Extended Analysis of a Rank Order Scoring Model and the Multi-Facet Hypothesis with the Federal Zone Comparison Technique

Raymond Nelson and Mark Handler

Abstract

Archival scores were used to evaluate the multi-facet hypothesis regarding test questions in the Federal ZCT when scored via a rank order test data analysis (TDA) model. Data were analyzed using bootstrapping and multivariate analysis. Dimensional profiles and statistical confidence intervals were calculated for criterion accuracy using different decision rules. Results from these analyses do not support the multi-facet hypothesis, which worsened criterion accuracy. Criterion accuracy was highest when the rank order scores were interpreted as a single issue. Unweighted combined decision accuracy for grand total scores was 87.6% excluding inconclusive results, and a 12.1% inconclusive rate. Results of the rank order model were compared to three other TDA algorithms: OSS-2, OSS-3, and a replication of the Probability Analysis algorithm. All three computer algorithms achieved decision accuracy levels over 90% with inconclusive rates less than 20%. Results based on rank order scores were also compared to results from a previous study in a cohort of inexperienced scorers evaluating the same confirmed case sample with the Empirical Scoring System. There was no advantage to the rank order model. Issues surrounding test item variance in rank order and nonparametric models are discussed. Possible improvements to a rank order scoring model were evaluated, and a weighted rank order model achieved an accuracy level of 90% with 19% inconclusives.

Introduction

The multi-facet hypothesis suggests that criterion accuracy would be highest for examinations conducted with the Federal Zone Comparison Test (ZCT) format for psychophysiological detection of deception (PDD) exams when the subtotal scores of the evidence-connecting question is scored independently from the other questions. Gordon (personal communication, 1/6/2011), suggested that the Federal ZCT is a multi-facet exam, due to the use of an evidence connecting relevant question (RQ), and posited that the first two RQs (R5 and R7) of the Federal ZCT represent a distinct single issue, while the third RQ (R10) represents a separate evidence connecting issue. An example of a direct involvement question would be "did you do it," where "it" describes the examinee's behavioral involvement in the issue of concern. An example of an evidence connecting question would be "did you help do it," "did you plan it," or "did you participate in it," where "it" describes the incident of concern in a manner that is detached from the action verbs which describe the examinee direct involvement. Gordon further suggested that the first two questions should be scored and evaluated independently of the third question, and that a deceptive result for the evidence-connecting questions should supersede the results of the first two questions if they are non-deceptive. According to Gordon, a non-deceptive result can be achieved by the subtotal scores of R5 + R7, as long as the score for R10 is not deceptive. Krapohl, Gordon & Lombardi (2008) published the results of a study using a sample of Federal ZCT exams and the rank order scoring model described by Gordon (1999), and by Gordon and Cochetti (1987), showing a combined decision accuracy of 84% using the rules and cut-scores proposed by Gordon, and 86% using the grand total score and optimal cut-scores described by Krapohl, Dutton and Ryan (2001). We evaluated Gordon's multi-facet hypothesis regarding the Federal ZCT using archival rank order scores from the Krapoh et al. study.

Rank order transformation models have at times been described as a "Horizontal Scoring System," and the Academy of Scientific Investigative Training (ASIT)

Horizontal Scoring System ([HSS], Gordon & Cochetti, 1987; Gordon et al., 2006). The term horizontal refers to either the arbitrary arrangement of the table matrix of scores of the observed reactions, or to the fact that rank order scores are assigned to physiological sensor data which are displayed along a horizontal x-scale or time axis. PDD scoring table matrices are correctly understood as three-dimensional numerical matrices, though not geometric, consisting of 1) multiple presentations, of 2) a series of test questions, and 3) an array of physiological sensors. Like other test data analysis (TDA) models, rank order scores of several iterations of a sequence of test questions are aggregated to achieve subtotal and grand total values. Values arranged horizontally could just as easily, and just as arbitrarily, be arranged vertically with no impact on the numerical test results. The descriptive term rank order is used throughout this paper, instead of the metaphorical term horizontal. The transformation method will be easily recognized by statisticians and model developers as a nonparametric rank transformation scheme in which the naturally occurring variance in response will be discarded and replaced with a uniform rank order variance. Rank order models, and nonparametric methods, although weaker in statistical power compared to parametric counterparts, are useful and informative when evaluating problematic data that do not conform to parametric assumptions or requirements.

Nonparametric Methods

Rank models are nonparametric and make no attempt to use the linear shape or parametric qualities of the data. Nonparametric coding models are therefore somewhat blunt, with weaker statistical power and less precision than their parametric counterparts. It is not uncommon to employ a combination of parametric and nonparametric methods, so long as parametric requirements are satisfied when parametric models are used. As a general principle, parametric models are preferred over nonparametric models.

Nonparametric statistical methods have been used in the past when parametric methods were unattainably complicated. However, the availability of powerful computers has substantially changed the ability to make complex calculations, and nonparametric are now primarily used only when data are of poor quality or when parametric requirements cannot be satisfied. It is inevitable, when removing or discarding noise or uncontrolled variance in this manner, that some potentially useful diagnostic information will also be lost. Information that is discarded is no longer available for analysis or interpretation. For example: in horse racing, rank transformations can be used to replace event timings to tell us which horses have placed first, second, or third in a race, but the placement ranks themselves have lost the ability to tell us the distance between the horses or the finishing times. When using a rank order coding scheme, test items will appear to be forced apart when the natural variance or strengths of reactions are very similar, and test items will be forced together when the natural variance or strength of reaction is very different.

All data can be thought of as a combination of diagnostic information that is accounted for or explained by the construct of interest, along with random uncontrolled variance that cannot be accounted for. This is sometimes referred to as signal and noise, but the underlying concerns pertain to the variance of the data. When conducting PDD examinations, the data of interest are physiological responses to psychological stimuli that are presented in the form of test questions to which the examinee will verbally answer “no” with the understanding that the answer will be interpreted as either truthful or deceptive regarding the examinees' denial of involvement in a behavioral concern. Diagnostic information must be empirically correlated with the criterion categories, and diagnostic features should be supported by statistical analysis and statistical modeling of their structural validity. The goal of any testing procedure is to make the diagnostic information accessible and reduce the influence of uncontrolled variance. In all forms of scientific testing, TDA is ultimately a matter of variance.

A test result can be classified with respect to the criterion when the variance of reactions to the test stimulus differs from normally expected variance at a statistically

significant level. It is therefore necessary to study and describe the normal variance of the data, so that statistical norms can be used to calculate the level of significance or probability that an individual test result has occurred due to error or random chance. In comparison question test (CQT) PDD examinations, diagnostic and screening efficiency is determined by calculating the level of statistical significance of measured or observed differences in the variance or strength of response to two different types of test stimuli: relevant and comparison stimuli.

Rank Order Transformations and PDD Examinations

Rank order TDA models have been previously described in the published literature (Gordon, 1999; Gordon & Cochetti, 1987; Gordon et al., 2006; Honts & Driscoll, 1987, 1988; Krapohl, Dutton & Ryan, 2001; Krapohl, et al., 2008; Miritello, 1999), however their use in field settings is less common than the seven-position scoring method (Backster, 1963a, 1963b; Bell, Raskin, Honts & Kircher, 1999; Department of Defense, 2006a; Swinford, 1999), and the three-position model (Blackwell, 1998; Department of Defense, 2006; Harwell, 2000; Krapohl, 1998; Van Herk, 1990). Rank order scores, and rank order variance, in CQT PDD exams are imposed on the population of responses to RQs and comparison questions (CQs) for each component sensor within each examination chart. Rank values are assigned to PDD scores for each component sensor (i.e., pneumograph, electrodermal [EDA], cardiograph), treating the two pneumograph sensors as a single channel. Questions that produce the strongest reaction are assigned the highest rank value, equivalent to the number of questions in the population (i.e., the number of combined RQs and CQs). Rank scores for RQs are not compared to the stronger of the nearest comparison as in other TDA models (Department of Defense, 2006). Instead of identifying the greater reaction within each pair of relevant and comparison stimuli, rank ordering requires examiners to evaluate the greatest of all reactions within a chart, then the second greatest reaction, then the next greatest reaction, etc., until all reactions are arranged in a hierarchy of reaction strength.

It is common in rank models to assign an average of the tied rank values when reactions are of similar value. Honts and Driscoll (1987, 1988) first described the use of tied ranks in a rank order PDD model, and Gordon (1999) later adopted the same procedure. The use of tied ranks is imposed because the rank order paradigm mandates that there can be only one item in each rank position even if two items are of equal value. The rank values are summed and divided by the number of tied values, and the resulting average or tied rank value is assigned to the items that produced the equivalent values. Neglecting this procedure would result in arbitrary judgments about which item to assign the strong rank value, and would potentially contaminate the results. Equally concerning would be the potential contamination that would result from the assignment of the higher rank score to both items, as this would result in the fabrication of increased integer rank scores. Rank order transformations, although conceptually simple, are procedurally complex in that they require the simultaneous evaluation of all test stimuli.

Two rank order TDA models have been described in the published literature. One model was described by Gordon (1999), Gordon, Fleisher, Morsie, Habib and Salah (2000), Gordon and Cochetti (1987), and also by Krapohl et al. (2008). The other model was described by Honts and Driscoll (1987, 1988). The two rank order models differ in their physiological features, cut-scores, and decision rules. The model proposed by Honts and Driscoll (1987, 1988) is based on primary physiological features for which there are multiple published studies that provide evidence of their statistical development and validity (Harris, Horner & McQuarrie, 2000; Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1988, 2002; Raskin, Kircher, Honts & Horowitz, 1988). These features are sometimes referred to as “Kircher features” (Dutton, 2000; Krapohl & McManus, 1999).

The rank order scores in this study were obtained using the features described by Gordon (1999), which appear to have been developed atheoretically and without published description of statistical analysis

or structural model coefficients which support their validity.

The rank order models, proposed by Honts and Driscoll (1987, 1988) and Gordon and Cochetti (1987) appear to differ in their mathematical transformations. In the rank order model introduced by Honts & Driscoll (1987, 1988), the rank order scores are partitioned into subtotal scores for the relevant question (RQ) scores and comparison question (CQ) scores. The RQ subtotal is then subtracted from the CQ subtotal. When the RQs produce generally stronger reactions, a phenomenon correlated with deceptive persons, the final score will be a negative integer value. When the CQs produce generally stronger reactions, a phenomenon correlated with truthful persons, the final score will be a positive integer.

Gordon (1999), Gordon et al. (2006) and Krapohl, et al. (2008) use a different procedure, in which RQs are assigned a negative sign value while CQs are designated with a positive sign value. Each RQ score is added to the score of the preceding CQ. Following that, subtotal scores are summed for each RQ and for the examination as a whole, if it is a single issue examination. The difference between these methods is procedural only. No integer scores are changed, created or lost in either of these methods. No sign values are changed, and the total resulting scores will be identical for the two rank order aggregation models as long as the examination consists of an equivalent number of relevant and comparison stimuli.

The hypothesized advantage of the more complex HSS (Gordon & Cochetti, 1987; Gordon, 1999; Gordon, et al., 2000) Gordon et al., (2006) method of handling the examination scores is the potential application of a rank order scoring model to multi-facet examinations for which the variance of evidence connecting RQs is thought to be independent. Independence, in the realm of scientific testing, refers to the notion that the variance of the individual test items is not influenced by the variance of other test items. Independent test items in PDD examinations are therefore thought to provide differential diagnostic information about the examinee's role or level of involvement in the issue under investigation.

Honts and Driscoll (1988) describe the comparison of each relevant question to the average of summed rank scores of the comparison questions, and provided data from a statistical analysis, recommending cut-scores of +/2 as an optimal solution, but ultimately concluded that rank order transformations offer no criterion advantages and are more complicated to execute than the seven-position and three-position transformation models based on the work of Backster (1963a, 1963b). Nelson, Krapohl & Handler (2008) described the application of a Kruskal-Wallace nonparametric ANOVA to evaluate between question variance when scoring multi-issue and multi-facet exams with the Objective Scoring System, version 3, algorithm. However, Nelson et al., (2008) employ the nonparametric ANOVA to evaluate variance only between the RQs, and not between the individual RQs and CQs. Neither Honts and Driscoll (1988) nor Nelson et al., (2008) describe a method for statistically evaluating independent variance among rank ordered question scores. Nelson et al. (2008) describe that they first evaluate the level of significance of the individual RQ sub-total. They then use the Kruskal-Wallace ANOVA as a second stage to show that test questions do not differ significantly before proceeding to make a classification of a screening test result as a whole.

In addition to the absence of statistical decision models and statistical classifiers, none of the previous studies on rank order scoring models for PDD have included any description of the statistical confidence intervals surrounding the reported accuracy estimates. Furthermore, previous studies have generally not included adequate descriptions of the normative data, with the exception of Krapohl, Dutton and Ryan (2001), from which statistical confidence intervals can be calculated. Statistical confidence intervals are important in the validation of scientific test methods because it is impossible to obtain data from the entire population of persons and accuracy calculations are therefore estimates, with corresponding levels of statistical confidence, of what would be observed if it were possible to test the entire population of persons.

Previous studies on rank order TDA models have sometimes used decision rules and cut-scores that are not based on scientific studies. Gordon and Cochetti (1987) reported the use of +/-18 cut-scores for a ZCT with three questions and three charts. Gordon (1999) and Gordon et al. (2006) describe cut-scores of +/-13 for total scores and +/-4.5 for subtotal scores. Krapohl et al. (2001), in an optimization study on rank order scoring, reported that cut-scores of 13/0, for grand total scores, provided optimal criterion validity. Later, Krapohl et al. (2008) published additional evidence that the 13/0 cut-scores provided better criterion accuracy than the +/-13 cut-scores.

Independence of the variance of test stimuli

All CQTs involve the evaluation of differential response between the RQs and CQs, and this is ultimately a matter of variance. Rank order models replace the natural variance of the population of all scored questions with a uniform rank variance. Rank variance is uniform in that the distance between items is always the same regardless of the natural variance or observed difference between items. Following the assignment of rank order scores, the rank order variance is partitioned into two portions: variance that describes the subset of RQs, and variance that describes the subset of CQs. Variance of these two groups can then be evaluated for statistically significant differences. Assumptions about the independence of test stimuli, as in multi-facet and multi-issue exams, require that the variance of RQs is further partitioned into variance belonging to the individual stimulus targets.

Rank order scoring models present nontrivial theoretical and statistical challenges when applied to multi-facet and multiple-issue examinations, for which Gordon (personal communication, 1/6/2011) has argued the test items vary independently. No mathematical solution has ever been described for the calculation of the statistical significance of independent variance of individual test items in a rank order scoring model. Rank order models are theoretically handicapped in that between-question variance is lost or nullified during the rank transformation. More importantly, rank order transformation violates the independence of test items by allowing the response magnitude variance of each question to affect the rank variance of every other test question. The variance of rank order scores are, by definition, non-independent. It is therefore unclear whether the expected improvement will result in increased criterion accuracy. Miritello (1999) described a procedure for rank ordering of individual questions, but provided no statistical decision model. Similarly, the procedural method described by Gordon (1999), for applying the rank order model to multi-facet and multi-issue exams, has no solution for the calculation of a statistical classifier of the rank ordered response variance at the subtotal or question level, and is therefore a sorting procedure only.

Krapohl et al. (2008) recommended research for further optimization of the rank order TDA model. Those suggestions included the possibility of component weighting, refinement of physiological features, evaluation of single issue and multi-facet examination formats, adjustment of cut-scores, and comparison with other algorithmic models. The present replication and extension of the Krapohl et al. (2008) study is intended to address these suggestions, in addition to Gordon's multi-facet hypothesis.

Method

Data

Archival scores, including subtotal and grand total scores, were obtained from the Krapohl et al. (2008) study. Scores for the Krapohl et al. (2008) study were provided by the third named author in that study, who was reportedly selected for his expertise in the use of the rank order scoring model described by Gordon (1999), Gordon and Cochetti (1987), and Gordon et al. (2006). Krapohl et al. (2008) reported their sample as size N = 100. However, 99 examination scores were provided to the investigators and first author of this extension study. Results of the missing case were coded as inconclusive for these analyses.

The examinations in this study (N = 100) were conducted using the Federal ZCT technique (DoDPI, 2006; Light, 1999), a PDD technique that is widely taught at polygraph

schools accredited by the American Polygraph Association (APA) and recognized by the American Association of Police Polygraphists (AAPP). The three-question Federal ZCT, based on the Backster ZCT (Backster, 1963a 1963b), is intended for event specific or evidentiary testing, and is considered to be among the most accurate diagnostic PDD techniques available at this time. All examinations consisted of three RQs, three probable-lie CQs, and three test charts.[1] RQs in the Federal ZCT are named R5, R7 and R10. Confirmation for all examinations exists in the form of extra-polygraphic evidence such as physical evidence of guilt or innocence, physical evidence of guilt of an alternative suspect, or the confession of an alternative suspect.

Analysis

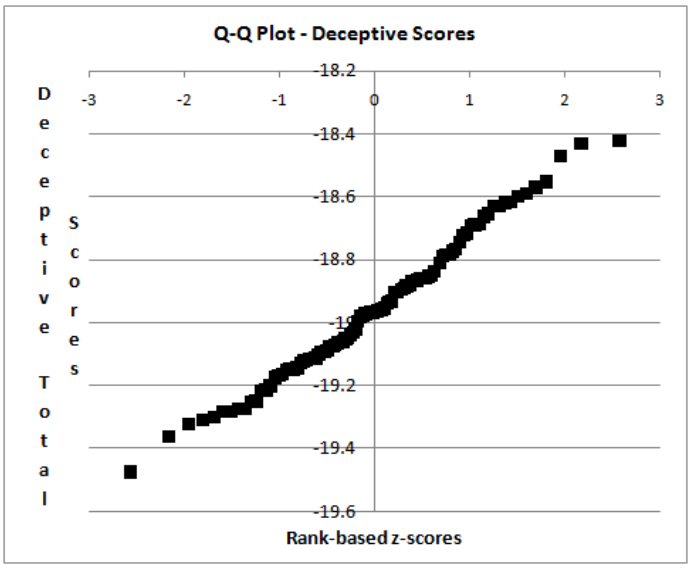
Rank-order scores were evaluated for normality, and normative data were calculated for use in a Gaussian-Gaussian signal discrimination model, as described by (Barland, 1985). Bootstrap resampling was also used to calculate the unbiased sample variance of dimensional profiles of criterion accuracy using different decision rules. Bootstrap variance statistics were used to calculate the sums of squares for a series of one-way and two-way ANOVAs that were used to investigate the statistical significance of the effects of the multi-facet hypothesis on criterion accuracy.

Results

Normative parameters

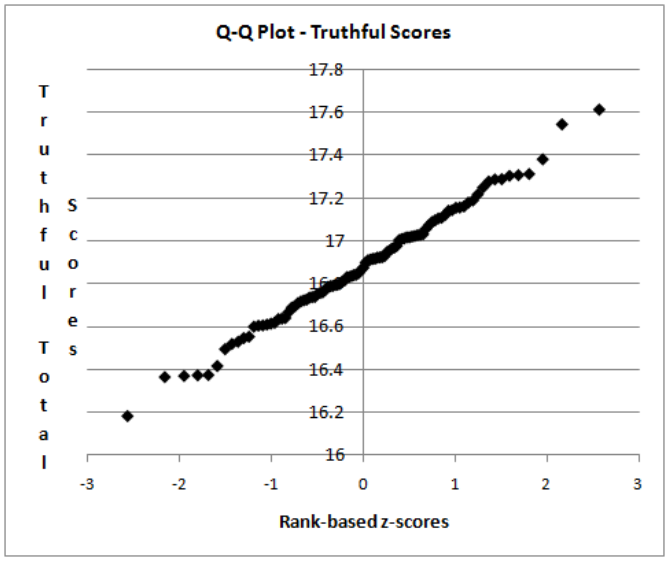
A bootstrap of 10,000 iterations was used to calculate a distribution of bootstrap mean and standard deviations that were evaluated for normality. Bootstrap distributions are known to be normally distributed when the underlying data are normally distributed. Quantile plots, shown in Figures 1 and 2, show that the distributions of bootstrap means were sufficiently normal to assume the underlying distribution of scores to be normally distributed, and to proceed with the calculation of statistically optimal cut-scores for use in a Gaussian-Gaussian signal discrimination model.

Figure 1. Q-Q plot for bootstrap mean deceptive scores.



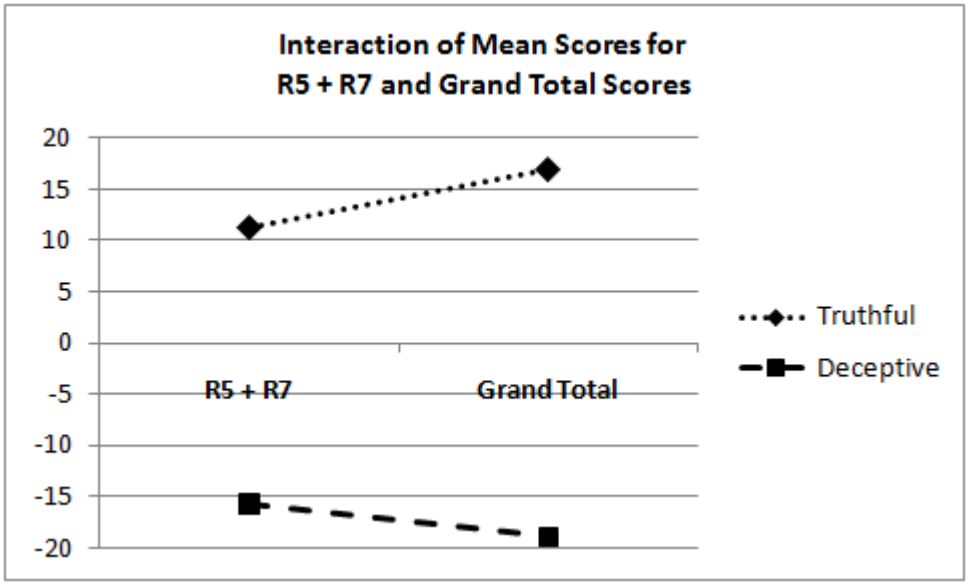
[1] Senter and Dollins (2004) showed that inconclusives can be reduced by recording up to two more test charts, with no reduction in decision accuracy, and current field practices allow for the completion of three to five test charts, with no change in decision rules or cut-scores.

Figure 2. Q-Q plot for bootstrap mean truthful scores.



The mean deceptive grand total score was -18.920 (SD = 16.529). The mean truthful grand total score was 16.948 (SD = 18.097). These values were mapped to the standard normal distribution to calculate lookup tables for the p-values and alpha levels for all possible cut-scores. Normative data for deceptive and truthful total scores are shown in Appendix A. Normative data were also developed for the sum of the first two RQs. The mean deceptive subtotal score for question R5 + R7 was -15.739 (SD = 12.095). The mean truthful subtotal score for R5 + R7 was 11.279 (SD = 15.296). Normative data for deceptive and truthful scores of R5 + R7 are shown in Appendix B. Figure 3a shows a plot of the interaction of mean scores for the truthful and deceptive cases of the R5 + R7 and Grand Total scores. A two-way ANOVA, model x status, showed that there was no statistically significant interaction and no significant main effects for total scores of the R5 + R7 + R10 model and the R5 + R7 model.

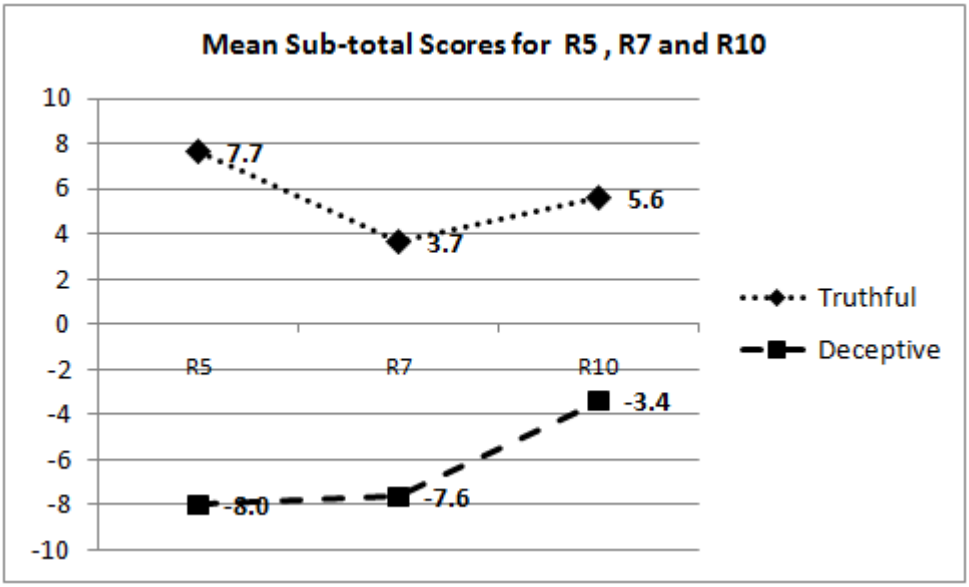
Figure 3a. Mean scores for R5 + R7 and Grand Total models.



Subtotal scores

The bootstrap mean for deceptive subtotals was -6.349 (SD = 8.794), and the bootstrap mean for truthful subtotals was 5.648 (SD = 9.134). Figure 3b shows the interaction of mean subtotal scores for the truthful and deceptive cases. A 2 x 3 ANOVA comparison, question x status, showed there was a significant interaction effect for the difference in the R5, R7 and R10 scores between the truthful and deceptive cases ($F_{1,294} = 5.980$, $p = .015$). The significant interaction of question and case status precludes any interpretation of the main effects without further analysis. Figure 3b shows that the pattern of cell means for the relevant questions was different for truthful and deceptive cases, and that R10 may produce a different pattern of reaction than the other questions. Scores for R10 shifted in a positive direction for both truthful and deceptive cases. These trends may have more to do with the order of the questions in the sequence than the test question language. It is also possible the effect has more to do with semantic language than the behavioral target. It may be possible to mitigate position related effects through the rotation of the questions within successive test charts. One-way post hoc ANOVAs showed that the within-group differences were not statistically significant for either group with ($F_{2,147} = 0.124$, $p = .884$) for the deceptive cases and ($F_{2, 147} = 0.057$, $p = .944$) for the truthful group.

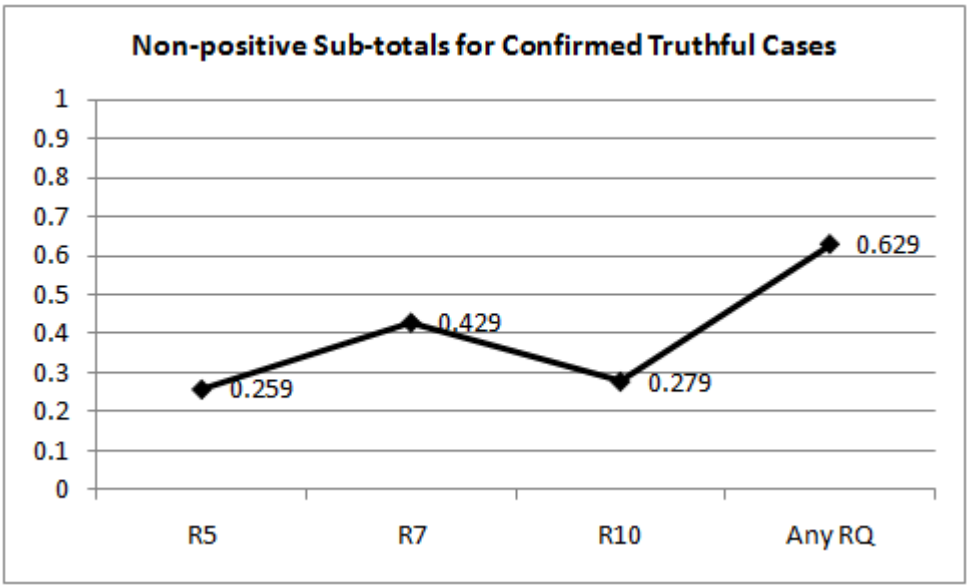
Figure 3b. Interaction of mean subtotal scores.



Non-positive subtotal scores for confirmed truthful cases

Because subsequent planned analysis evaluated decision rules for which the subtotal scores may be permitted to trump the total score when making deceptive classifications, subtotal scores were evaluated for their sign value. Evaluation of the rank order scores from the experienced scorer revealed that 63.9% (95% CI = 62.1% to 65.6%) of the confirmed truthful cases had at least one non-positive (i.e. zero or negative integer) subtotal score. The greatest proportion of non-positive scores occurred at R7, for which 41.9% of the truthful cases had a non-positive score. Figure 4 shows a plot of the non-positive subtotals for the confirmed truthful cases. A one-way ANOVA shows there was no statistically significant difference in the rate of non-positive subtotal scores for the three RQs ($F_{2,147} = 2.530$, $p = 0.080$). However, this result was approaching a statistically significant level, with more non-positive subtotal scores at the second of three RQs.

Figure 4. Proportion of non-positive subtotal scores for confirmed truthful cases.



Criterion accuracy and decision rules

Normative data were used to determine statistically optimal cut-scores, with alpha = .05 for both truthful and deceptive decisions. Cut-scores were determined for several decision rules, including: the Horizontal Scoring System (HSS) rule, Grand Total Rule (GTR), Two-stage Rules (TSR) (Senter, 2004; Senter & Dollins, 2008a, 2008b), the R5 and R7 Rule (R57), and the Spot Score Rule (SSR) (Light, 1999).

Decision rules and statistically optimal cut-scores for the HSS rule were as follows:

1. If any subtotal score is -22 (Bonferonni corrected alpha = .017 * 3 = .05) or lower, a decision of Deception Indicated (DI) is made,
2. If the sum of R5 and R7 is -14 or lower (alpha = .05), a decision of DI is made,
3. If the grand total scores is -13 (alpha = .05) or lower, a decision of DI is made,
4. If the subtotal of the first two RQ together is 5 or greater (alpha = .05), and no subtotal is -22 or lower, a decision of No Deception Indicated (NDI) is made,
5. If the grand total score is 9 (alpha = .05) or greater, a decision of NDI is made, and
6. All other results are inconclusive (INC).

Decision rules and statistically optimal cut-scores for the GTR were as follows:

1. If the grand total scores is -13 (alpha = .05) or lower, a decision of DI is made,
2. If the grand total score is 9 (alpha = .05) or greater, a decision of NDI is made, and
3. All other results are INC.

Decision rules and statistically optimal cut-scores for the TSR were as follows:

1. If the grand total scores is -13 (alpha = .05) or lower, a decision of DI is made,
2. If the grand total score is 9 (alpha = .05) or greater, a decision of NDI is made,
3. If the grand total is INC, and any subtotal score is -22 (Bonferonni corrected alpha = .017 * 3 = .05) or lower, a decision of DI is made, and
4. All other results are INC.

Decision rules and statistically optimal cut-scores for the R57 rule were as follows:

1. If the sum of R5 and R7 is -14 (alpha = .05) or lower, a decision of DI is made,
2. If the sum of R5 and R7 is 5 (alpha = .05) or greater, a decision of NDI is made, and
3. All other results are INC.

Decision rules and statistically optimal cut-scores for the SSR were as follows:

1. If any subtotal score is -9 (uncorrected alpha = .05) or lower, a decision of DI is made,
2. If all subtotal scores are 0 (corrected alpha = .05) or greater, a decision of NDI is made, and
3. All other results are INC.

A bootstrap of 10,000 resampled iterations of the N = 100 test results was used to calculate the standard deviations and statistical confidence intervals for the dimensional profiles of criterion accuracy achieved by the different decision rules. Bootstrapping was used because it allows for easy calculation of a variance statistic for the proportion of categorical results while using a single sample. Table 1 shows the bootstrap accuracy profiles for the five decision rules. Accuracy profiles include mean, standard deviation, and 95% confidence intervals for several dimensions of criterion accuracy, including: overall percent correct, total inconclusives (INC), inconclusive truthful cases (T INC), inconclusive deceptive cases (D INC), sensitivity to deception, specificity to truthfulness, false-negative rate (FN), false-positive rate (FP), positive predictive value (PPV), negative predictive value (NPV), percent correct for deceptive cases (D Correct), percent correct for truthful cases (T Correct), and the unweighted average of the percentage of correct decisions, excluding inconclusives, for truthful and deceptive cases. The unweighted average is more robust against

Table 1. Accuracy profiles for different rank order decision rules with alpha = .05.

	Mean (SD) [95% Confidence Intervals]				
	GTR .05/.05	**HSS .05/.05**	**TSR .05/.05**	**R57 .05/.05**	**SSR .05/.05**
Correct	.876 (.035) [.806 to .945]	.860 (.037) [.788 to .931]	.876 (.035) [.806 to .945]	.876 (.039) [.799 to .953]	.766 (.048) [.671 to .860]
INC	.121 (.032) [.057 to .184]	.080 (.027) [.028 to .133]	.121 (.032) [.057 to .184]	.281 (.044) [.194 to .367]	.230 (.042) [.148 to .313]
D INC	.059 (.034) [<.001 to .125]	.040 (.027) [<.001 to .093]	.059 (.034) [<.001 to .125]	.298 (.064) [.172 to .424]	.059 (.033) [<.001 to .124]
T INC	.179 (.053) [.075 to .283]	.119 (.045) [.031 to .208]	.179 (.053) [.075 to .283]	.258 (.061) [.138 to .379]	.397 (.069) [.261 to .533]
Sensitivity	.812 (.055) [.704 to .92]	.812 (.055) [.704 to .920]	.812 (.055) [.704 to .920]	.594 (.069) [.458 to .729]	.931 (.035) [.864 to >.999]
Specificity	.713 (.063) [.590 to .836]	.753 (.060) [.636 to .870]	.713 (.063) [.590 to .836]	.653 (.067) [.522 to .784]	.237 (.060) [.119 to .356]
FN	.118 (.046) [.029 to .208]	.138 (.049) [.042 to .234]	.118 (.046) [.029 to .208]	.098 (.043) [.015 to .182]	<.001 (<.001) [<.001 to <.001]
FP	.098 (.042) [.015 to .181]	.118 (.046) [.028 to .207]	.098 (.042) [.015 to .181]	.078 (.038) [.004 to .153]	.356 (.068) [.224 to .489]
PPV	.892 (.047) [.800 to .983]	.873 (.049) [.776 to .970]	.892 (.047) [.800 to .983]	.883 (.056) [.773 to .993]	.722 (.056) [.613 to .831]
NPV	.858 (.054) [.753 to .964]	.846 (.054) [.74 to .951]	.858 (.054) [.753 to .964]	.870 (.055) [.762 to .978]	>.999 (<.001) [>.999 to >.999]
D Correct	.873 (.049) [.777 to .969]	.855 (.051) [.754 to .955]	.873 (.049) [.777 to .969]	.858 (.060) [.74 to .976]	>.999 (<.001) [>.999 to >.999]
T Correct	.879 (.052) [.778 to .981]	.865 (.052) [.763 to .967]	.879 (.052) [.778 to .981]	.893 (.052) [.792 to .994]	.399 (.091) [.222 to .577]
Unweighted Avg.	.876 (.036) [.806 to .946]	.860 (.037) [.787 to .932]	.876 (.036) [.806 to .946]	.875 (.040) [.797 to .954]	.700 (.045) [.611 to .789]

difference in the sample size of deceptive and truthful cases and differences in test sensitivity and specificity; it is therefore considered to be a more generalizable estimate of accuracy than the simple percentage of correct decisions.

Evaluation of the means, standard deviations, and statistical confidence intervals in Table 1 revealed that the R57 model and SSR model differed significantly from the other models with significantly greater inconclusives for the R57 model, along with significantly greater inconclusives and weaker decision accuracy for the SSR model. The R57 and SSR models were removed from further analysis.

Figures 5 and 6 show the criterion accuracy levels for the HSS, GTR, and TSR. A series of 2 x 3 ANOVAs, case status x decision rule, was conducted to evaluate the differences in decision accuracy, error, and inconclusives rates for the three single issue decision rules: HHS, GTR, and TSR. No significant interaction or main effects were found for correct decision or errors. A significant interaction was found ($F_{1,294} = 14.956$, $p < .001$) for inconclusive results. Post hoc one-way ANOVAs showed there were no significant one-way effects, with ($F_{2,147} = 0.205$, $p = .814$) for the deceptive cases and ($F_{2, 147} = 0.001$, $p = .482$) for the truthful group. It is clear from Table 1 that inconclusive rates are significantly higher for truthful cases than deceptive cases.

Figure 5. Accuracy of rank order scores with deceptive cases using different decision rules.

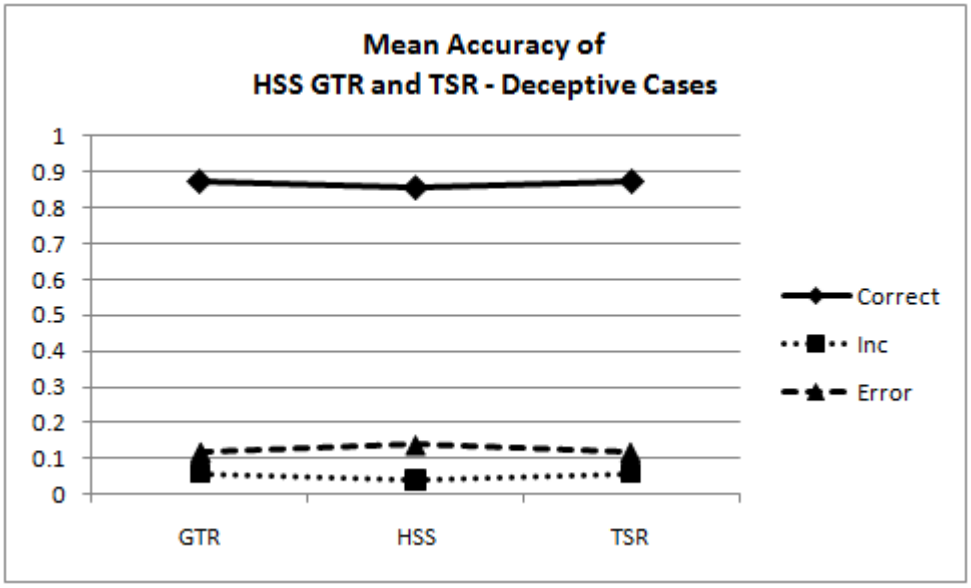
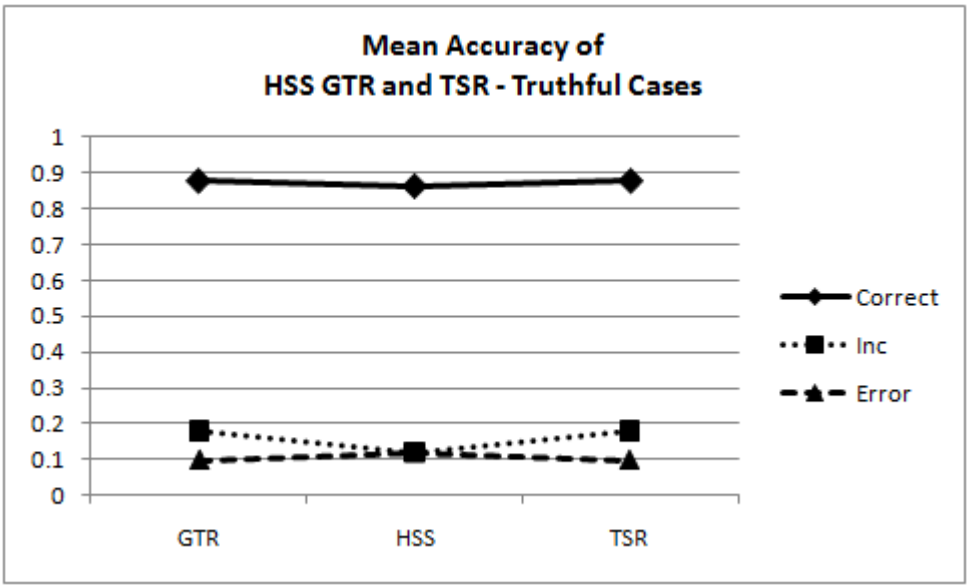


Figure 6. Accuracy of rank order scores with truthful cases using different decision rules.



Computer algorithm results

The confirmed case sample was then evaluated using three computer scoring algorithms: the Objective Scoring System, version 3 ([OSS-3], Nelson et al., 2008), the Objective Scoring System, version 2 ([OSS-2], Krapohl & McManus, 1999; Krapohl, 2002), and a replication of the Probability Analysis algorithm (Kircher & Raskin, 1988; 2002; Nelson et al., 2008; Raskin et al., 1988). Results of this analysis are shown in Table 2, along with the accuracy profile for the rank order model, and reveal that all three of the algorithms achieved decision accuracy levels over 90% and inconclusive rates less than 20%.

Table 2. Algorithm results with the confirmed case sample (N = 100) using recommended alpha levels.

	Percentile Scores			
	OSS-3 .1./05	**OSS-2 .07/.06[2]**	**PA replication[3] (.70/.30)**	**Rank Order GTR .05/.05**
Correct	.907 (.032) [.844 to .97]	.914 (.031) [.853 to .975]	.908 (.035) [.839 to .976]	.876 (.035) [.806 to .945]
INC	.059 (.025) [.010 to .108]	.143 (.035) [.074 to .211]	.167 (.037) [.095 to .239]	.121 (.032) [.057 to .184]
D INC	.081 (.038) [.007 to .155]	.124 (.045) [.036 to .213]	.171 (.054) [.065 to .277]	.059 (.034) [<.001 to .125]
T INC	.038 (.030) [<.001 to .096]	.160 (.050) [.062 to .258]	.161 (.052) [.060 to .263]	.179 (.053) [.075 to .283]
Sensitivity	.845 (.056) [.737 to .954]	.785 (.064) [.661 to .910]	.715 (.070) [.577 to .853]	.812 (.055) [.704 to .92]
Specificity	.862 (.049) [.765 to .958]	.783 (.054) [.677 to .889]	.799 (.061) [.680 to .918]	.713 (.063) [.590 to .836]
FN	.074 (.038) [<.001 to .149]	.09 (.041) [.011 to .169]	.113 (.051) [.013 to .214]	.118 (.046) [.029 to .208]
FP	.100 (.041) [.020 to .180]	.057 (.032) [<.001 to .120]	.040 (.027) [<.001 to .093]	.098 (.042) [.015 to .181]
PPV	.892 (.043) [.807 to .977]	.932 (.038) [.857 to >.999]	.946 (.037) [.873 to >.999]	.892 (.047) [.800 to .983]
NPV	.921 (.042) [.839 to >.999]	.897 (.049) [.801 to .992]	.877 (.055) [.769 to .985]	.858 (.054) [.753 to .964]
D Correct	.919 (.042) [.838 to >.999]	.897 (.047) [.804 to .989]	.863 (.06) [.745 to .982]	.873 (.049) [.777 to .969]
T Correct	.896 (.043) [.812 to .980]	.932 (.037) [.859 to >.999]	.952 (.033) [.887 to >.999]	.879 (.052) [.778 to .981]
Unweighted Avg.	.908 (.032) [.844 to .971]	.914 (.031) [.853 to .975]	.908 (.035) [.839 to .976]	.876 (.036) [.806 to .946]

Decision accuracy, inconclusive rates, and error rates for the algorithm results were subject to a series of 2 x 4 ANOVAs, model x status, for which the mean percentages of correct decisions, errors, and inconclusive results are shown in Table 3 and Figures 7 and 8. There was a significant interaction between the scoring algorithm and case status ($F\ 1,392 = 79.964$, $p < .001$) for correct decisions. One-way post hoc ANOVAs were completed to further investigate the difference in correct decisions for the algorithms within the truthful and deceptive groups. Within-group differences were not statistically significant for either group, with ($F\ 2,147 = 0.708$, $p = .548$) for the deceptive cases and ($F\ 2,147 = 1.433$, $p = .234$) for the truthful group, indicating there were no differences in overall decision accuracy for the four algorithm models.

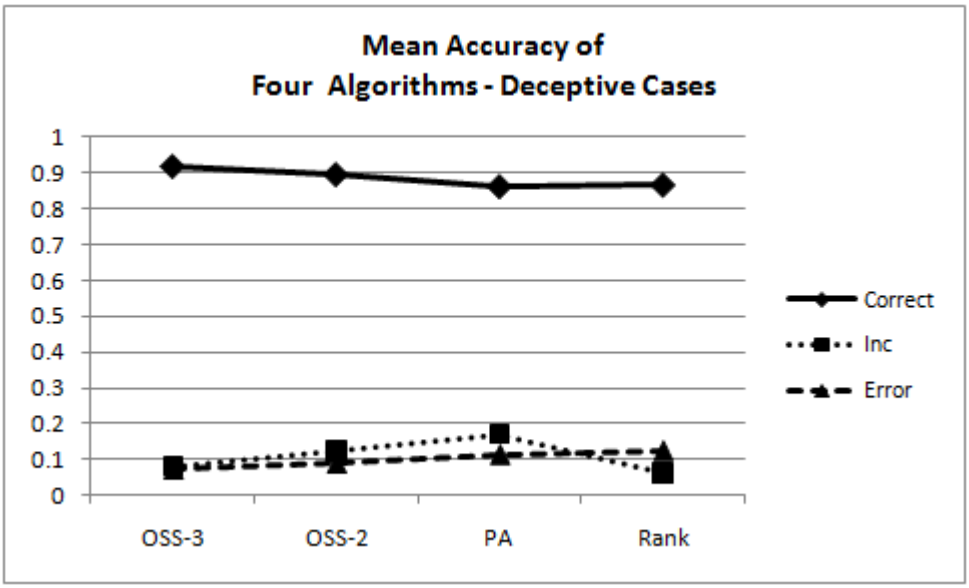
[2] Alpha boundaries of .07 and .06 correspond to traditional cutscores of +6 and -6, normally used with the OSS-2.

[3] The Probability Analysis algorithm, was replicated by Nelson et al., (2008) from information available in published studies. The developers of the Probability Analysis algorithm have not published their discriminate function. The replication was trained independently through discriminate analysis with the OSS development sample used by Krapohl and McManus, 1999.

Table 3. Correct decisions, errors, and inconclusive rates for four algorithms

Deceptive Cases				
	OSS-3	**OSS-2**	**PA**	**Rank Order**
Correct	0.919	0.897	0.863	0.867
Inc	0.080	0.124	0.171	0.063
Error	0.073	0.090	0.113	0.124
Truthful Cases				
	OSS-3	**OSS-2**	**PA**	**Rank Order**
Correct	0.895	0.932	0.951	0.876
Inc	0.038	0.160	0.161	0.180
Error	0.100	0.057	0.040	0.101

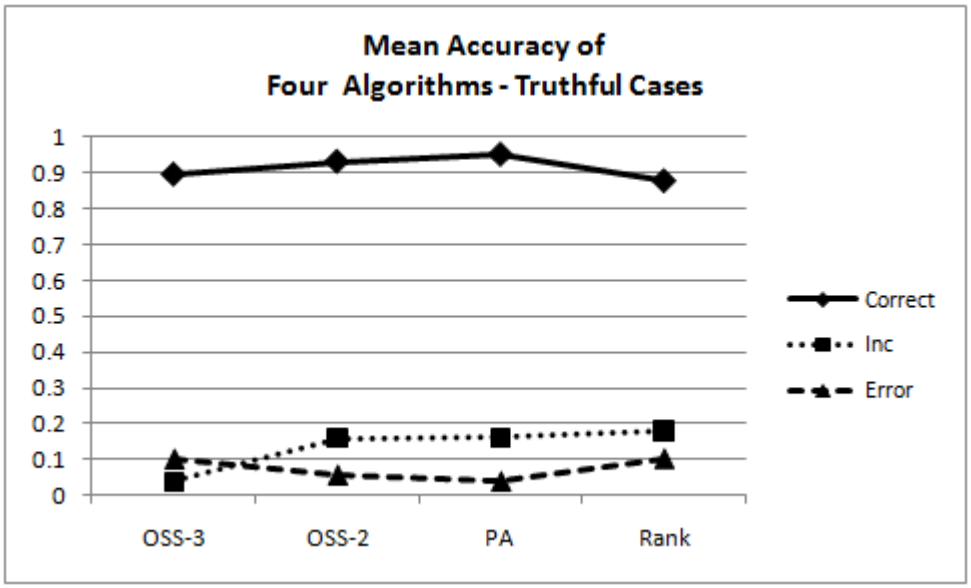
Figure 7. Mean decision accuracy of four algorithms with deceptive cases.



A second two-way ANOVA, algorithm x status, for errors showed there was a significant interaction between the scoring algorithm and case status (F 1,392 = 77.990, $p < .001$) for errors. Post hoc one-way ANOVAs showed that Within-group main effect differences in errors were not statistically significant for the deceptive cases (F 2, 147 = 0.656, $p = .580$) or for the truthful cases (F 2, 147 = 1.422, $p = .238$), suggesting that overall error rates did not differ significantly for the four algorithms.

A third two-way ANOVA, algorithm x status, for inconclusive results showed there was a significant interaction between the scoring algorithm and case status when evaluating inconclusive results (F 1,392 =181.029, $p < .001$). The main effects for scoring algorithm were not statistically

Figure 8. Mean decision accuracy of four algorithms with truthful cases.



significant at the .05 level but were approaching statistical significance (F 1,392 = 3.154, p .077). Post hoc one-way ANOVAs showed there were statistically significant differences in inconclusive rates for the four algorithms for both groups, with (F 2,147 = 2.699, p = .047) for the deceptive cases and (F 2,147 = 5.039, p = .002) for the truthful group. The OSS-3 algorithm produced low inconclusive rates for both truthful and deceptive groups, while the OSS-2 and PA algorithms had higher inconclusives for both groups. The rank order algorithm produced low inconclusives for deceptive cases and higher inconclusive rates for truthful cases. A one-way ANOVA showed that the OSS-3 algorithm produced significantly fewer inconclusives for truthful cases (F 1,98 = 5.360, p = .022) when compared to the rank order model.

Comparison of the rank order model with the Empirical Scoring System.

Rank order scores from the experienced scorer were then compared to the scores from a previous study by Handler, Nelson, Goodson and Hicks (2011) that involved a cohort of 19 inexperienced polygraph examiner trainees who used the Empirical Scoring System ([ESS], Blalock, Cushman & Nelson, 2009; Krapohl, 2010; Nelson & Handler, 2010; Nelson, Blalock, Oelrich & Cushman, 2011 in press; Nelson & Krapohl, 2011; Nelson et al., 2008) to evaluate the same confirmed case sample.[4] Table 4 shows the accuracy profiles for the ESS and rank order models, and Figures 9 and 10 show the interaction of the percentage of correct decisions with inconclusives, errors, and inconclusive results.

[4] Nelson, the principal investigator and first author of the present study, has published studies on the ESS but has no financial or proprietary interest in the ESS. Nelson is a psychotherapist and field polygraph examiner employed as a researcher with the Lafayette Instrument Company which provides computer programming expertise and sales support for the ASIT HSS algorithm. Nelson is also the principal investigator and developer of the Objective Scoring System, version 3, a free and open source computer scoring algorithm for which he has no financial or proprietary interest. Handler, the second author, has also published studies on the ESS and OSS-3 algorithm, as a principal investigator and second author, and also has no proprietary interest in the OSS-3 or the ESS.

Table 4. Comparison of rank order and ESS criterion accuracy profiles

Mean (SD) [95% CI]		
	Rank Order GTR .05/.05	**ESS .10/.05**
Correct	.876 (.035) [.806 to .945]	.902 (.030) [.842 to .958]
INC	.121 (.032) [.057 to .184]	.030 (.017) [<.001 to .070]
D INC	.059 (.034) [<.001 to .125]	.029 (.024) [<.001 to .087]
T INC	.179 (.053) [.075 to .283]	.031 (.024) [<.001 to .087]
Sensitivity	.812 (.055) [.704 to .920]	.867 (.049) [.762 to .959]
Specificity	.713 (.063) [.590 to .836]	.883 (.045) [.784 to .960]
FN	.118 (.046) [.029 to .208]	.088 (.040) [.020 to .178]
FP	.098 (.042) [.015 to .181]	.102 (.044) [.021 to .196]
PPV	.892 (.047) [.800 to .983]	.908 (.040) [.821 to .980]
NPV	.858 (.054) [.753 to .964]	.896 (.044) [.804 to .977]
D Correct	.873 (.049) [.777 to .969]	.895 (.046) [.800 to .978]
T Correct	.879 (.052) [.778 to .981]	.909 (.041) [.814 to .980]
Unweighted Average	.876 (.036) [.806 to .946]	.902 (.030) [.839 to .958]

Figure 9. Accuracy for HSS and ESS with deceptive cases

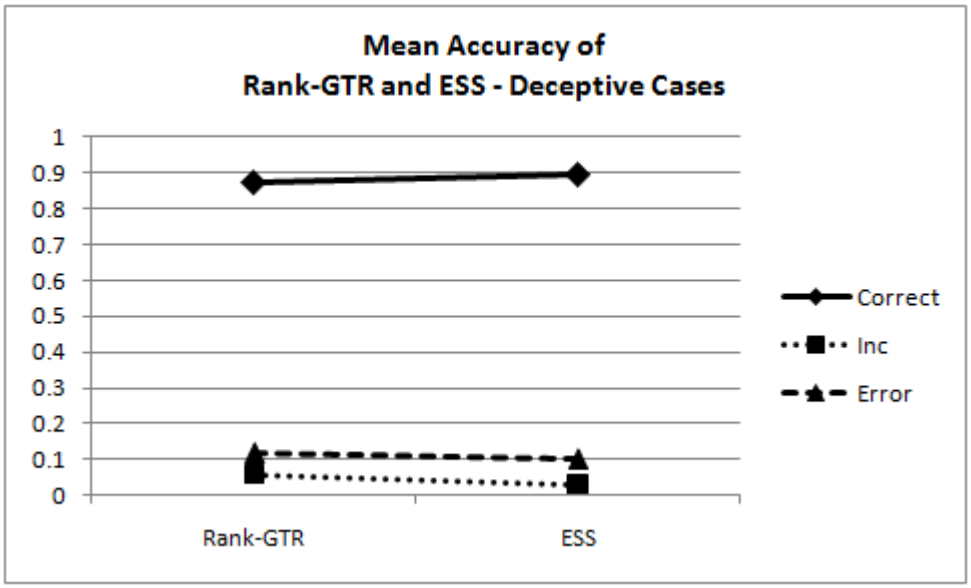
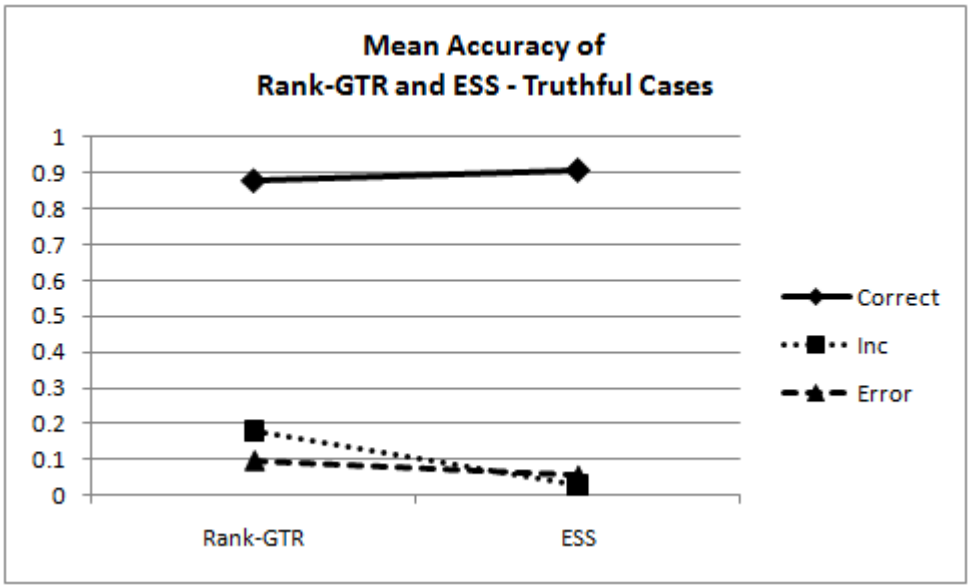


Figure 10. Accuracy for HSS and ESS with truthful cases



There were no statistically significant interaction effects or main effects for decision accuracy or errors for the rank order and ESS scoring models. However, there was a statistically significant interaction between scoring model and case status for inconclusive results ($F_{1,196} = 144.171$, $p < .001$). Also, the main for scoring model effect was approaching a significant level for inconclusive results, ($F_{1,196} = 3.102$, $p = .080$). Post hoc ANOVAs were completed to further investigate the difference in inconclusive rates for the ESS and rank order models within the truthful and deceptive groups. Within-group differences were not statistically significant for the deceptive group ($F_{2, 147} = 0.467$, $p = .496$). Inconclusives were different at statistically significant rates for the truthful group ($F_{2, 147} = 6.607$, $p = .012$). The ESS model produced significantly fewer inconclusives with truthful cases.

Potential improvements to the rank order scoring model

The confirmed case sample was evaluated using a rank order scoring model based on Kircher features, for which there is replicated evidence of validity (Harris et al., 2000; Kircher & Raskin, 1988, 2002; Kircher, et al., 2005; Krapohl & McManus, 1999; Raskin et al., 1988). Data were measured using the Extract software (Harris, 1998), which makes automated measurements of the Kircher features, which were then assigned rank order scores via an automated implementation of the rank order model described by Gordon (1999), and Gordon and Cochetti (1987).

The EDA has been shown to provide the strongest diagnostic signal (Handler, Nelson, Krapohl & Honts, 2010; Harris & Olsen, 1994; Harris, et al., 2000; Kircher & Raskin, 1988, 2002; Kircher et al., 2005; Krapohl & McManus, 1999; Raskin et al., 1988). The results of a linear discriminate analysis were published in a previous study (Nelson, Krapohl & Handler, 2008), showing the optimal weighting coefficients to be as follows: pneumograph = .19, electrodermal = .53, cardiograph = .28). In consideration of the emphasis on simplicity at the integer level, all EDA rank scores were multiplied by 2 before the rank scores were partitioned into relevant and comparison groups. This practical result of doubling all EDA scores was to increase the average mathematical weight placed on the EDA data to ~.50.

A bootstrap resample of 10,000 iterations was completed to calculate variance statistics and normative distribution parameters from the single sample. Truthful cases produced a mean weighted rank score of 14.9 (SD = 23.5), while the deceptive cases produces a weighted mean rank score of -34.9 (SD = 26.5). Appendix C shows the normative data for the weighted mean rank total scores. Weighting the EDA data produced a larger change in the normative data for the deceptive group than the truthful group, moving the deceptive mean score further from zero.

Two weighted mean rank models were evaluated. One model with alpha = .05 for deceptive and truthful classifications, with cut-scores at -24 and +9 for deceptive and truthful decisions. The other weighted rank model used alpha = .10 for deceptive and truthful classifications, with cut-scores at -16 and 0 for deceptive and truthful decisions. Table 5 shows the accuracy profiles for the two weighed rank models, along with the accuracy profile for the unweighted rank model. Figures 11 and 12 show the interaction of mean percentiles for decision accuracy, errors and inconclusives for the weighted and unweighted rank order models.

Table 5. Accuracy profiles for a weighted rank order models

Weighted Rank Order Model, mean, (SD), [CI]			
	Unweighted Rank GTR .05/.05	**Weighted Rank – GTR .05/.05**	**Weighted Rank – GTR .10/.10**
Correct	.876 (.035) [.806 to .945]	.938 (.031) [.878 to .998]	.901 (.036) [.831 to .971]
INC	.121 (.032) [.057 to .184]	.362 (.047) [.27 to .453]	.194 (.036) [.123 to .266]
D INC	.059 (.034) [<.001 to .125]	.322 (.056) [.212 to .431]	.190 (.050) [.091 to .289]
T INC	.179 (.053) [.075 to .283]	.399 (.077) [.248 to .55]	.198 (.052) [.097 to .3]
Sensitivity	.812 (.055) [.704 to .92]	.637 (.060) [.519 to .755]	.727 (.063) [.602 to .851]
Specificity	.713 (.063) [.590 to .836]	.563 (.076) [.415 to .711]	.725 (.061) [.606 to .845]
FN	.118 (.046) [.029 to .208]	.042 (.029) [<.001 to .099]	.083 (.046) [<.001 to .173]
FP	.098 (.042) [.015 to .181]	.038 (.024) [<.001 to .086]	.076 (.036) [.006 to .146]
PPV	.892 (.047) [.800 to .983]	.944 (.037) [.871 to >.999]	.904 (.047) [.811 to .996]
NPV	.858 (.054) [.753 to .964]	.931 (.048) [.837 to <.999]	.899 (.054) [.792 to 1.005]
D Correct	.873 (.049) [.777 to .969]	.938 (.043) [.855 to >.999]	.897 (.056) [.788 to 1.006]
T Correct	.879 (.052) [.778 to .981]	.937 (.04)0 [.858 to >.999]	.905 (.044) [.818 to .992]
Unweighted Average	.876 (.036) [.806 to .946]	.938 (.030) [.878 to .997]	.901 (.035) [.832 to .97]

Figure 11. Accuracy of weighted and unweighted rank order models with deceptive cases.

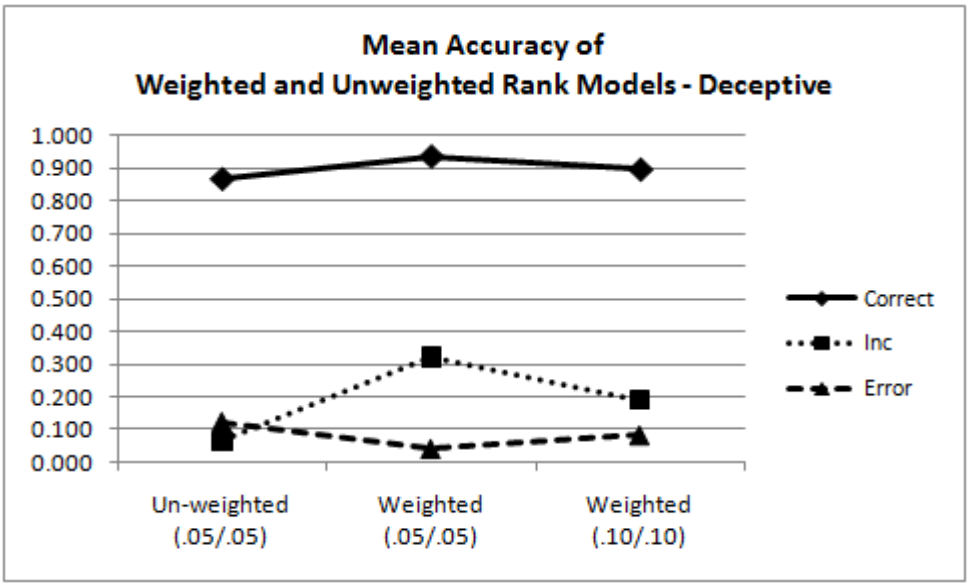
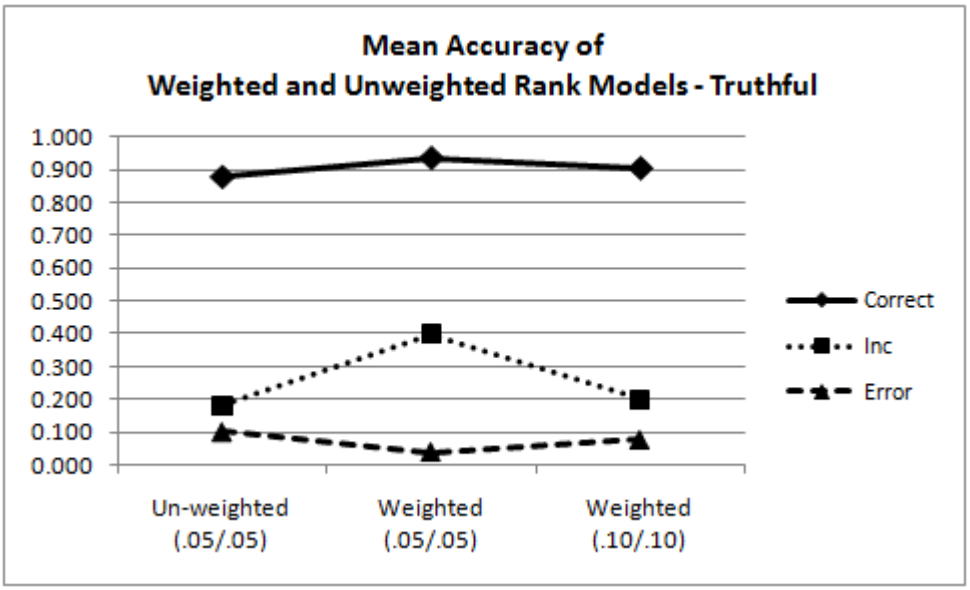


Figure 12. Accuracy of weighted and unweighted rank order models with deceptive cases.



A series of 2 x 3 ANOVAs, status x model, was completed to evaluate any differences between the weighted and unweighted rank order models using the GTR. There were no significant interactions and no significant main effects for correct decisions.

The interaction of model and status was significant for errors (F 1,294 = 33.508, p

< .001). The main effect for model was approaching a statistically significant level (F 1,294 = 2.735, p = .099). Post hoc one-way ANOVAs showed that within-group differences in error rates were not statistically significant for the truthful group (F 2,147= 0.657, p = .520), though the weighted models did produce fewer FP errors. Error rates were different at statistically significant rates for the deceptive cases (F 2,147 = 3.232, p = .042). Table 12 shows that the use of electrodermal weighting can be expected to produce fewer FN errors.

There was a significant interaction of model and status for inconclusive results (F 1,294 = 51.643, p = .001), and the main effect for model was also significant (F 1, 294 = 10.087, p = .002). Post hoc ANOVAs for within-group differences were statistically significant for both groups with (F 2,147 = 14.003, p = <.001) for the deceptive, and (F 2,147 = 8.547, p = <.001) for the truthful group. The weighted model with alpha = .05 produces more inconclusives than the other models.

Discussion

These results do not support the multi-facet hypothesis regarding rank order scores with the Federal ZCT. No independent rank order variance was found among the different RQs within the groups of truthful and deceptive cases. Results from these analyses do not support the hypothesis that interpreting between-question response variance will increase criterion accuracy, and do not support the hypothesis that R10 represents a distinct issue that should be scored separately from the other questions.

Evidence of support for the multi-facet hypothesis would have to be observed as improved criterion accuracy when using specialized decision rules intended to partition and interpret independent between-question variance and differential meaning of the evidence connecting question. Instead, HSS decision rules weaken the criterion accuracy of the Federal ZCT. Criterion results using the SSR provide additional evidence that rank order numerical transformations are ineffective at partitioning and using independent between-question variance to increase test accuracy. The SSR does produce a high level of test sensitivity and low FN errors, but the cost is weak overall test accuracy and test specificity that is so low that the chance of a truthful person passing the test is significantly less than chance (p < .001). The high proportion of non-positive subtotal scores among the confirmed truthful cases (63.9%) will mean that any decision rule that permits a subtotal score to supersede the grand total score could result in test specificity levels that are weaker than chance.

The GTR, involving the sum of the R5, R7 and R10 subtotals, provides the highest level of criterion validity, as measured by decision accuracy, inconclusives, and error rates for the truthful and deceptive groups. It is unlikely that field examiners will be able to achieve any increase in criterion accuracy through attempts to make use of observed differences in reactions to individual RQs in the Federal ZCT when using the rank order model. Observed differences in RQs may be attributable to the position of the question in the test question sequence, and may also be due to semantic language and not the behavioral concern. Positional effects might be mitigated by rotation of the RQs in subsequent test charts. Additional research is needed in this area.

Although the multi-facet hypothesis is not supported, a number of important findings do begin to emerge from the results of this study. These analyses show that a rank order model is capable of producing numerical scores for which the variance of grand total scores is sufficiently normally distributed to develop normative data that can be used to calculate an inferential level of statistical significance or probability of error for individual test results. Normative data can be used to make evidence-based decisions about the selection of statistically optimal cut-scores that will satisfy requirements for decision accuracy and inconclusive rates.

Results from these analyses prompt a question as to whether the failure of the multi-facet hypothesis to increase criterion accuracy is attributable to deficiencies in the ability of a rank order transformation model to effectively partition between-question variance, or to deficiencies in the effectiveness of multi-facet questions to elicit

between-question variance. A tentative answer can be formulated through a convergence of existing knowledge. It is known that rank order transformation schemes systematically replace the assumed independent between-question variance with non-independent rank order variance before attempting to evaluate between-question variance. It is also known that the TSR does reduce inconclusives with other TDA models. The absence of any criterion effect for the TSR with rank order scores is therefore cautiously interpreted as an artifactual result of the overall bluntness of rank order transformations and the replacement of between-question variance with uniform rank variance for which the variance of each question affects every other question.

The high rate of non-positive subtotal scores among the truthful cases provides further evidence that the rank order model is ineffective at partitioning and using between-question variance to increase criterion accuracy. If the study data are considered to be representative of rank order scores that can be expected in field circumstances then it may be unwise with our current evidence to employ decision rules that allow the subtotal scores to supersede the grand total score.

An array of presently available computer algorithms appear to provide criterion accuracy that is as good as or better than the results achieved by the single expert scorer who provided rank order scores for the Krapohl, Gordon and Lombardi (2008) study. There were no significant differences in decision accuracy or error rates for the rank order model and three algorithms: OSS-3, OSS-2, and a replication of the Probability Analysis algorithm. However, there are differences in inconclusive rates among the four TDA models, with fewer inclusive results using the OSS-3. The absence of differences in accuracy and error rates indicates the observed differences inconclusives are related to increases in test sensitivity and test specificity rates for the OSS-3 algorithm. This should be the focus of future research.

Although there is some observed effect for improved decision accuracy for the ESS, compared to the rank order model, the difference is not significant. The rank order model has significantly more inconclusives for truthful cases compared to the ESS, and this corresponds to weaker test specificity. Of great importance is that the ESS scores were obtained not from a single experienced scorer but from a cohort of inexperienced polygraph examiner trainees. The results of a single expert scorer may be less likely to generalize to field settings, in which skill and experience vary considerably, than results from a cohort of inexperienced scorers.

Weighted and unweighted models produce significantly different criterion accuracy profiles, and the results indicate that component weights may affect errors and inconclusives differently among truthful and deceptive cases. Weighting the EDA data produces a larger change in the normative data for the deceptive group than the truthful group, moving the deceptive mean score further from zero. There is no significant effect for differences in decision accuracy for the weighted and un-weighted models. There is, however, a significant reduction in false-negative errors for the weighted EDA rank order model, indicating that weighting the EDA data increases test accuracy for deceptive examinees. All rank order models in this study produce more FN errors than FP errors, and the difference is greatest for the unweighted model. Weighted models produce higher rates of inconclusive results than the unweighted model. This difference is loaded on the deceptive cases and was greater when the data were evaluated at conservative alpha boundaries (a = .05). This is interpreted as an artifact of the bluntness of the nonparametric rank order transformation model.

The optimal weighted rank model is achieved with alpha = .10 for both deceptive and truthful cases. This model achieves a bootstrap mean rate of correct decisions at 90% with 19% inconclusives. It is unlikely that further optimization or the use of asymmetrical alpha boundaries will produce more favorable results. More conservative alpha boundaries do not increase criterion accuracy and only increase the occurrence of inconclusive results. We again attribute this to the overall bluntness of the rank order transformation model.

Limitations

The present study, like all studies, is limited in some unavoidable ways, the first of

which is the sample itself which was constructed as a matched random sample selected from the confirmed case archive held at the Department of Defense. Extant information includes only simple demographics including age and gender of the examinee, general health and medications. Available information also includes the agency, and decision results from the original examiner and quality control personnel, along with the type of information available to confirm the status of each case. It is known that some of the results for the original examiners were incorrect and that some of the cases represent false-positive or false-negative error potentials in field settings. Some field examiners, including Gordon (1/6/2011, personal communication), have registered comments about a high level of difficulty in scoring the sample examinations. The fact that a number of scoring models, including a rank order model, and scores with widely varied levels of experience, have achieved a high level of criterion accuracy serves to reduce our concerns about the difficulty and representativeness of the sample data. For the purpose of acquiring new knowledge from the present study however, the sample is assumed to be representative.

A second, more troubling, limitation involves the cohort of participants who provided scores for this study. A single selected expert scorer cannot be considered representative of the population of average examiners working in various field settings with varying levels of experience, supervision, and continuing education. Criterion results and normative data based on scores from a cohort of scorers with a more average level of experience and training can be expected to generalize more effectively to field settings than norms based on data obtained from a single highly regarded expert. Additionally, it is not possible to study interrater reliability with a single scorer. None of the existing published studied on the rank order model, with the exception of Honts and Driscoll (1987), include any evidence or statistical description of interrater agreement, which limits our ability to consider it as a generalizable scoring model.

A related limitation is that the rank order model requires more than a simple dichotomous choice between the stronger of two reactions. It requires a complex process involving the comparison and ranking of all reactions at once. This raises concerns about reliability if the procedures for measurement and ranking are not automated. As there is no remedy for this limitation, other than automating the measurement and ranking procedures, we accept optimistically that the scores from a single experienced examiner are still somewhat informative.

An overarching limitation exists in the form of naive modeling, in which normative data are developed from the same data that is used to provide evidence of model effectiveness. Naive modeling is considered to be optimistic, and there is an increased tendency for the model and normative data to "overfit" a single sample used for both development and validation of normative data. The result will be an optimal fit between a normative model and the data used to demonstrate the effectiveness of the normative model. The results of these conditions will be weak generalizability when compared with models for which validity is demonstrated using a holdout or validation sample that is independent of the development sample.

Computer based analytic models such as bootstrap resampling, Monte Carlo models and randomized methods do not correct sampling deficiencies. Bootstrap resampling was used in this analysis to calculate statistics such as sampling variance and statistical confidence intervals that cannot otherwise be obtained from a single sample without exhaustive difficulty. These statistics are used to describe the range of expected bias and potential variance that can be expected in validation experiments that use other data. It is always preferable to perform model development activities on one sample and then complete subsequent validation samples using a separate sample. It is reasonable to expect some shrinkage or loss of effectiveness and criterion accuracy rates when the normative data and cut-scores are applied to a different sample of rank order scores from a different cohort of scorers. These analyses should be replicated in future studies based on independent sample data.

Another limitation of the present study is that the normative data developed in

this study pertain only to event-specific ZCT examinations with three RQs, three CQs, and three test charts. Application of these statistical norms to other techniques is not warranted. Additionally, the rank order model requires an equal number of relevant and comparison questions, and is considered non-robust against missing, artifacted or uninterpretable data. Future research should evaluate alternatives for achieving some form of replacement that will improve the robustness of the rank order model.

Some readers will inevitably note that the ASIT HSS was designed for use with the Integrated Zone Comparison Technique (IZCT (Gordon et al., 2000, Gordon et al., 2006)[5] which adds many complex assumptions to the basic structure of the single issue ZCT. This study is not intended to evaluate the IZCT, and is limited to an extended analysis of the previous study by Krapohl et al. (2008). To the extent that the basic structure and principles of the event-specific three-question ZCT questions sequence is considered simple and robust, it is hoped that the information in this study can provide some useful and generalizable information. In the strictest sense, the application of these normative data to the IZCT is unknown. Although caution is always warranted, there is little to be gained from a rigid perspective that prevents the acquisition and application of new knowledge. To the extent that both the Federal ZCT and the IZCT are event specific single issue examination formats, the use of these normative data may be somewhat justifiable in the absence of normative data specific to the IZCT.

Another, more troubling, complication regarding the application of these statistical norms to the IZCT will be that the IZCT testing procedures involve the reversal of the comparison-relevant question sequence during the third test chart (Gordon et al., 2000, 2006), resulting in relevant questions that are presented following a neutral question and before a comparison stimulus. It is theorized by Gordon et al. that reversal of the presentation of comparison and relevant stimuli reduces testing bias. However, Gordon et al. provide no data to support this claim and neglect to acknowledge that no test can be biased until the results are interpreted using normative data. The expected effect from the reversal of the comparison-relevant question sequence will be a shift of scores in the negative direction, increasing test sensitivity to deception. Corresponding decreases in test specificity, along with increased FP errors and increased inconclusives among truthful cases, can also be expected under these circumstances. Krapohl (2006) reported that relevant questions immediately preceded by neutral questions produce significantly lower scores than relevant questions preceded by comparison questions for both truthful and deceptive examinees. In this respect, the IZCT violates the basic principle of structural validity that comparison stimuli should precede the relevant stimuli. Because the normative data in this study were developed from Federal ZCT examinations, which conform to the basic principle of presenting comparison before relevant stimuli, the application of these norms to examinations that include the altered question sequence may not be justifiable. However, in consideration of the robustness of the event-specific three-question ZCT, there may be little difference between the effectiveness of the Federal ZCT and the IZCT if the IZCT is conducted without the alteration of the comparison-relevant sequence in the third test.

It is unknown to what degree normative data for a weighted rank order model based on Kircher features (Kircher & Raskin, 1998, 2002; Krapohl & McManus, 1999; Raskin, et al., 1988) will generalize to examinations scored using the alternative features described by Gordon (1999), for which there is no published statistical evidence of their structural coefficients or structural validity. Examination scores that

[5] The IZCT is a proprietary PDD format that is taught at and used by examiners trained at the Academy for Scientific Investigative Training, for which Gordon, the primary author of the 1987, 1999 and 2006 studies, and second author of the Krapohl et al (2008) study, has a financial and proprietary interest.

are derived from alternative physiological features may lead to unpredictable results. In consideration of the volume of scientific evidence in support of the validity and effectiveness of the Kircher features, the suggestion for use of an alternative feature model without substantial scientific proof seems ill-advised. A conservative argument would hold that scores obtained from an alternative feature model should not be evaluated using normative data based on Kircher features. A more generous argument might suggest that regardless of the unsubstantiated assumptions of the alternative features, the alternative features appear to place primary emphasis on the same robust primary features as the Kircher features. When considering the bluntness of rank order models, it may be unlikely that different rank order scores would be achieved by the alternative physiological features. Regardless of their practical similarity, an evidence-based approach to PDD test validation would have to require that any suggestion to use an alternative feature model must be supported not only by proven theoretical assumptions but by a volume of published and credible research evidence at least equivalent to that which supports the validity of the Kircher features. Construct and structural validity, and the structural coefficient model, of the alternative physiological features from Gordon (1999) should be the focus of future research, and this should be completed before advocating the use of the alternative features in field settings.

Another unstudied concern will be the effect of conducting additional PDD test charts when the results of the first three charts are inconclusive. Field testing protocols for seven-position and three-position numerical scoring (DoDPI, 2006a; DoDPI, 2006b, Handler, 2006; Handler & Nelson, 2008; Raskin & Honts, 2002) allow the collection of additional test charts if the results are inconclusive for the first three charts. Field practices for other TDA models do not involve changing the decision thresholds when data are collected from additional test charts. When considering that rank order total scores tend to be much larger than scores from other numerical scoring models, normative data for rank order scores from three charts should not be considered generalizable to exams that include additional test charts. Future research should investigate the normative data and optimal cut-scores for scoring additional test charts with the rank order model.

Recommendations

Of great importance is the potential for improvements to the rank order TDA model, and a number of recommendations can be made from this study. Until such time as there is evidence of increased criterion or construct validity for an alternative set of physiological features, rank order models should employ the Kircher Features. Weighting the EDA component will also improve the criterion accuracy profile of the rank order model, as does the use of statistical norms and optimal cut-scores. Because the criterion accuracy level of the rank order TDA model hangs precariously close to the boundaries of 90% or better decision accuracy with 20% or fewer inconclusives, changing or neglecting to use statistically optimal cut-scores will most likely result in suboptimal results, including the potential for decreased decision accuracy, increased errors, and increased inconclusives. Decision rules should be based on the simple and robust grand total rule, which also provides the highest level of criterion accuracy. Attempts to use subtotal scores when interpreting the results of rank order scores are not justified mathematically, and are not supported by these data.

Future research should focus on the refinement of improved statistical norms for manual scores of a weighted rank order model. Normative data should be obtained from a cohort of scorers for whom it can be reasonably anticipated their scores will be representative of, and generalizable to, the average examiner working in field settings. Attempts to use a single expert scorer to calculate generalizable criterion levels or normative data are misguided. Future research should also further investigate the validity of the multi-facet hypothesis regarding the Federal ZCT when scored with other TDA models. Additionally, the observed proportion of non-positive subtotals among the truthful cases in this study suggests the need for similar analysis with other samples

and other TDA models. Because the normative data from this study pertain only to three-question ZCT examinations, future research on rank order TDA models should develop normative data for different PDD exam formats. All developers of PDD TDA models face an obligation to develop and publish normative data and statistical evidence of both structural and criterion validity. It will be increasingly important to provide evidence in support of the validity of all aspects and all assumptions that underlie a PDD TDA model.

Conclusions

Rank order transformation models, while statistically simple, may increase procedural complexities in field settings, as they require field examiners to make a larger number of comparative decisions than do simpler transformation models. Despite their procedural complexity rank order models are considered blunt, and nonparametric methods are known to have weaker statistical power compared to parametric alternatives. In addition, rank order models lack both mathematical justification and evidence in support of the multi-facet hypothesis for decisions based on subtotal scores. A simpler procedural model may provide important advantages over a more complex model if the criterion accuracy rates can be demonstrated to be as good or better. Of course, computers can automate complex procedures with perfect reliability.

Although the use of computerized statistical and development methods will become increasingly common, it will remain important that field testing protocols intended for manual TDA tasks be distilled to their simplest procedural solutions. Procedural simplicity will help to optimize interrater reliability, skill acquisition, and skill retention. As the number of procedural requirements increase, so do the opportunities for errors and interrater disagreement. Correspondingly, interrater agreement is known to increase as procedural requirements decrease. A minor increase in procedural demands may be acceptable if there are corresponding significant increases in criterion accuracy.

In this study, three computer algorithms and one other manual scoring protocol produced raw frequencies that exceeded those of the rank order model; though the differences were not significant. The procedurally complex rank order model did not outperform any of the other models in terms of raw frequencies or statistically significant increases on criterion accuracy.

Finally, although the evidence does not support the multi-facet hypothesis, and suggests that the rank order model provides no advantages over other TDA models, the results of this extension study do support the validity of the rank order TDA model as potentially capable of providing a high level of decision accuracy and inconclusive rates that are within acceptable limits when applied to the Federal three-question ZCT. Results of this study further suggest that rank order TDA models are most effective when based on scientifically valid features, validated structural models involving weighted EDA scores, and grand total decision rules with cut-scores determined via statistical norms. It will be important to replicate this study with a cohort of less experienced scorers before endorsing these results and statistical norms as highly generalizable to field settings.

References

Backster, C. (1963a). Do the charts speak for themselves? New standards in polygraph chart interpretation. *Law and Order*, 11, 6768, 71.

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

Barland, G. H. (1985). A method of estimating the accuracy of individual control question polygraph tests. In, Antiterrorism; forensic science; psychology in police investigations: Proceedings of IDENTA'85 (pp. 142-147).

Bell, B. G., Raskin, D. C., Honts, C. R. & Kircher, J.C. (1999). The Utah numerical scoring system. *Polygraph*, 28(1), 19.

Blackwell, J. N. (1998). PolyScore 33 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. Reprinted in *Polygraph*, 28, (2) 149-175.

Blalock, B., Cushman, B. & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38, 281-288.

Department of Defense (2006a). Federal psychophysiological detection of deception examiner handbook. Reprinted in *Polygraph*, 40(1), 266.

Department of Defense (2006b). Test Data Analysis: DoDPI numerical evaluation scoring system. Retrieved from http://www.antipolygraph.org/documents/federalpolygraphhandbook02102006.pdf on 3312007.

Gordon, N. J. (1999). The academy for scientific investigative training's horizontal scoring system and examiner's algorithm system for chart interpretation. *Polygraph*, 28, 56-64.

Gordon, N. J., Fleisher, W. L., Morsie, H., Habib, W. & Salah, K. (2000). A field validity study of the integrated zone comparison technique. *Polygraph*, 29, 220-225.

Gordon, N. J. & Cochetti, P.M. (1987). The horizontal scoring system. *Polygraph*, 16, 116-125.

Gordon, N. J., Mohamed, F. B., Faro, S. H., Platek, S. M., Ahmad, H. & Williams, J.M. (2006). Integrated zone comparison polygraph technique accuracy with scoring algorithms. *Physiology & Behavior*, 87, 2514.

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2010 in press). Empirical Scoring System: A Cross-cultural replication and extension study of manual scoring and decision policies. Polygraph, 39 , [in press].

Handler, M., Nelson, R., Krapohl, J. & Honts, C. (2010). An EDA primer for polygraph examiners. *Polygraph*, 39, 68-108.

Harris, J. (1998). Extract.exe. [software developed for the Department of Defense Polygraph Institute].

Harris, J. C. & Olsen, D.E. (1994). *Polygraph automated Scoring System*. Patent Number: 5,327,899. U.S. Patent and Trademark Office.

Harris, J., Horner, A. & McQuarrie, D. (2000). An evaluation of the criteria taught by the Department of Defense Polygraph Institute for interpreting polygraph examinations. Johns Hopkins University, Applied Physics Laboratory. SSDPORPOR007272.

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

Honts, C. R. & Driscoll, L.N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16, 241-257.

Honts, C. R. & Driscoll, L.N. (1988). A field validity study of rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17, 115.

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

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

Kircher, J. C., Kristjansson, S. D., Gardner, M. K. & Webb, A. (2005). Human and computer decision-making in the psychophysiological detection of deception. University of Utah.

Krapohl, D. J. (1998). A comparison of 3 and 7 position scoring scales with laboratory data. *Polygraph*, 27, 210-218.

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

Krapohl, D. J. (2002). Short report: Update for the objective scoring system. *Polygraph*, 31, 298-302.

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

Krapohl, D. J., Dutton, D. W. & Ryan, A.H. (2001). The rank order scoring system: Replication and extension with field data. *Polygraph*, 30, 172-181.

Krapohl, D., Gordon, N. & Lombardi, C. (2008). Accuracy demonstration of the Horizontal Scoring System using field cases conducted with the federal Zone Comparison Technique. *Polygraph*, 37, 263-268.

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

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

Miritello, K. (1999). Rank order analysis. *Polygraph*, 28, 74-76.

Nelson, R. & Handler, M. (2010). *Empirical Scoring System*. Lafayette Instrument Company.

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

Nelson, R. & 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*, 40.

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.

Raskin D C, H. C. R. (2002). The Comparison Question Test. In M. Kleiner (Ed.), *Handbook of Polygraph Testing*. San Diego: Academic Press.

Raskin, D., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988). A study of the validity of polygraph examinations in criminal investigations. Final Report, National Institute of Justice, Grant No. 85IJCX0040.

Senter, S. & Dollins, A. (2004). Comparison of question series and decision rules: A replication. *Polygraph*, 33, 223-233.

Senter, S. M. & Dollins, A.B. (2008a). Exploration of a two-stage approach. *Polygraph*, 37(2), 149-164.

Senter, S. M. & Dollins, A.B. (2008b). Optimal decision rules for evaluating psychophysiological detection of deception data: an exploration. *Polygraph*, 37(2), 112-124.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph*, 28, 10-27.

Van Herk, M. (1990). Numerical evaluation: Seven point scale +/6 and possible alternatives: A discussion. *The Newsletter of the Canadian Association of Police Polygraphists*, 7, 28-47. Reprinted in *Polygraph*, 20(2), 70-79.

Appendix A

Normative data for rank order grand total scores of three question ZCT examinations

Mean Truthful Score: +16.9 (SD = 18.1)
Mean: Deceptive Score: -18.9 (SD = 16.5)

Total Deceptive Scores		Total Truthful Scores	
Score	**p-value**	**Score**	**p-value**
-25	0.010	20	0.009
-24	0.012	19	0.011
-23	0.014	18	0.013
-22	0.016	17	0.015
-21	0.018	16	0.017
-20	0.021	15	0.020
-19	0.023	14	0.023
-18	0.027	13	0.027
-17	0.030	12	0.031
-16	0.034	11	0.035
-15	0.039	10	0.040
-14	0.044	9	0.046
-13	0.049	8	0.052
-12	0.055	7	0.058
-11	0.061	6	0.066
-10	0.068	5	0.074
-9	0.076	4	0.083
-8	0.084	3	0.092
-7	0.093	2	0.103
-6	0.102	1	0.114
-5	0.113	0	0.126
-4	0.124	-1	0.139

Appendix B

Normative lookup data for rank order R5 + R7 scores

Mean Truthful Score: +11.3 (SD = 15.3)
Mean: Deceptive Score: -15.7 (SD = 12.1)

R5 + R7 Deceptive Scores		R5 + R7 Truthful Scores	
Score	**p-value**	**Score**	**p-value**
-25	0.009	20	0.002
-24	0.011	19	0.002
-23	0.013	18	0.003
-22	0.015	17	0.003
-21	0.017	16	0.004
-20	0.020	15	0.006
-19	0.024	14	0.007
-18	0.028	13	0.009
-17	0.032	12	0.011
-16	0.037	11	0.014
-15	0.043	10	0.017
-14	0.049	9	0.020
-13	0.056	8	0.025
-12	0.064	7	0.030
-11	0.073	6	0.036
-10	0.082	5	0.043
-9	0.092	4	0.051
-8	0.104	3	0.061
-7	0.116	2	0.071
-6	0.129	1	0.083
-5	0.144	0	0.097
-4	0.159	-1	0.112

Appendix C

Normative data for weighted rank order scores

Mean Truthful Score: +14.9 (SD = 23.5)
Mean: Deceptive Score: -34.9 (SD = 26.5)

Weighted Deceptive Scores		Weighted Truthful Scores	
Score	**p-value**	**Score**	**p-value**
-25	0.045	25	0.012
-24	0.049	24	0.013
-23	0.053	23	0.014
-22	0.058	22	0.016
-21	0.063	21	0.017
-20	0.069	20	0.019
-19	0.074	19	0.021
-18	0.080	18	0.023
-17	0.087	17	0.025
-16	0.094	16	0.024
-15	0.101	15	0.023
-14	0.109	14	0.032
-13	0.117	13	0.035
-12	0.126	12	0.038
-11	0.135	11	0.042
-10	0.144	10	0.045
-9	0.154	9	0.049
-8	0.164	8	0.053
-7	0.175	7	0.057
-6	0.186	6	0.061
-5	0.198	5	0.066
-4	0.210	4	0.071
-3	0.223	3	0.076
-2	0.235	2	0.082
-1	0.249	1	0.088
-0	0.262	0	0.094
-1	0.276	1	0.100
-2	0.290	2	0.107