-position TDA methods: test results were classified as deceptive when any subtotal score was -3 or lower, andtest results were classified as truthful when allsubtotal 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, *etal.*, 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 guiltyand innocent persons (Nelson *et al.*, 2011, Nelson, 2017, Nelson & Rider, 2018). The maindifference 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 fortruth-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 truthfulwhen 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 testresults 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 sub-total score equals or exceeds the cutscore for deceptive classifications, and are classified asindicative of truth-telling – using the term *no significant reactions* – when all subtotal scoresequal 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 per- mitted 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-tellingto other RQs within the individual cases in theMonte 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 werecompleted with a level of significance set at alpha = .05. These may be found in AppendicesA 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 meth- od 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 becomputed 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 MGQTexams 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 methodswith 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.

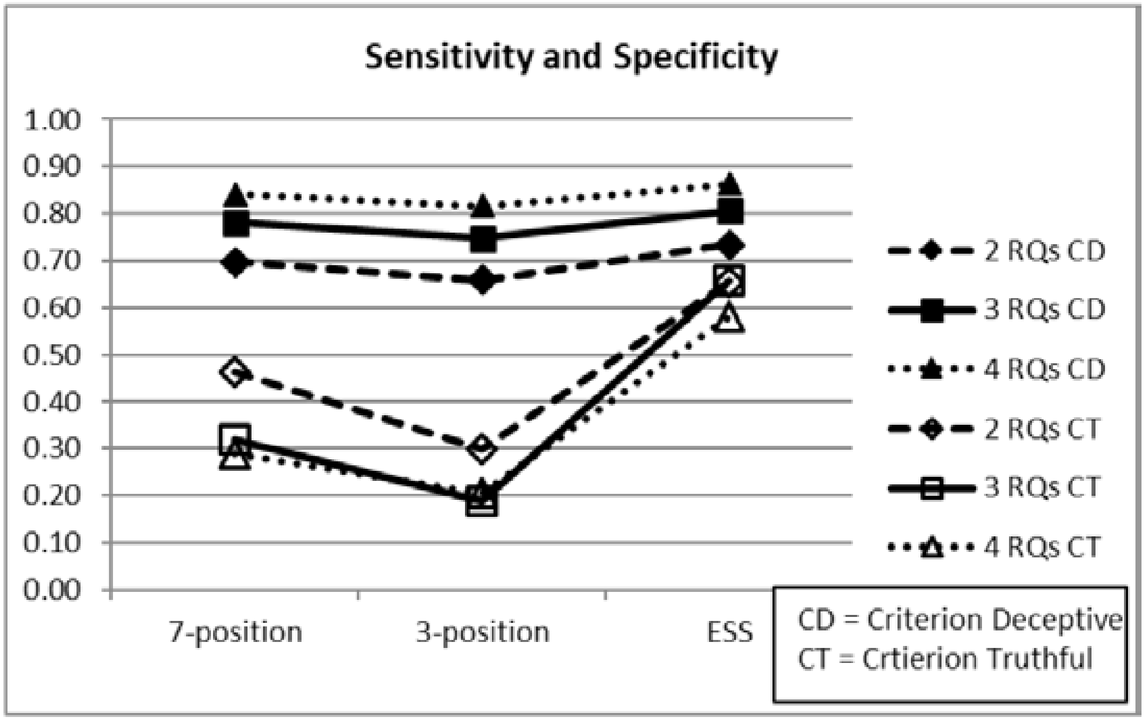


Figure 1 shows that mean test sensitivity to deception exceeded chance (.5) for all three scoring methods, while mean test specificity to truth-telling did not exceed chance for the seven-position and three-position scoring methods.

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 scoringmethod (*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 interactionof TDA method x number of RQs was significant for the truthful cases *F* (1,441) = 5.669, *p* = <.001). One-way ANOVAs showed that maineffects for the truthful cases were not significant for the number of RQs (p = .799) or for thescoring 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 sourceof 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 methodx 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 shownin 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.

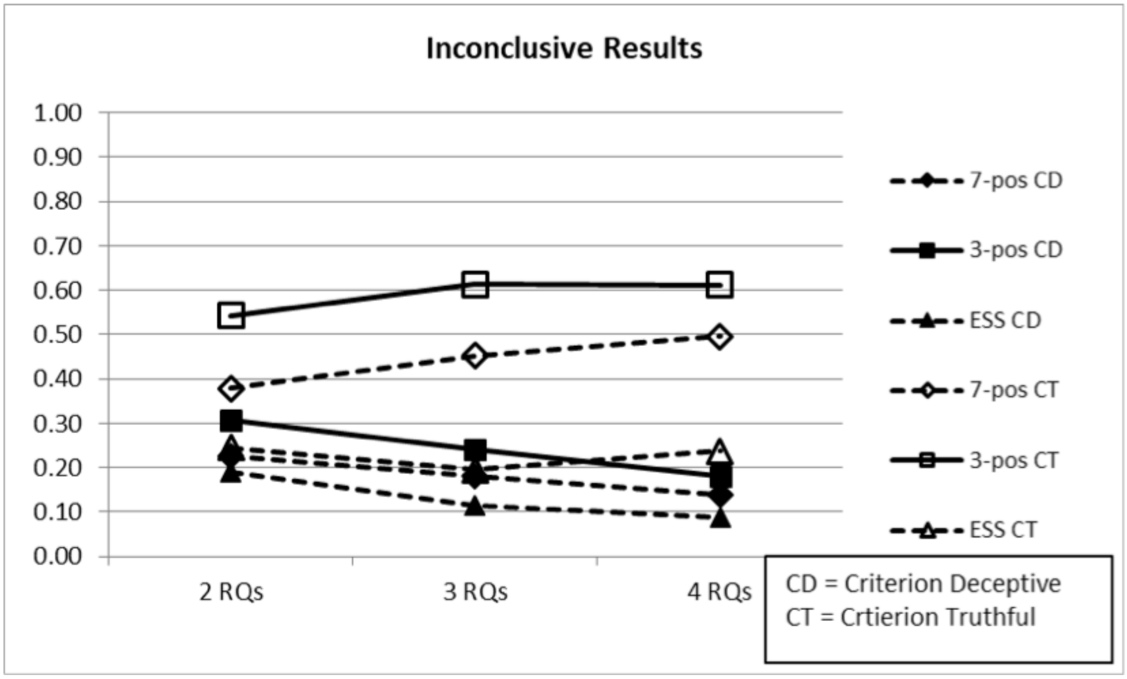
Table 5. Three-way ANOVA summary for inconclusive results (RQs x TDA method x criterion state)

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 sev en-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 inter- action effects inconclusive results as a function of RQs x scoring method can be attributed to differences between the scoring methodswith 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 ANO- VA contrast was calculated for the seven-position and three-position scores, excluding the ESS scores. The three-way interaction for in- conclusive 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 meth- od 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.

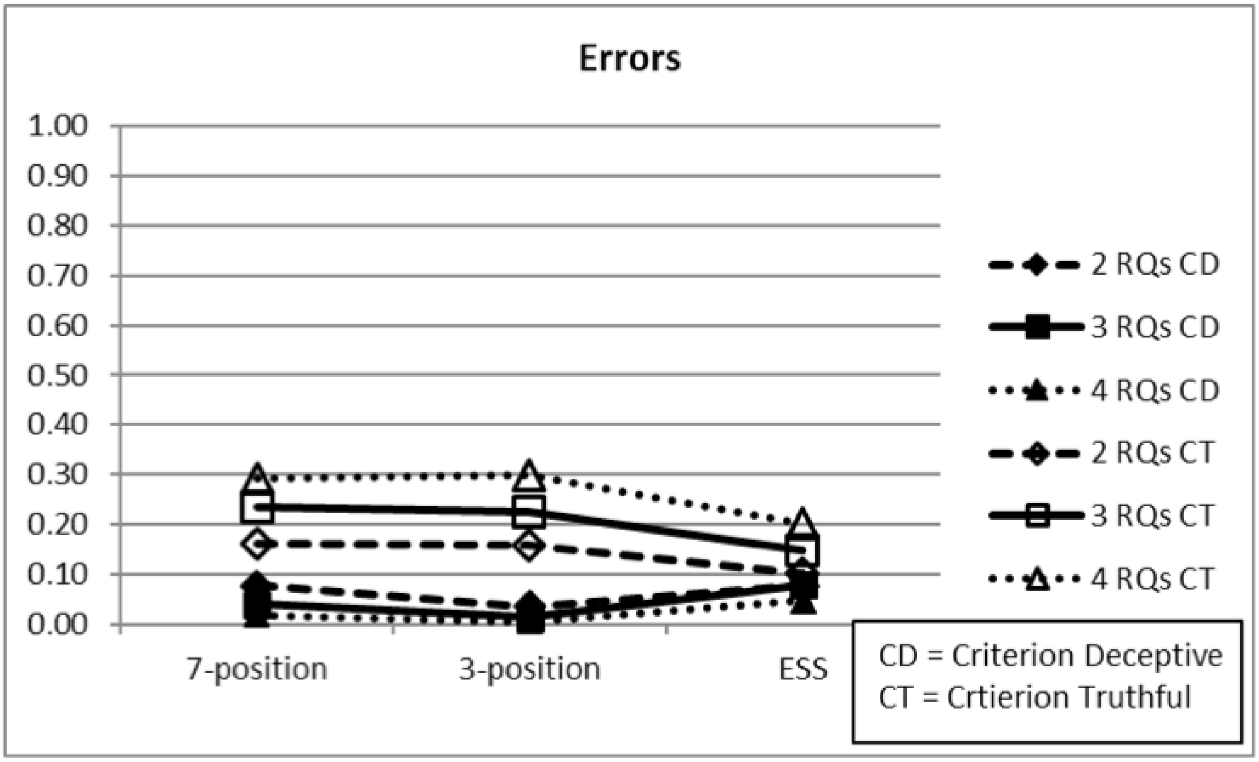


Table 6. Three-way ANOVA summary for errors (RQs x TDA method x criterion state)

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 summaryshown in Table 7 indicates a significant inter-action [*F*(1,894) = 104.051, (*p* < .001)] for decision errors as a function of the number of RQsand criterion state.

Although errors appear to increase with number of RQs for criterion truthful cases and de-crease 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.

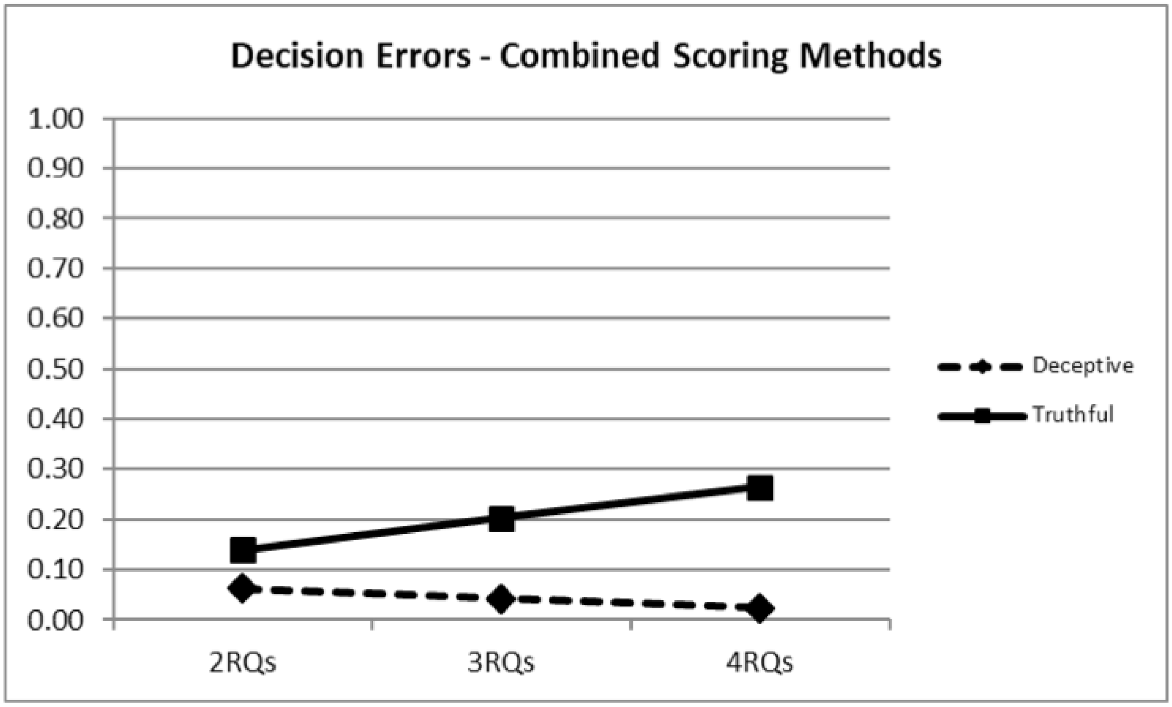


Table 7. Two-way ANOVA summary for decision errors with 7 position scores (RQs x criterion state).

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 ifone actually existed. This suggests that the observed interaction can be attributed to the factthat, although the difference for two, three or four RQs are not significant within the truthful or deceptive cases, the likelihood of testingerror for multiple issue polygraphs increases with the number of RQs for criterion truthful cases while decreasing for criterion deceptive cases.

Unweighted average accuracy.

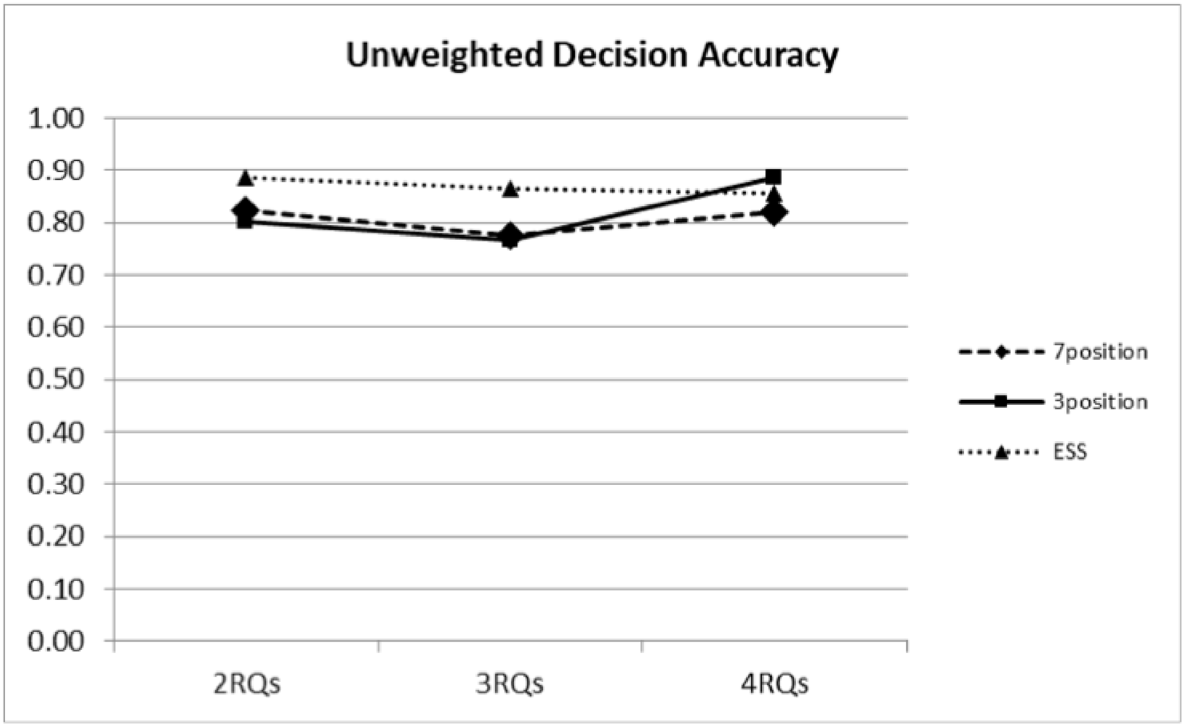
Unweighted decision accuracy excluding in- conclusive results is shown in Table 2, and was significantly greater than chance (.5) for all three TDA methods with two, three, and four RQs (*p* < .05). Table 8 also shows that variation in test accuracy increases as a function of the number of RQs for all three scoring methods. A two-way. Similarly, as shown in the appendices, both false-negative andfalse-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 functionof 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. Theseresults indicate there is no real difference in unweighted accuracy for PDD results with2RQs, 3RQs or 4RQs, excluding inconclusive results.

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

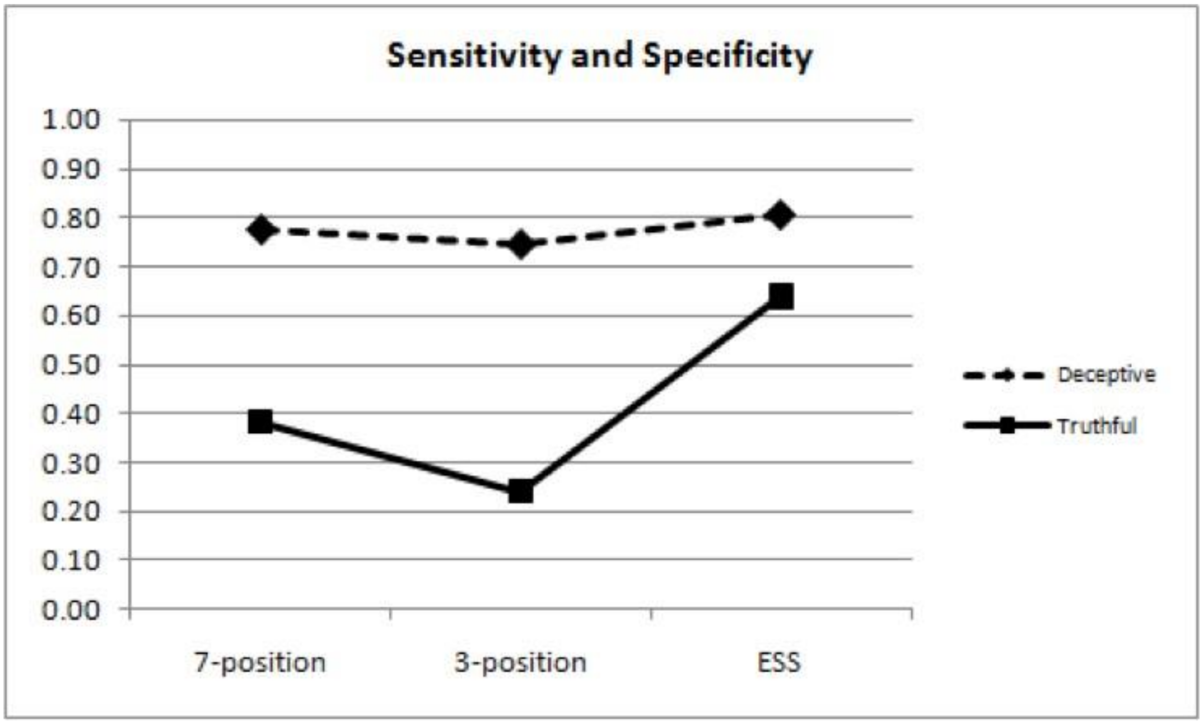
Three additional Monte Carlo models wereused to further understand any differences between the seven-position, three-position, and ESS scoring methods while randomizing the number of RQs for each case in the Monte Carlo space. For each case, the number of RQs was varied randomly from two, three, or four 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 resultsthat 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 MGQTexams with randomized two, three, or fourRQs.

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.

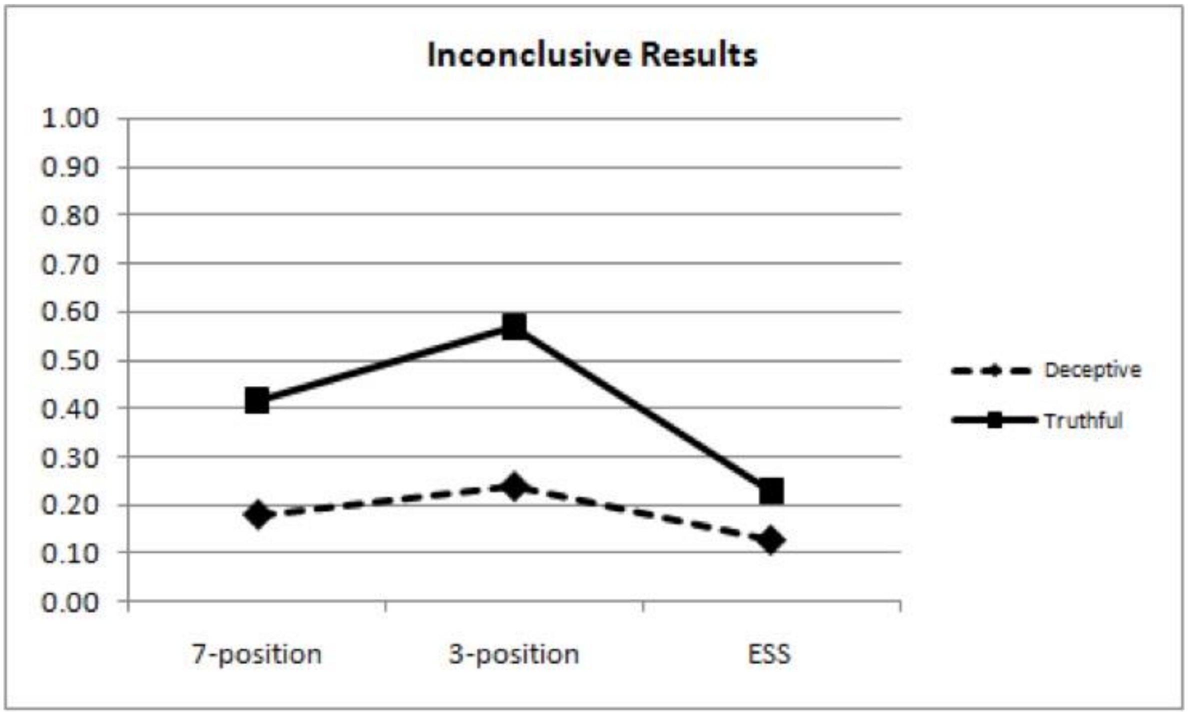


Evaluation of the simple main effects for scoring method showed that the difference in detection of deception differed significantly fromdetection of truth-telling for the seven-positionscoring method [*F*(1,98) = 8.307, (*p* = .005)] and for the three-position scoring method [*F*(1,98) = 19.438, (*p* < .001)]. The simple maineffect for criterion deceptive and criterion truthful cases was not significant for the ESS(p = .222). These results indicate the two-wayinteraction can be attributed to differencestest sensitivity and test specificity for the ESSscoring method compared to the seven-position and three-position methods. As shown inAppendix 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 examswith randomized two, three, or four RQs.

A two-way ANOVA for inconclusive results(scoring method x criterion status) showedsignificant 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 TDAmethods. Simple main effects for inconclusiveresults were not significant for the deceptive cases (*p* = .185) or truthful cases (*p* = .177).

Figure 7. Monte Carlo mean estimates for inconclusive rates.



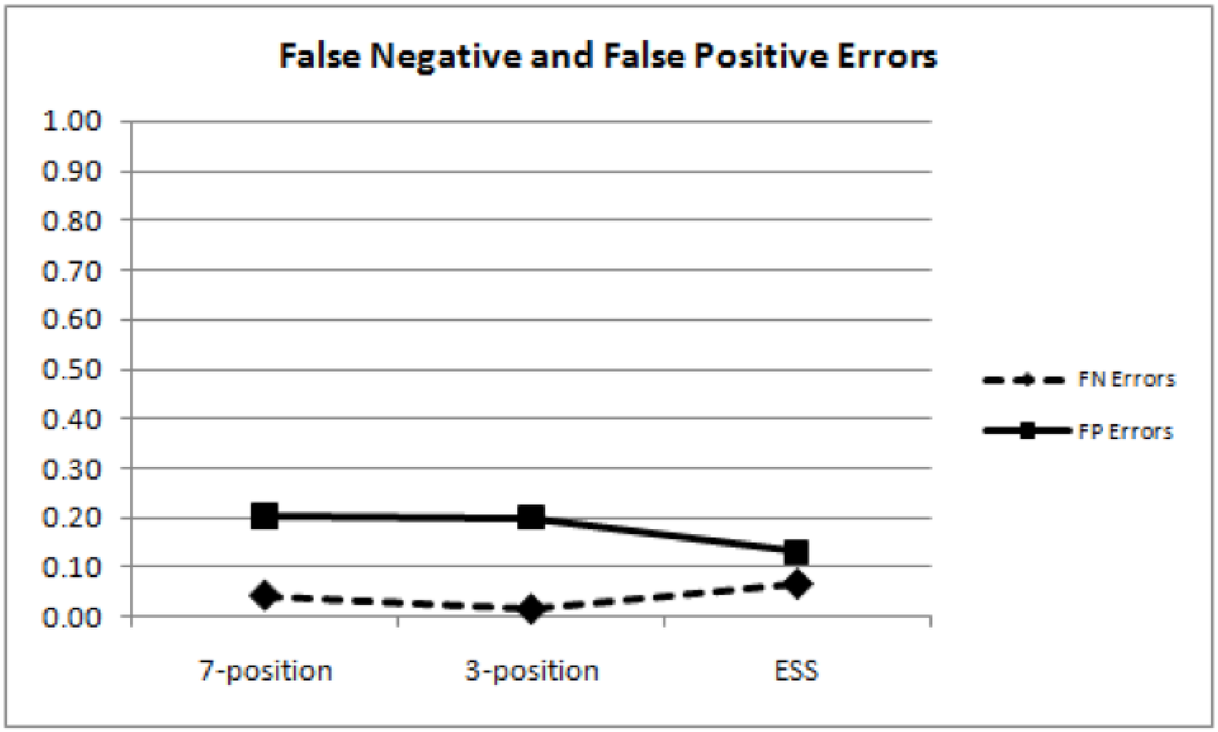
The simple main effect, for differences in inconclusive rates with criterion deceptive and criterion truthful cases, was significant forthe three-position scores [*F*(1,98) = 5.147, (*p* = .025)], but not for seven-position scores (p = .084) or the ESS (p = .413). These results indicate that the observed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

Decision errors for USAF MGQT exams withrandomized 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 thethree TDA methods.

Figure 8. Monte Carlo mean estimates for inconclusive rates.



Simple main effects for were not significant forfalse-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 linesis 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 thetest 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 equalnumber 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) forall three TDA methods.

Despite the fact that unweighted accuracy didnot 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 withthe number of RQs for criterion truthful casesand 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 moreclosely to the trend exhibited by the seven-position and three-position results in this study.

No significant differences were found infalse-positive or false-negative error rates as afunction 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 criteriondeceptive cases.

In addition to the investigation of criterion ac-curacy 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 lev-el (.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 functionof differences in numerical cutscores. Resultsfor 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. ESScut-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 (orif 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 sensitivityand 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. MonteCarlo models, although insufficient to providea 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 fieldstudies. Use of subtotal seed parameters thatwere 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 representthe 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 supportthe 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 thanchance 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 rulesfor 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 set- tings, 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}