Literature Review and Analysis of the Multi-facet Hypothesis and the Evaluation of Independent Target Questions

Raymond Nelson,

Mark Handler,

and David C. Raskin

Abstract

Published literature was reviewed for studies investigating the Multi-Facet Hypothesis (MFH). It postulates responses to PDD target questions vary independently when the questions address different levels of involvement in known or alleged incident. Implicit within the MFH is responses to individual questions serve to discriminate deception and truth-telling at rates greater than chance. The MFH suggests PDD test questions will not only discriminate guilty from innocent persons, but may also discriminate the behavioral role or level of involvement of a guilty person. The MFH is essentially the hypothesis of a multiple issue diagnostic exam, and is based on decision rules using question subtotal scores. Overall accuracy with the MFH was (.78) and lower than accuracy using grand total scores (.89). These results provided limited support for high test sensitivity with guilty persons when classifications were made under the MFH, but with test specificity for innocent persons not exceeding the range of chance. In contrast, decisions using grand total scores provided test sensitivity and specificity that were significantly greater than chance. The hypothesis of effective role discrimination using multiple issue diagnostic exams is inconsistent with the clear trend in the published scientific literature and is therefore not supported. Furthermore, these results do not support polygraph field practices that attempt to determine that a person has been deceptive to one or more questions and also truthful to one or more questions within a single examination. This literature review and associated analysis supports the evidence that multi-facet questions are non-independent.

The authors are indebted and grateful to Dr. Joe Stainback and Dr. Stewart Senter for their reviews, edits, comments and critique on earlier versions of this manuscript.

The Multi-Facet Hypothesis (MFH) posits that responses to psychophysiological detection of deception (PDD) test questions vary independently when the test questions address different behavioral roles or different levels of involvement in a known or alleged incident. This hypothesis suggests that the effectiveness of discrimination of guilt and innocence can be optimized through the use of decision rules that treat the questions independently. Multi-facet exams are a form of diagnostic exam[1] conducted in the context of known or alleged events. They differ from other event-specific diagnostic examinations in that they attempt to assess both guilt vs. innocence and the level of involvement of guilty examinees. The MFH holds that asking about different roles or actions is sufficient for responses to vary independently[2]. For example, a thief, accomplice, and lookout may all be involved in a crime though they have different roles. Multi-facet test questions may also be formulated to describe either primary/direct involvement or secondary/indirect involvement, such as knowledge of the identity of the guilty criminal, plans or details regarding a known or alleged crime, or the details of a known or alleged crime. Similarly, use of words such as *helping, planning and participating* in a crime may be thought of as three different roles (or different levels of involvement) that may be distinct from, or secondary to, primary or direct involvement in a behavioral action such as a theft, robbery, assault, or other crime.

Field practices have traditionally emphasized the scoring and interpretation of multi-facet criminal investigation exams in the same manner as multiple-issue screening exams for which both criterion variance and response variance[3] of the target questions are assumed to be independent (Department of Defense, 2006; Department of Defense, 2006b). In practical terms, the MFH is the hypothesis of a multiple issue diagnostic exam .[4] The null-hypothesis to the MFH will maintain that the use of different action verbs is insufficient to achieve independent response variance to PDD test stimuli, and that interpreting responses to test questions with an assumption of independent variance does not optimize the discrimination between guilt and innocence, compared to interpretation with no assumption of independence, nor does it identify the behavioral role or level of involvement among guilty persons at rates greater than chance (Department of Defense, 2006; Department of Defense, 2006b).

---

[1]Similar to other forms of testing, psychophysiological detection of deception (PDD) examinations can be thought of as belonging to the category of either diagnostic or screening exams (American Polygraph Association, 2011; Krapohl & Stern, 2003). PDD diagnostic tests are limited in scope to a single known or alleged behavioral incident, and interpreted at the level of the test as a whole. In contrast, PDD screening examinations are those tests conducted in the absence of a known or alleged problem. Screening exams are often formulated to simultaneously investigate multiple issues of concern for which the external criterion state is assumed to vary independently.

[2]An assumption of independent variance, in the scientific context, holds that responses to individual questions are influenced by no factors in common with each other – that responses to different stimuli have no shared sources of variance. This assumption is inherently compromised by the fact that responses to all individual questions have in common the examinee. The practical result of this is that it is not possible to conclude, with any scientific confidence, that an examinee has been deceptive to one or more questions while being truthful to other questions within the same examination. All comparison question test formats represent a form of single-subject, repeat-measures design, in which differential responses to different types of test stimuli permit the analysis of response variance with each examinee serving as a control set for oneself.

[3]Response variance is distinct from criterion variance, and criterion variance and criterion state refer to the actual disease state of the issue we are attempting to diagnose, (i.e., guilt or innocence). Variance can be thought of as either independent (i.e., having no causal factors in common with, and therefore not influenced by, the other test items) or non-independent (i.e., subject to potential influence from factors in common with other test items) when the requirements for independence are not satisfied.

[4]The MFH in field practice has been a source of ambiguity and discussion. And example of this is the Federal ZCT format, for which the third relevant question is formulated as an evidence-connecting or indirect involvement question. Despite the fact that results for the Federal ZCT format are given at the level of the test as a whole, there has been discussion among field practitioners about whether this format is an example of a multi-facet exam, with the implication that the third relevant question may vary independently. As with the MFH in general, validity of this idea would require evidence of increased accuracy as a result of interpretation at the question subtotal compared to results at the level of the test as a whole.

Deception and truth-telling are amorphous phenomena that are without physical substance and which therefore cannot be physically measured. Response features used when scoring and interpreting PDD examinations are criteria that are correlated with deception and truth-telling, and can be recorded, aggregated, normed, and interpreted categorically based on the probabilistic strength of the information. Numerous published studies have shown that responses to relevant and comparison stimuli vary significantly as a function of the criterion states of deception and truth-telling (APA, 2011; Honts, Thurber & Handler, 2020; Kircher & Raskin, 1988; National Research Council, 2003; Offe & Offe, 2007; Podlesny & Truslow, 1993; Raskin, Kircher Honts & Horowitz, 1988). PDD results are both categorical and probabilistic statements for which response features serve as proxy data that are correlated with the criterion states of the issues being tested – similar to the way that hormones or antibodies can serve as proxies for the presence of a pregnancy or disease state. All correlations are imperfect, and all physiological responses are multi-purposed (i.e., all physiological activities serve multiple functions and therefore have multiple external correlates. An inevitable feature of probabilistic information is that test data is always a combination of diagnostic variance, also referred to as controlled variance, explained variance or signal information, and error variance, also referred to as random variance, uncontrolled variance or noise.

In contrast to probabilistic tests and probabilistic models, are those that are deterministic for which randomness, uncertainty, and uncontrolled variance play no role. Deterministic models are based on physical substance and thus offer the potential for mechanical measurement – subject only to measurement error. Probabilistic test results are based on proxy data that cannot be interpreted with deterministic expectations of perfection, and for this reason scientific tests are not expected to be infallible. Understanding the accuracy of probabilistic models is a matter of quantifying the degree of uncertainty associated with a model or test result. All probabilistic test results based on proxy data will include some margin of potential error. Scientific tests achieve their effectiveness by quantifying whether the margin of error conforms to stated requirements for accuracy, generally expressed in the form of the tolerance for error or *alpha boundary*. Practical effect sizes can be described as the observed or expected proportions of correct or incorrect test results, and can also be described as the degree of improvement over chance outcomes. The MFH suggests that error variance can be minimized through the assumption that test question response variance varies independently, commensurate with the criterion state. Implicit within the multi-facet independence hypothesis is the expectation that responses to individual relevant stimuli will also vary significantly from responses to comparison stimuli as a function of guilt or innocence[5].

In field practice, assumptions about the independence of reactions to relevant questions are operationalized through the use of decision rules that use subtotal scores. Ideally, these subtotal scores would be considered with cut-scores that satisfy requirements for statistical significance. Using the subtotal scores for multi-facet or multiple issue exams, the examinee is considered deceptive if the subtotal score for *any* individual relevant question equals or exceeds the normative cut-score, while truthful classifications are made when *all* subtotal scores equal or exceed the normative cut-score for a truthful result. If

[5] PDD test effectiveness, as in other forms of diagnosis and screening, is achieved through development and validation of structural models composed of statistically optimal combinations of physiological response features that have been shown individually to have significant correlation with the criterion. Implicit in this correlation is an understanding that no relationship between test data and criterion can be uniform or perfect, and that test results are probabilistic and not deterministic. Test data analysis, test accuracy, and test validation are a process of developing predictive classification models by quantify differential responses in physiology that occur in response to test stimuli. Measured physiological responses are aggregated mathematically and used to make probabilistic calculations of how well a test result fits our expectations according to a statistical reference model that may represent either the null-hypothesis or the hypothesis. The principles of scientific hypothesis testing and decision theory can then be used to make categorical and probabilistic classifications of the test results regarding deception and truth-telling.

neither of these two conditions is satisfied, the result is considered inconclusive. [Refer to Nelson (2018a) for a discussion of various decision rules used in polygraph field practice.] If the subtotal score for any question meets the criterion for deception, then the numerical results of other question subtotal scores that do not meet the criterion for deception are uninterpretable.[6] This contrasts with decisions based on the grand total score, in which deceptive and truthful classifications are made by comparing the grand total score to normative cut-scores and statistical reference models without regard to subtotal scores.[7]

The MFH suggests several potential benefits, including: 1) increased discrimination between deception and truth-telling, 2) discrimination at the level of the individual target questions regarding the behavioral role or level of involvement among guilty examinees, 3) increased confidence that a reasonable and complete array of stimuli was presented to the examinee during testing, and 4) improved context setting for post-polygraph discussion, interview, and investigation. The third and fourth of these hypothesized effects can be thought of as *staging effects* intended to clarify for examinees and referring professionals the semantic and logical meaning of the test stimuli. The literature review performed herein is limited to the first two, which can be thought of as *criterion effects* because they are related to criterion accuracy, (i.e., whether decisions at the level of the individual questions can increase the criterion accuracy of the test, and whether the test can determine involvement or non-involvement in the distinct behavioral targets of the investigation). Criterion accuracy can be measured along several dimensions, including decision accuracy with guilty and innocent cases, sensitivity and specificity rates, false-positive and false-negative error rates, inconclusive rates with guilty and innocent persons, and Positive and Negative Predictive Values (outcome confidence in the test result). Consideration of these dimensions of test accuracy is important because different risk-assessment and risk-management contexts may prioritize different aspects of the risk-benefit ratios that may be informed by the test result.

The following research questions were investigated in this literature review: 1) does the interpretation of test data at the level of the individual target question provide any increase or advantage to criterion accuracy, 2) do subtotal scores for individual target questions discriminate deception and truth-telling, and 3) do subtotal scores for individual target questions discriminate the behavioral role or level of involvement of guilty examinees. Investigation of these effects will require the identification of published studies that reported results using both grand total and subtotal scores.

Validity of the MFH will be supported if published evidence reveals increased test effectiveness as a result of the interpretation of physiological responses at the level of the individual target questions when compared to the interpretation of responses at the level of the test as a whole. The null-hypothesis (responses to individual questions do not vary independently and decisions based on subtotal scores do not optimize the discrimination of deception and truth-telling) can be rejected if subtotal scores discriminate deception vs. truth-telling at rates significantly greater than chance, and if increases in test accuracy are observed as a result of the scoring and interpretation of subtotal scores in comparison to results based on grand total scores. The related hypothesis regarding discrimination of the behavioral role or level of involvement of guilty persons can be supported and the corresponding null hypothesis (responses to individual questions do not effectively discriminate the behavioral role or

[6]The decision rule is described here in the context of a norm-referenced statistical hypothesis test. However, traditional cut-scores appear to have sometimes been determined through a combination of empirical and heuristic analysis without regard for statistical reference distributions or the level of statistical significance. It is possible that test accuracy might be further optimized through the use of statistically determined norm-referenced cut-scores.

[7]Combinations of the SSR and GTR have also been described, including the simultaneous use of the SSR and GTR (Department of Defense, 2006), and the sequential combination of the SSR and GTR (Bell, Raskin, Honts & Kircher, 1999; Handler & Nelson, 2008; Senter, 2003; Senter & Dollins, 2003;). All assumptions about the criterion, including assumptions about independence and non-independence, are embedded in pragmatic and procedural test operations.

level of involvement of guilty persons) can be rejected if the observed combinatoric decision accuracy rate for individual questions is significantly greater than chance for both deceptive and truthful classifications.

History and evolution of independent target questions

Keeler (1930) provided one of the earliest examples of the use of a multi-facet target selection and question formulation approach.[9] He described a procedure in which four or five irrelevant [10] questions are presented, e.g., *Do you live in Chicago?* Following the irrelevant questions, several relevant questions would be asked that described multiple facets of a crime incident. Keeler gave the following as examples of relevant questions: *Did you dine with Jones Tuesday night?, Did you return to Jones' apartment that night?, Did you owe Jones some money?, Did you discuss this indebtedness?, Have you ever been in California?, Did you shoot Jones?, and Do you own a Savage forty-five?*

Keeler (1933) described a burglary in Los Angeles during 1923, in which the owner of a second story apartment came home to surprise a burglar attempting to open a safe. Reportedly, the burglar attempted to open a window to the fire escape, became entangled in the curtains before shooting and killing the owner, and then left through the door. Keeler described the presentation of the following questions that had no bearing on the case: *Is your name Jones?, Do you live in Los Angeles?, Do you own an automobile? The following relevant questions were also presented: Do you live on Maple Street? (the street of the burglarized apartment.) Do you live in a second story apartment?, Have you heavy draperies in your apartment?, Have you a safe in your apartment?*[11] Keeler explained that innocent persons showed no difference in reactions to the relevant questions compared to the other questions, while the guilty person reacted stronger to the questions about the burglarized apartment.

Summers (1939) described the use a series of *significant questions*[12] and provided these examples from a mock crime laboratory experiment: *Do you know who took the money?, Did you take the money?, Have you the money on your person?* These questions were included in sequence with these matter of fact questions:[13] *Are you wearing a black coat?, Did you eat breakfast this morning?* Summers also described the use of relevant questions regarding whether an examinee had helped to kill X, including: *Were you in the home of X on the day of the murder?, Did you kill X?, and Do you know who killed X?* Importantly, Summers included questions called *emotional standards,*[14] for which he gave the following examples: *Were you ever arrested, Are you living with your wife?, "Do you own a revolver?* [15] Emotional standard questions preceded each of the significant questions in the test question sequence, and the sequence of questions was presented three times.[16] Summers explained that if an examinee showed generally greater reactions to the significant questions than to the emotional standards, the examinee was attempting to deceive the examiner; if the examinee showed reactions to the significant questions that were generally not greater than reactions to the emotional standards, the examinee was answering truthfully. This

[8] In the combinatoric approach to the hypothesis of role discrimination, a test result is correct if the results are correct for all individual questions.

[9] Keeler's technique, though he does not appear to have used the term at the time, forms the basis of what is referred to today as the relevant-irrelevant technique (Department of Defense, 2006).

[10] Referred to today as neutral questions.

[11] Although Keeler's method of target selection is unlikely to be used today, the use of seemingly neutral relevant questions is noteworthy as an innovative early solution to challenges of test data analysis.

[12] Referred to today as relevant questions. Use of the term significant generally implies statistical significance.

[13] Similar to the irrelevant questions described by Keeler and neutral questions of today.

is one of the earliest examples of the current conception of the polygraph technique that strength of reactions to one group of questions or the other (relevant or comparison) is a function of deception or truth-telling to the relevant questions (Kircher & Raskin, 1988; Nelson, 2015a; Nelson, 2015b).

Reid (1947) introduced a revised questioning technique that included relevant questions, irrelevant questions, and two types of comparison questions.[17] He suggested the use of the following irrelevant questions: *Have you ever been called 'Red'? (the examinee had used one or more alias names including the name Red). Did you have something to eat today?, and Did you ever smoke?* (the examiner had seen the examinee smoking). Affirmative truthful answers were solicited from the examinee regarding these irrelevant questions. Reid provided these examples of comparative response questions to which the examiner was reasonably sure the examinee would lie: *Have you ever been arrested before?, Since you got out of the penitentiary, have you committed any burglaries?, Have you ever lied on the witness stand?, Have you ever stolen anything?, Have you ever cheated on your income tax returns?, Have you ever committed adultery?*, and *Beside that five dollars you told me about, have you ever stolen any money? Reid stated, The examiner must feel reasonably sure, as a result of his preliminary interrogation, that the subject will answer "No" to any of the above suggested questions used for comparative response purposes.* Reid gave the following as examples of relevant questions: *Do you know who shot John Jones?, About two months ago did you kill a man during a burglary at 112 State Street?, Did you stay in Chicago last night?, Do you know who shot John Jones?, Did you kill John Jones last Saturday night?, Did you steal a diamond ring from John Jones' room last Saturday night?, and Were you present when John Jones was shot?* These questions represent a retention of earlier attempts to stimulate the guilty examine with a variety of target questions that describe distinct aspects of a known or alleged incident under investigation.[18] However, field practice with the Reid technique involved one decision about the test as a whole, and published studies on this technique involved uniform decisions at the level of the question and therefore did not attempt to interpret the individual questions with an assumption that they may vary independently.

Kubis (1962) reported on a series of three experiments on various aspects of lie detection, including whether response variance could be scored and interpreted independently for individual test questions. One experiment suggested the potential for role discrimination in which blind scorers were able to differentiate between thief, lookout, and innocent suspects in a simulated theft. Relevant questions were: *Were you an accomplice to the thief?, Did you take the money from the coin box?, Do you have the coin box money with you?* A separate series asked these questions: *Do you know how much*

[14]Later referred to as emotional controls, then control questions and more recently comparison questions.

[15]These questions are unlikely to be viewed today as satisfactory comparison questions.

[16]Summers' question formulation method, and the technique as a whole - with three relevant target questions, three comparison questions, and three iterations of the test question sequence - bears early resemblance to contemporary versions of the family of zone comparison techniques.

[17]One type of comparison question was a guilt complex question based on a fictitious crime of the same type being investigated, and suggested that any reaction to the fictitious crime that is greater than or about the same as the actual crime question would indicate the examinee is innocent. Reid further suggested that reaction to the crime questions without reaction to the fictitious crime would indicate guilty knowledge or responsibility rather than nervousness or other factors. Raskin, Barland, and Podlesny (1978) reported that guilt complex questions were less effective than comparison questions at determining truthfulness or deception. These questions are not used in current PDD techniques.

[18]Backster (1963) recognized that the use of several behaviorally distinct questions introduces complex and troublesome implications regarding attention, salience, interpretation of test data and test accuracy. Backster began to advocate for the use of a series of single-issue test questions that forgo attempts to differentiate role involvement and instead describe the most important aspect of a known or alleged event. Barland Honts and Barger (1989) later showed that a series of single issue exams offered no advantages – or may suffer from the same psychological and statistical multiplicity complications – compared to the use of multiple issue exams.

money was in the coin box?, Did you act as a lookout for the thief?, and Do you have the coin box money with you? However, blind scorers in two later experiments, a crime scenario and a classified-information scenario, were informed that the examinees were deceptive to two of four target questions, but were unable to discriminate deceptive from truthful questions.[19]

The Department of Defense (2006a; 2006b) described the use of a variety of types of relevant test questions in the context of the Comparison Question Test (CQT) format, a generic category of the test that encompasses variations of the Modified General Question Technique (MGQT), including versions developed by the United States Army and United States Air Force (US Customs and Border Protection, 2010). Stimulus questions in this format are formulated to serve different purposes. These include both primary relevant questions that test possible direct involvement and secondary relevant questions that test possible involvement in a crime such as helping, planning, or any participation. Secondary relevant questions can also describe issues of guilty knowledge such as seeing, hearing, or knowing specific crime details. Consistent with the trend toward quantitative decision models, these examination formats employ a structured numerical scoring method with standardized protocols for feature identification and numerical transformations. Interpretation of CQT/MGQT test question stimuli is accomplished through structured decision rules that assume reactions vary independently for each relevant question. The test results are reported deceptive if reactions to any relevant question are determined to be indicative of deception, and are reported as truthful when reactions to all question are determined to be indicative of truth-telling.

Method of Literature Review and Evaluation

Ten published studies were found in the literature to provide quantitative information regarding effect sizes for decisions using both grand total and subtotal scores. One of these studies described the result of two experiments using two different samples, and the sample data from one study was used in another larger analysis. In all, eleven separate experiments were identified as providing results using both grand total and subtotal scores. Included studies are listed below:

1 Horvath and Reid (1971)

2. Hunter and Ash (1973)

3. Slowik and Buckley (1975)

4. Wicklander and Hunter (1975)

5. Raskin, Kircher, Honts, and Horoitz (1988)

6. Barland, Honts, and Barger (1989)

7. Podlesny and Truslow (1993)

8. Krapohl and Norris (2000)

9. Senter (2003)

10.Senter and Dollins (2003)

Two-way unbalanced ANOVAs were calculated using the method described by Cohen (2002) to test the level of significance of differences among results using grand total and subtotal scores for criterion guilty and criterion innocent groups, including decision accuracy, sensitivity, specificity, false-positive and false-negative errors, and inconclusive results. Unbalanced ANOVAs were used due to difference in study sample sizes. Standard

[19]Kubis (1962) is also interesting for a number of other practical and historical reasons, including the use discriminate analysis to show that the structural correlation of electrodermal responses was greater than respiratory and cardiovascular responses. Kubis also provided one of the earliest published references to numerical scoring, including the use of a Likert type numerical transformation to assign integer scores 0, 1, 2, and 3 to relevant and comparison stimuli using a rubric of "non-significant," "doubtfully significant," "significant," and "very significant."

deviation estimates were calculated using binomial approximation to the normal distribution, using the arithmetic mean of the sample sizes.[20] One-way, post-hoc ANOVAs were used as needed to investigate significant interactions. Random chance probability of a correct or incorrect classification was assumed to be .5 for decisions at the levels of the test as a whole and the individual questions. Statistical tests were conducted with statistical significance set at $\alpha = .05$.

Chronology of research pertaining to the independence hypothesis

Horvath and Reid (1971) reported the criterion accuracy of blind evaluation of confirmed field examinations conducted by Horvath using the Reid Technique (Reid, 1947; Reid & Inbau, 1977).[21] This study involved 10 examiners who provided blind subjective judgments of a sample of confirmed criminal investigation exams after removing exams that were regarded as very easy or very difficult to interpret. Half of the examinees (n = 20) were reportedly confirmed guilty[22] and the other half (n = 20) were reportedly confirmed innocent.[23] Unweighted decision accuracy was reported as 87.5% for the test as a whole. In addition, blind evaluators provided categorical judgments for 164 individual questions, for with the criterion state was coded uniformly for each case. The external criterion was coded innocent for half of the questions and the other half was coded as guilty. Decision accuracy for individual test questions was reported as 88.2% .[24]

Hunter and Ash (1973) reported the results of a study in which seven examiners from the staff of John E. Reid and Associates (Chicago, Illinois) provided blind judgments of deceptive or truthful, regarding a sample of N= 20 confirmed field cases that were conducted using the Reid Technique. Field cases included: theft, official misconducted, sexual assault, and homicide. Half of the sample cases were reportedly verified innocent, and the other half were reportedly verified guilty. Judgments were made by evaluating whether the consistency of physiological response was greater to the relevant or comparison questions. Unweighted decision accuracy was reported to be 87.1% for judgments made at the level of the test as a whole. Decision accuracy was reported to have been 92.1% for judgments regarding the individual relevant questions with criterion states coded uniformly for all questions within each case.[25]

Slowik and Buckley (1975) reported the results of a study in which seven examiners from the staff of John E. Reid and Associates provided blind categorical judgments of deception and truth-telling regarding a sample of

[20]Mean sample size was used to calculate the binomial approximation of the sample standard deviation for both grand total and subtotal scores. This method was preferred because the assumptions of frequentist hypothesis testing assert that sampling error is a function of sample size when the sample is randomly selected from independent members of the population. Because subtotal scores within each examination violate the assumption of independence, attempts to use the N of subtotals were expected to overestimate the precision of the sampling statistic.

[21]This technique is not presently being taught at any accredited polygraph school that we are aware of, but forms part of the basis of the family of MGQT formats.

[22]The terms guilty and innocent are used to refer to the criterion state to avoid potential confusion when using the terms deceptive and truthful to discuss categorical test results. These terms are not intended to convey assumptions or implications about legal or criminal culpability in either laboratory or field settings.

[23]Among the important differences between this technique and contemporary polygraph techniques is that the Reid technique employed a clinical approach in which judgments were made via consideration of a combination of whether physiological responses to test stimuli were interpreted as loaded onto either relevant or comparison stimuli (without fixed numerical cutscores), together with evaluation of the case background information and information from behavioral observation of the examinee during testing.

[24]Uniform coding of the criterion state in this manner, together with interpretation at the level of the individual question, is not an expression of the MFH or the assumption that the both the criterion state and responses of individual questions will vary independently, and cannot address the hypothesis of role discrimination among the guilty cases.

[25]Uniform coding does not address MFH.

N = 30 field cases that were selected from the reportedly verified case files at John E. Reid and Associates. Half of the sample cases were reported as confirmed as innocent, and half were reported as confirmed guilty. Cases involved theft, industrial sabotage, drug abuse, and sexual assault. The sample consisted of 141 individual relevant questions for which 71 questions were coded as confirmed innocent, and 70 were coded as confirmed guilty. Decision accuracy was reported as 86.7% for judgments made at the level of the individual test questions, with criterion states coded uniformly for all questions within each case.[26] Accuracy was reported to have been 88.9% for judgments made at the level of the overall test.

Wicklander and Hunter (1975) reported the results of a study in which six examiners from the staff at John E. Reid and Associates provided categorical judgments regarding a sample of 20 reportedly confirmed field cases that were conducted by the authors. Half of the sample cases were reported as confirmed guilty, and half were reported as confirmed innocent. Sample examinations consisted of 89 individual relevant questions, for which 43 were reported as verified innocent and 46 were reported as verified guilty. Criterion accuracy for categorical judgments of individual relevant questions was reported to have been 92.8% when the criterion state for individual questions was coded uniformly for each case. [27] Criterion accuracy for judgments made at the level of the overall test was reported as 93.9%.

Raskin et al. (1988) analyzed 76 examinations that had been conducted by the United States Secret Service during FY83, FY84, and FY85. The criterion states were confirmed (through an exhaustive records search, described in the study) by a combination of admissions and reliable evidence independent of the polygraph examinations. The examinations were independently scored by six experienced examiners employed by the United States Secret Service and one psychophysiologist-examiner. The cases were divided into Pure Verification and Mixed Verification cases of which 37 Pure Verification examinees were confirmed deceptive to all questions and 26 were confirmed truthful to all questions. The Mixed Verification cases consisted of 13 cases for which the examinees were confirmed as deceptive to at least one question and also confirmed as truthful to at least one question. There were 23 questions for which the examinees were confirmed deceptive and 19 questions for which the examinee were confirmed truthful. Accuracy of blind numerical scores for Pure Verification cases in which questions were not independent was 90.1%. Unweighted average accuracy for the Mixed Verification cases was 75.6%[28] for decisions at the level of the individual test questions. Raskin et al. concluded that test accuracy is maximized by formulating test questions that can be interpreted as non-independent.

Barland et al., (1989) investigated the ability to detect deception and truth-telling at the level of the individual questions. Federal polygraph instructors tested 100 participants who completed none, one, two, or three mock espionage and sabotage behaviors. Participants were randomly assigned to guilty or innocent groups. Fifty examinees were tested using a multiple-issue examination consisting of criterion independent target questions describing possible involvement in different mock espionage and sabotage activities. The other 50 examinees were tested with a series of three single-issue exams, each of which described

[26]Uniform coding does not address the potential for role discrimination among guilty cases.

[27]Uniform coding does not address MFH and cannot be used to investigate the potential for role discrimination among guilty cases.

[28]Also noteworthy in this study was that there was no effect for experience or type of training, with both highest and lowest criterion accuracy rates resulting from blind scores provided by experienced field examiners. The main effect for decisions was significant ($F [1, 212] = 1340.26$, $p < .001$), but the main effect for confirmation was not significant ($F [1, 212] = 1.57$, ns) suggesting that decisions did not vary as a function of confirmation. The authors also reported that the computer algorithm model generally outperformed blind numerical scores and most human scorers. They recommended the use of algorithms as a form of quality control.

a single crime with two relevant questions in a sequence repeated three times. Unweighted accuracy was 87.3% for the participants tested using the series of single-issue exams and 83.9% for those participants who were evaluated using a multiple-issue examination format.[29] The difference in accuracy was not significant for tests using target questions that varied independently vs. decisions made using a series of examinations for which decisions were made at the level of the test as a whole. However, unweighted accuracy was only 47.8% when decisions were made at the level of the individual target questions, and only 33% of the results on individual crimes were correct when attempting to interpret the results of individual target questions. Although overall accuracy was high for guilty subjects, they were unable to determine which behavior had been committed by the guilty examinees.

Podlesny and Truslow (1993) reported the results of a laboratory study designed to investigate the ability of the polygraph test to identify guilty and innocent examinees and to determine the role or level of involvement of each examinee . Participants were recruited from the local community through a temporary employment agency and were paid for their participation, and were reported as not suffering from lack of sleep or illness. Ninety-six participants were randomly assigned equally to *innocent, perpetrator, accomplice,* or *confidant* and were tested using a Modified General Question Technique. Relevant questions included items about direct involvement, secondary involvement, knowledge, and innocence or truth-telling in general. Data were evaluated manually by 10 polygraph instructors from the Department of Defense.[31] *Innocent* participants produced a significantly positive mean grand total and significantly positive mean subtotal scores to all relevant questions except knowledge. Each of the guilty groups differed significantly from the innocent group, but there was no significant difference among the guilty groups. Guilty *perpetrators* produced significantly negative mean scores to questions about perpetration, general truth, and knowledge, *but not to participation/involvement.* Guilty *accomplices* also had significant negative mean scores for questions regarding general truth, perpetration, and knowledge, *but not participation/involvement.* Guilty *confidants* produced significant negative scores to knowledge questions but failed to produce significant negative scores to questions about general truth. Guilty *confidants* also produced a higher rate of false negative errors and failed to produce significant negative mean total scores

Podlesny and Truslow reported that questions about *participation/involvement* did not discriminate among any of the guilty roles.[32] Unweighted accuracy was 89.7% using the grand total scores and 79.1% for subtotal (i.e., question) scores. The authors concluded that the evidence provided general support for the hypothesis that guilty participants would produce responses to individual relevant questions that varied significantly from responses to comparison stimuli, but did not generally support the hypothesis of role discrimination. The authors cautioned that although some discrimination among guilty groups might be possible, attempts to sub-categorize deception or to differentiate perpetration from participation/involvement and guilty knowledge may place overambitious demands on the testing context and examinee.

[29]Barland, Honts and Barger (1989) also reported the results of a two-way ANOVA showing a significant main effect for criterion state ($F [1, 96] = 30.4$, $p < .001$) but no significant interaction between criterion state and type of testing (i.e., multiple issue exam or series of single issue exams) for decisions made at the level of the test as a whole.

[30]Podlesny and Truslow (1993), in this study, also replicated the feature extraction and structural model development of Kircher and Raskin (1988) and Raskin, Kircher, Honts and Horowitz (1988), and provided replication and extension of the significant loading effect and numerical score differences for members guilty and innocent groups in a CQT test paradigm.

[31]Data were also scored using a computer algorithm similar to the one described by Kircher and Raskin (1988), and hypothesis testing was completed using linear discriminate analysis.

[32]Podlesny and Truslow (1993) also reported the results of a multivariate analysis that showed a significant interaction for role x relevant question ($F [9, 219)$, $= 1.57$, $p < .05$), and a significant main effect for group ($F [3, 92] = 30.2$, $p < .01$).

Krapohl and Norris (2000) reported the results with 16 reportedly confirmed innocent and 16 reportedly confirmed guilty criminal investigation tests that were conducted using the Army MGQT format (Department of Defense, 2006b).[33] The series of test questions for the Army MGQT exam is structurally similar to that of the Reid Technique.[34] These examinations consisted of several different types of relevant questions including: 1) primary relevant questions that address direct involvement in the crime or issue under investigation, 2) secondary relevant questions that describe helping, planning or participating indirectly in the crime or issue, or secondary involvement such as seeing hearing, or knowing details about the crime or issue, 3) evidence-connecting relevant questions designed to determine if the examinee was involved with any of the evidence of the crime, or is aware of the nature or location of various items of evidence, and 4) guilty-knowledge relevant questions that are used to determine if the examinee has any knowledge of who committed the incident under investigation (Department of Defense, 2006a). Data were scored by three experienced examiners using a seven-position scoring model. Decisions made at the level of the test as a whole resulted in an accuracy rate of 75.5% but only 56.9% when decisions were made at the level of the individual questions.

Senter (2003) published the results of a study of decision rules with the Army MGQT format (Department of Defense, 2006b). Data were 205 verified criminal investigation tests and included 161 reportedly confirmed guilty cases and 44 reportedly confirmed innocent cases that were used in four previous research samples. Twenty-six of the exams were conducted by the United States Army Criminal Investigative Division (USACID). Forty-seven exams were conducted by the Bureau of Alcohol Tobacco and Firearms (BATF). Thirty-two exams were used in a sample used by Krapohl and Norris (2000). One-hundred MGQT cases were from a stratified matched sample used by Blackwell (1998). Cases from the BATF and USACID were scored by the original examiners. Results from the three examiners in the Krapohl and Norris (2000) study and the three examiners from the Blackwell (1998) study were averaged to achieve a single set of scores for each of those samples. There were 644 questions for which the criterion state was coded as guilty and 176 questions coded as innocent, with the criterion state coded uniformly for all questions within each case. Using traditional cutscores and decision rules, in which categorical results are determined at the level of the individual target questions, the unweighted mean for criterion accuracy of guilty and innocent groups was 66.1%. Unweighted accuracy using only the grand total scores was 83.8%.

Senter and Dollins (2003) studied the criterion accuracy of decisions based on grand total and subtotal scores using the Federal ZCT test format (Department of Defense, 2006a; 2006b) which makes use of primary relevant questions, that describe the possible direct involvement of an examinee in a known incident or allegation, along with secondary relevant questions that describe indirect involvement or guilty knowledge regarding an issue under investigation. Laboratory sample data were aggregated from previous studies, including data from Senter and Dollins (2004) [35] involving a sample of 50 criterion guilty cases and 50 criterion innocent cases that were scored by five examiners, along with data from two

[33]This technique is no longer used in field settings, and was not included in the American Polygraph Association (2011) meta-analytic survey of criterion accuracy. Results are included here because the study design addressed the issues surrounding assumptions about independent or non-independent variance, and decision rules.

[34]One important difference between the Reid format and the Army MGQT is that decisions are made by comparison of subtotal scores for individual target stimuli to fixed cut-scores when using the Army MGQT. This is in contrast to clinical judgments made at the level of the test as a whole when using the Reid technique. Both formats may use a combination of target questions developed around both primary and secondary involvement in addition to evidence-connection or guilty-knowledge questions.

[35]Note that the publication dates are out of sequence due to apparent differences in project and publication timelines.

experiments described by the Department of Defense Polygraph Institute Research Division Staff (2001) involving a sample of 32 Federal ZCT exams (16 guilty cases and 16 innocent cases) that were scored by three examiners, and another sample of 32 Federal ZCT exams (16 guilty cases 16 innocent cases) that were scored by seven examiners. Laboratory studies included a total of 82 guilty examinees and 82 innocent examinees, whose data were scored by 15 different examiners. Data consisted of 492 individual questions, including 246 questions for which the criterion state was guilty and 246 questions for which the criterion was innocent. Decisions made with the grand total resulted in an unweighted average accuracy rate of 91.8%. Unweighted accuracy was 86.8% for decisions based subtotal scores, with the criterion state coded uniformly for all questions within each case.

Senter and Dollins (2003) also studied decisions based on grand total and subtotal scores using data from two Department of Defense archival samples of reportedly confirmed field exams that were conducted using the Federal ZCT format. Field cases consisted of examinations scored by six different examiners, with 115 guilty examinees and 85 innocent examinees. One hundred of the Federal ZCT cases from the sample were previously used by Blackwell (1998) and included 65 criterion deceptive cases and 35 criterion innocent cases that were scored by three examiners. An additional 100 cases were from the sample used by Krapohl, Dutton and Ryan (2001). There were a total of 600 individual questions, including 345 questions for which the criterion state was guilty and 255 questions for which the criterion state was innocent. Unweighted accuracy for the Federal ZCT exams was 91.5% for decisions based on grand total scores, but unweighted accuracy was only 79.8% for decisions based on subtotal scores.[36]

Results

Discrimination of deception and truth-telling.

Data from the 11 studies was scored by 85 examiners who evaluated N = 888 individual cases consisting of n = 549 criterion guilty cases and n = 339 criterion innocent cases. There were a total of N = 2870 individual target questions, including n = 1748 criterion guilty questions and n = 1122 criterion innocent questions.[37] Sensitivity and specificity rates for these studies are shown in Appendix A, while false-positive and false-negative error rates are shown in Appendix B. Information from the sensitivity, specificity, and error rates, together with the sample size, provide the information for this analysis.

The weighted mean of unweighted accuracies for grand total scores was 88.9% (SEM = .011, 95% CI = .869 to .910) and the weighted mean of unweighted accuracy rates for decisions based on subtotal scores was 79.0% (SEM = .014, 95% CI = .763 to .816).[38] Decisions based on an assumption of non-independent response variance, using the grand total score, were significantly more accurate than decisions based on an assumption that physiological responses to individual questions, and resulting subtotal scores, will vary independently, $F(1,1775) = 16.737$, ($p < .001$).

Table 1 shows the weighted means, standard errors, and confidence intervals for guilty and innocent cases. Figure 1 shows the mean plot for guilty and innocent cases. Results from a two-way unbalanced[39] ANOVA showed a significant interaction between criterion state and decision rule, $F(1,1772) = 40990.342$, ($p < .001$). Post-hoc one-way unbalanced ANOVAs showed a significant one way effect for criterion state for decisions based on grand

[36]Senter and Dollins (2003) also reported results using a sequential combination of the grand total score and subtotal scores and that would provide similarly high decision accuracy (87.7% for laboratory cases and 86.4% for field cases) with increased test sensitivity and reduced inconclusive results.

[37]Data from Krapohl and Norris (2000) are excluded from these totals because the sample was included in the data reported by Senter (2003).

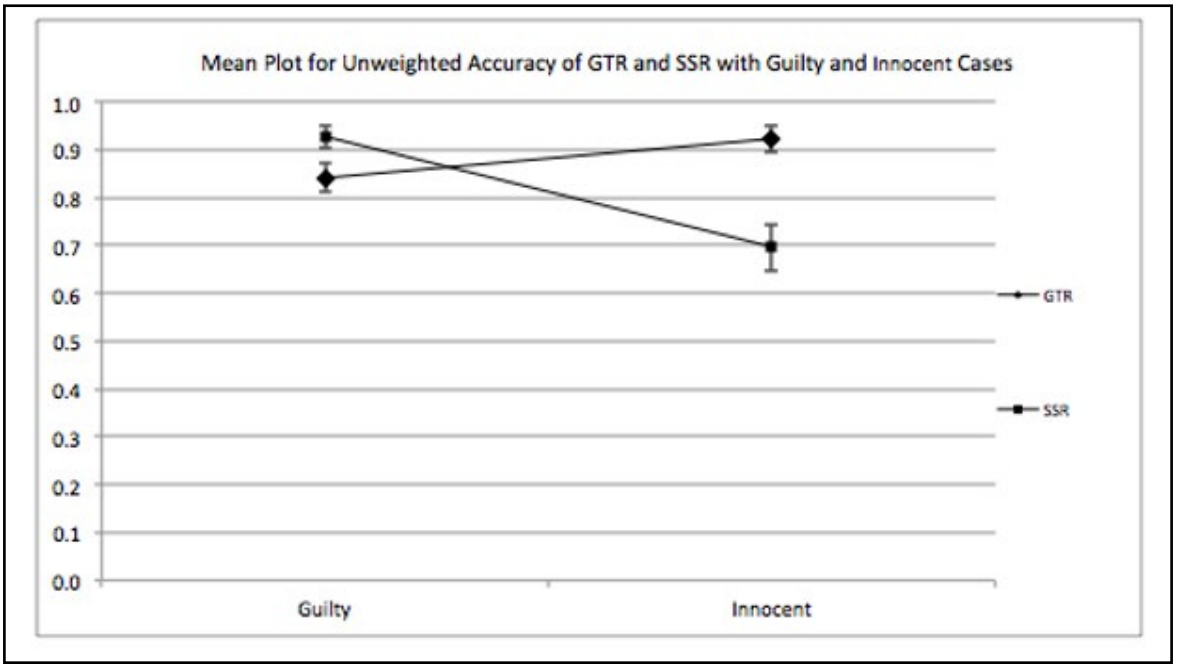
[38]The weighted mean is preferred in this situation, giving more importance to the statistical estimates from larger studies, due to the premise of frequentist inference that the precision and error with which a sampling statistic describes the population is a function of the sample size, with the assumption that samples are representative of the population if the sample cases are independent of each other and are randomly selected (i.e., all members of the population have an equal chance of being selected).

total scores, F (1,837) = 7.032, (p = .008) and for subtotal scores, F (1,837) = 39.724, (p < .001). The difference in accuracy with guilty and innocent cases was significant for decisions based on grand total scores, F (1,1097) = 9.923, (p = .002) and subtotal scores, F (1,677) = 31.154, (p < .001). Decisions based on grand total scores were more accurate with innocent cases, while decisions using subtotal scores were more accurate with guilty cases. However, unweighted decision accuracy was significantly greater than chance for both grand total and subtotal scores with both guilty and innocent cases.

Table 1. Mean accuracy for guilty and innocent cases using grand total and subtotal scores.

		Criterion guilty	Criterion innocent
Results based on grand total scores	Weighted mean	.841	.924
	SEM	.016	.014
	95% confidence interval	.811 to .872	.895 to .952
		Criterion guilty	Criterion innocent
Results based on subtotal scores	Weighted mean	.926	.694
	Standard deviation	.011	.025
	95% confidence interval	.904 to .948	.647 to .745

Figure 1. Mean plot for accuracy of guilty and innocent cases based on grand total and subtotal scores.



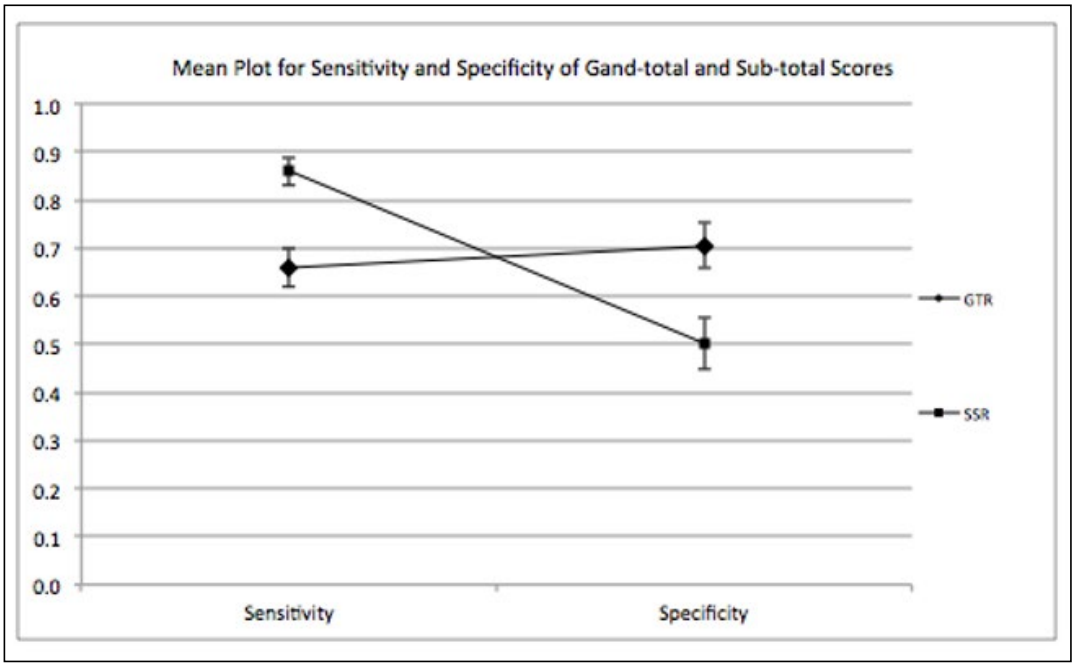
[39]Use of unbalanced ANOVAs, using the harmonic mean of sample sizes, was required due to inequality in the cell sizes. Unbalanced ANOVAs provide a slight reduction in statistical power, and are thought to be a more cautious test of significance under these circumstances.

Table 2 shows the sensitivity and specificity rates of decisions based on grand total and subtotal scores. Figure 2 shows the mean plot for test sensitivity and specificity. Results from a two-way unbalanced ANOVA showed a significant interaction between decisions and decision rule, F (1,1772) = 41642.631, (p < .001). Post-hoc one-way ANOVAs showed that the difference in test sensitivity and specificity was not significant for decisions based on grand total scores, F (1,837) = 1.05, (p = 0.306), but was statistically significant for results based on subtotal scores, F (1,837) = 72.713, (p < .001). The difference in test sensitivity to deception for grand total and subtotal scores was statistically significant, F (1,1097) = 32.047, (p < .001), and the difference in test specificity for grand total and subtotal scores was also significant, F (1,677) = 15.469, (p < .001). Use of subtotal scores was more accurate with guilty cases, while grand total scores were more accurate with innocent cases.

Table 2. Sensitivity and specificity cases using grand total and subtotal scores.

		Sensitivity	Specificity
Results based on grand total scores	Weighted mean	.660	.706
	Standard deviation	.020	.025
	95% confidence interval	.620 to .699	.658 to .755
		Sensitivity	Specificity
Results based on subtotal scores	Weighted mean	.834	.465
	Standard deviation	.015	.027
	95% confidence interval	.831 to .889	.449 to .555

Figure 2. Mean plot for sensitivity and specificity.

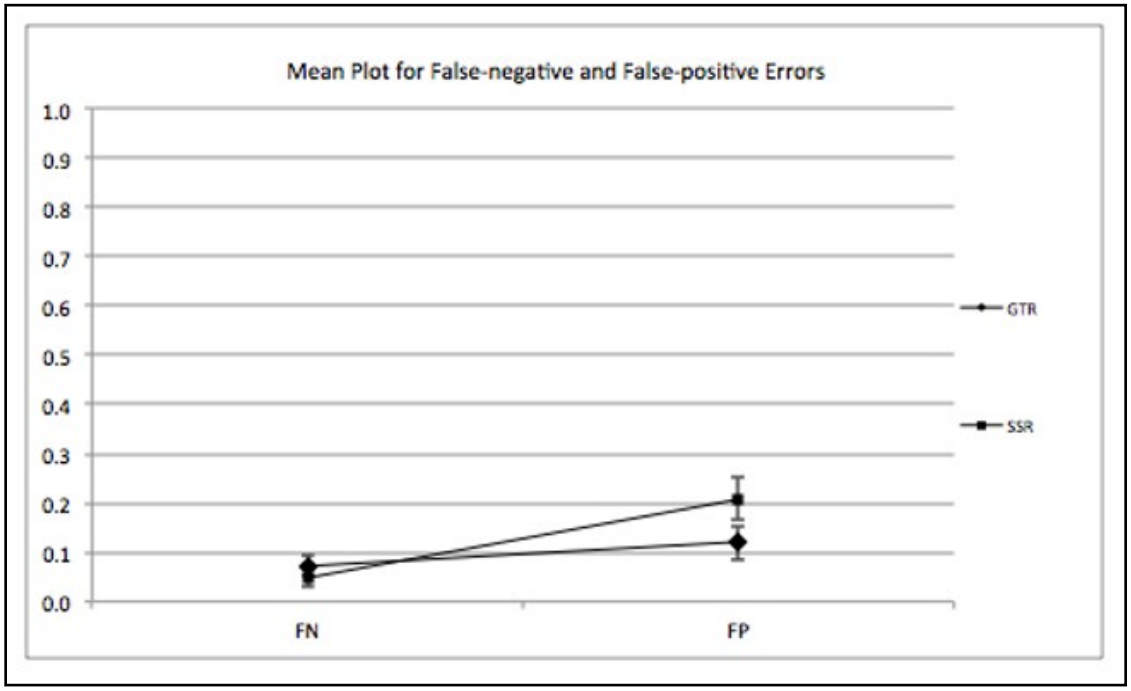


False-negative and false-positive errors are shown in Table 3, and Figure 3. A two-way unbalanced ANOVA showed a statistically significant interaction between errors and decision rule, F (1,1772) = 7979.41, (p < .001). Post-hoc analysis showed that false-negative and false-positive errors did not differ significantly at the .05 level for results using grand total scores, F (1,837) = 2.709, (p = .100). However, the difference in error rates was significant for results based on subtotal scores, F (1,837) = 25.567, (p < .001). False-negative error rates did not differ significantly for results based on grand total and subtotal scores, F (1,1097) = 1.374, (p = 0.241), but the difference in false-positive errors was significant for decisions based on grand total and subtotal scores, F (1,677) = 5.034, (p = 0.025). An assumption of independent variance contributed to a statistically significant increase in FP errors.

Table 3. False-negative and False-positive errors.

		False-negative errors	False-positive errors
Results based on grand total scores	Weighted mean	.072	.119
	Standard deviation	.011	.018
	90% confidence interval	.051 to .094	.085 to .154
		False-negative errors	False-positive errors
Results based on subtotal scores	Weighted mean	.048	.209
	Standard deviation	.009	.022
	90% confidence interval	.031 to .066	.166 to .252

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

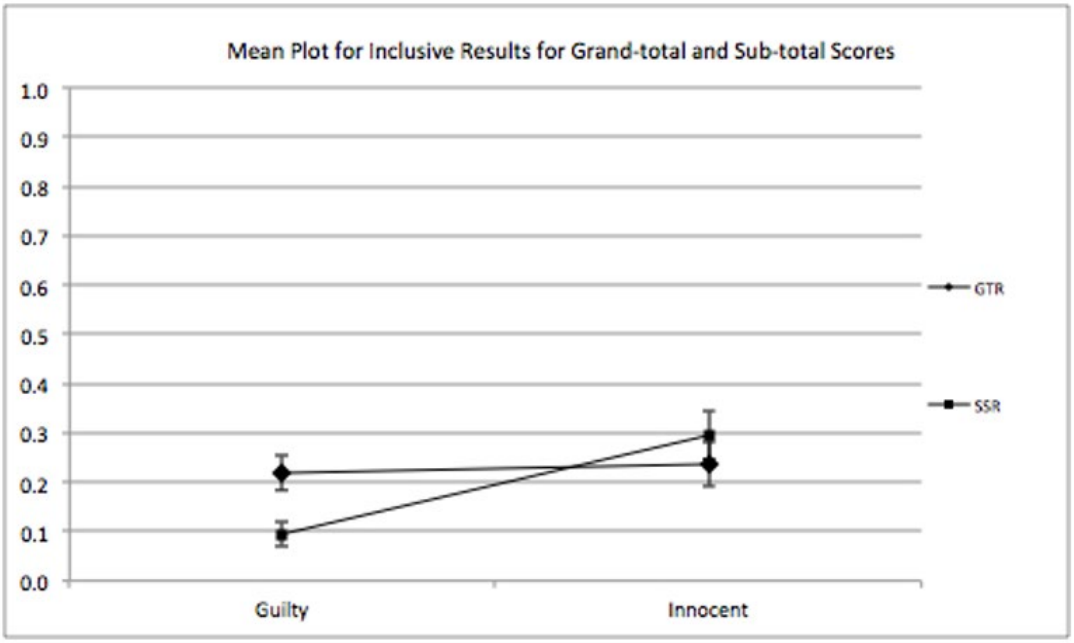


Inconclusive rates for results based on grand total and subtotal scores are shown in Table 4 and Figure 4. A two-way unbalanced ANOVA showed a significant interaction between the criterion state and decision rule for inconclusive results, F (1,1772) = 11102.575, (p < .001). Post-hoc one-way unbalanced ANOVAS showed that the difference in inconclusive rates for guilty and innocent cases was not significant for results based on grand total scores, F (1,837) = 0.218, (p = .641), but was statistically significant when results were based on subtotal scores, F (1,837) = 29.029, (p < .001). Inconclusive rates the innocent cases did not differ significantly for results based on grand total and subtotal scores, F (1,677) = 1.518, (p = .218). However, the difference in inconclusive rates was statistically significant for guilty cases when comparing results using grand total and subtotal scores, F (1,1097) = 16.39, (p < .001). Subtotal scores produced fewer inconclusive results with guilty cases.

Table 4. Inconclusive rates for results based on grand total and subtotal scores.

		Criterion guilty	Criterion truthful
Results based on grand total scores	Weighted mean	.218	.237
	Standard deviation	.018	.023
	90% confidence interval	.184 to .253	.192 to .283
		Criterion guilty	Criterion truthful
Results based on subtotal scores	Weighted mean	.120	.323
	Standard deviation	.024	.044
	90% confidence interval	.070 to .119	.248 to .345

Figure 4. Inconclusive rates for results based on grand total and subtotal scores.



Unweighted decision accuracy shown in Table 1 was significantly greater than chance with both guilty and innocent cases for grand total scores and for subtotal scores. For decisions using grand total scores, both test sensitivity to deception and test specificity to truth-telling were significantly greater than chance (0.5), with no significant difference in sensitivity and specificity. However, decisions using subtotal scores showed an increased in test sensitivity along with a reduction of mean test specificity to chance.

False-positive and false-negative error rates did not differ significantly for results based on grand total scores, and the mean false-negative error rate was not significantly different for results based on subtotal scores when compared with results of grand total scores. However, decisions based on subtotal scores produced significantly more false-positive errors than decisions using grand total scores. Inconclusive results did not differ significantly for guilty and innocent cases when results were based on grand total scores but were significantly lower for guilty cases when results were based on subtotal scores.

Results of these analyses provide general support for the hypothesis that PDD examination data can discriminate deception and truth-telling at rates significantly greater than chance for the test as a whole and at the level of individual question. The results do not support the validity of the MFH that interpretation of independent response variance for individual target questions optimizes or increases the effectiveness of event-specific polygraph examinations. Results using grand-total scores, based on an assumption of non-independent response variance produced greater mean accuracy and more balanced error rates compared to results using subtotal scores. However, observed differences were not uniform for the criterion guilty and criterion innocent cases. Based on an assumption of independent response variance, subtotal scores showed increased effectiveness with guilty cases but also showed a disproportionately larger decrease in effectiveness with innocent cases.

Discrimination of behavioral role or level of involvement.

Analysis of effect sizes for role discrimination or level of involvement of guilty persons requires that the guilt vs. innocence criterion state is known or set independently for each individual relevant question, in addition to a requirement that decisions are made using subtotal scores. Although results were reported for the study samples using both grand total and subtotal scores, a majority of these samples consisted of target questions for which the criterion state was coded uniformly (i.e., non-independently) for all questions within each case. In other words, the criterion state was coded as guilty to all questions or innocent to all questions within these sample cases. Uniform coding in this manner makes no assumption of independent criterion variance, and this imposes a substantial barrier to the interpretation of independent responses and the effect of the MFH for these sample cases.

Samples from three of the included studies (Barland et al., 1989; Podlesny & Truslow, 1993; Raskin et al., 1988) did include target questions for which the criterion states of the individual target questions were coded independently. An ancillary analysis was completed using data from these three studies in comparison to data from the seven studies (Horvath & Reid, 1971; Hunter & Ash, 1973; Krapohl & Norris, 2000; Senter, 2003; Senter & Dollins, 2003; Slowik & Buckley, 1975; Wicklander & Hunter 1975) for which the criterion states of the individual questions was coded uniformly.

Mean unweighted accuracy of studies for which the criterion state of the individual target questions was coded independently was .892, SEM = .021 (95% CI = .851 to .934) for decisions using grand total scores, and .792 SEM = .028 (95% CI = .737 to .847) for decisions using subtotal scores. Mean accuracy of study samples for which the criterion state of individual target questions was coded uniformly for within each case was .889, SEM .012 (95% CI = .865 to .912) for decisions using grand totals, and .789, SEM = .016 (95% CI = .758 to .820) for decision using subtotal scores. Differences were not statistically significant for grand total decisions of questions coded independently or uniformly, nor was the

difference statistically significant for subtotal decisions of questions coded independently or uniformly.

To test for differences in criterion accuracy for independent and uniform coding of guilt and innocence within each case, a series of two-way ANOVA contrasts was conducted for decision accuracy, sensitivity and specificity, false-negative and false-positive errors, and inconclusive results for guilty and innocent cases. Figure 5 shows the mean plot for unweighted decision accuracy, Figure 6 shows the mean plot for sensitivity and specificity, Figure 7 shows the mean plot for errors, and Figure 8 shows the mean plot for inconclusive results.

Figure 5.

Figure 6.

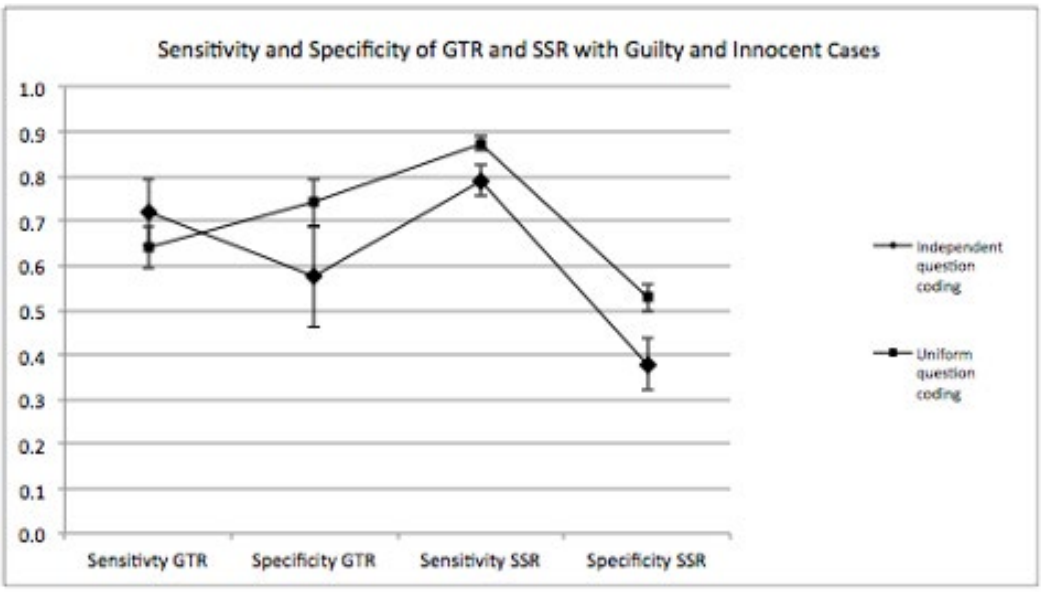


Figure 7.

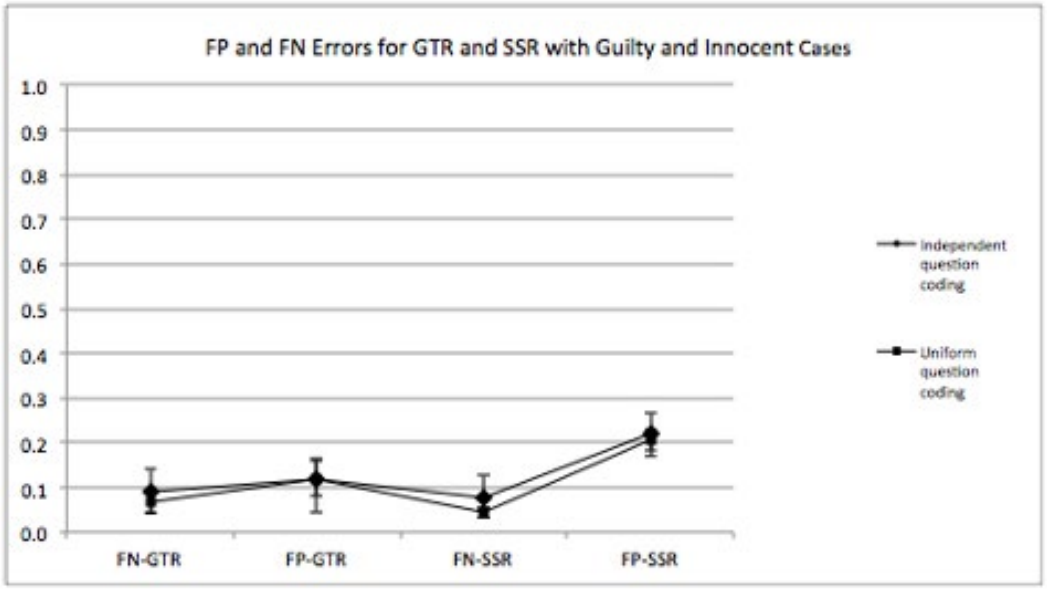
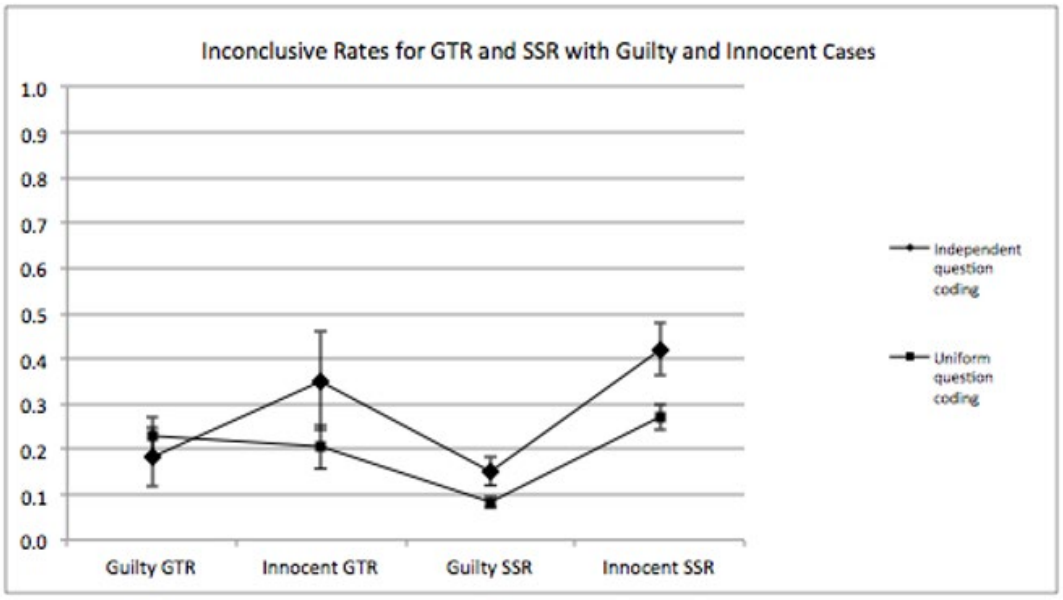


Figure 8.



The factor of interest for the planned contrasts was the method of coding of the criterion state for the individual target questions within each case, i.e., whether the criterion variance was independently coded or uniformly coded. The effect for differences in decision accuracy for independent versus uniform question coding, shown in Figure 5, was not significant, $F(1,1768) = 0.132$, ($p = .717$). Decision accuracy was similarly effective with cases for which the criterion state was coded independently for each question compared to cases coded uniformly. The one-way effect for test sensitivity and specificity, shown in Figure 6, was not significant at the .05 level, $F(1,1768) = 3.435$, ($p = .064$), The difference in errors, shown in Figure 7, was also not statistically significant for independent vs uniform coding of question criterion states, $F(1,1769) = 0.348$, ($p = .555$). Mean decision accuracy was lower for innocent cases when decisions were based on subtotal scores regardless of the method through which the criterion state was coded.

The one-way difference in inconclusive rates, shown in Figure 8, was statistically significant, $F(1,1768) = 4.077$, ($p = .044$) at the .05 level. Cases for which the criterion state of individual questions was coded independently produced more inconclusive results compared to cases for which questions were coded uniformly, and this difference was loaded on innocent cases. This is not surprising when considering that standard field practice when making decisions using subtotal scores is that the examinee is considered deceptive if the subtotal score for any individual relevant question equals or exceeds the normative cut-score, while truthful classifications are made when all subtotal scores equal or exceed the normative cut-score for a truthful result. Inconclusive results occur when neither of these two conditions is satisfied. A corollary to this procedure is that field polygraph practices prohibit examiners from rendering both deceptive and truthful classifications within a single exam. [Refer to Nelson, Blalock & Handler (2019) for a discussion of how test results are parsed for individual questions and for the test as a whole for single-issue and multiple-issue polygraphs.] As a consequence, when any single question within a test is classified as deceptive, results are deemed inconclusive for all other questions that are not statistically significant for deception. In other words, when the criterion state was coded independently for the relevant questions within each exam, and when the criterion states were mixed for the individual questions, with one or more questions coded guilty and one or more questions coded innocent, questions were classified as inconclusive if they were coded innocent in an exam for which any single questions was classified as deceptive.

As shown in Figure 5, decision accuracy was slightly higher for guilty cases when decisions were made using subtotal scores, but with a disproportionately larger decrease in accuracy with innocent cases. In addition, an interaction effect for decision accuracy can be observed in the results with grand total scores, shows in Figures 5, 6 and 8, between criterion state and whether the criterion states of individual questions were coded independently or uniformly when decisions were based on grand total scores. The increase in accuracy that resulted from the use of grand total scores with truthful cases was not observed when the criterion state was set independently for individual target questions but was observed when target questions were coded uniformly. When the criterion states of the individual questions were coded independently the use of grand total decisions resulted in less effective discrimination of deception and truth-telling compared to decisions using the grand total score with cases for which the criterion state of individual questions was coded uniformly.

Discussion

Important limitations of the present study include the form of analysis and the use of a series of unbalanced multivariate ANOVAs to evaluate several dimensions of test accuracy by using proportional statistics that describe categorical results. Calculation of variance estimates, statistical confidence intervals, and sums of squares was accomplished with binomial approximation of the normal distribution. A more precise analysis might be obtained through bootstrapping or Monte-Carlo simulation or other method. It is also possible that meta-analytic methods might lead to more interesting insights than were observed in this analysis.

Fundamentally, any analytic method that re-

lies on null-hypothesis significance testing is based on the assumptions that samples are collected in a random manner and are representative of the population. It is also assumed that sampling bias is a function of sample size, i.e., a sufficiently large sample will approximate the population with reasonable accuracy. However, large random samples are difficult to obtain. Furthermore, other sources of bias or error that may be introduced to the sample data and study results include sampling methodology, test administration (including test question construction and inter-rater reliability)and test data analysis. This analysis is premised on and assumes that the published sampling information is in some way representative of the larger population.

Some of the samples included field investigation cases for which case selection was contingent on non-random criteria involving the availability of confirmation data (i.e., confession or investigative evidence) that may not be independent of the PDD examination results. In other words, the PDD result may have contributed to the resolution of the field cases, and this can be expected to inflate the correlation between the test result and criterion state. A further limitation of field sampling procedures is the potential for the systematic exclusion of both false-positive and false-negative errors for which confirmatory confession or evidence may be difficult or impossible to obtain. Non-random sampling methodology may inflate observed test accuracy. Despite these known sampling confounds, field investigation samples offer the advantage of external and ecological validity.

Although ecological validity is not required to achieve the more important goal of external validity, laboratory samples have been subject to criticism because they may incompletely represent field testing contexts. Laboratory studies offer the important advantage of scientific control over the research questions, which cannot be achieved in field research. Results of laboratory studies have been shown to be more conservative but not significantly different from field studies (Honts, Thurber & Handler, 2020). Correspondence between the result of field and laboratory sampling data provides additional insight as to the stability or reproducibility of observed results.

An equally important limitation of the present analyses is that the results of the included studies were universally derived without regard for the normative distributions of the scores of guilty and innocent persons. This means that decision cut-scores were statistically under-informed and potentially sub-optimal as a result of unknown alpha levels for some of the sampling data. Also, statistical corrections were not applied to decision alpha nor and some of the included studies were completed without fixed numerical cutscores.

Individual field examinations represent a form of single-systems experiment, and examinations involving the simultaneous interpretation of multiple, potentially independent, sources of response variance represent the execution of multiple statistical comparisons within each test experiment. There are known statistical multiplicity effects that can occur when making multiple statistical comparisons, including the potential for inflation of alpha boundaries corresponding increase in Type I (i.e., false-positive) errors when making deceptive classifications, the potential for deflation of alpha boundaries, and corresponding increase in Type II (i.e., false-negative) errors when requiring multiple statistically significant truthful scores in order to make truthful classifications. For this reason, there is an unquantified possibility that sampling data from experiments with norm-referenced cut-scores and statistically informed alpha boundaries can produce different results.

Another limitation for this literature review is the small number of published studies that have evaluated the MFH. A number of the included studies made use of subtotal scores with uniform criterion coding that could not fully evaluate the MFH. Those that did attempt to code the criterion states of individual target issues point to a general conclusion against the notion of effective role discrimination, and there is a clear trend in the literature that scores based on uniform coding and grand total scores can maximize the criterion accuracy of PDD test data.

It was not possible to investigate whether the use of multi-facet questions contributes to any semantic increase in test sensitivity – which might result from describing a broader

range of crime-event related stimuli to which guilty persons may respond. Similarly, it was not possible to evaluate whether the staging and presentation of multi-facet test questions leads to increased effectiveness and resolution of post-test discussion and investigation activities or increased satisfaction among referring professionals. This review is limited to quantitative effects only. These hypothesized effects are beyond the scope of the present study but are nonetheless important and should be investigated in future studies.

Generalizability of these results to multiple-issue screening exams is neither completely certain nor completely unknown. Multiple-issue screening exams are conducted in a context where there is a strong assumption of independence of the criterion state for the individual questions. Greater independence of response variance in multi-issue testing may lead to differences in the ability of individual questions to discriminate deception and truth-telling. However, the degree to which the target questions of multiple-issue screening exams may include an increase in the independence of response variance is unknown. Practical experience in field polygraph settings has revealed potential problems with over-reliance on an assumption of independent response variance. As a result, field polygraph examiners are generally prohibited from making both deceptive and truthful classifications within a single exam. Also unknown are the effects of variable prior probabilities of guilt for multi-issue screening polygraphs compared to the priors associated with event-specific diagnostic exams. Few practicing field polygraph examiners are prepared to consider the effect that base-rates may have on Positive or Negative Predictive Values of the individual test question results.

Conclusion

The MFH holds that response variance is independent when the relevant questions employ different behavioral action verbs that will provide valid and reliable information about a person's behavioral role or level of involvement in a known allegation or incident. The MFH suggests that, in addition to the ability to partition response variance into the larger dimensions of relevant and comparison stimuli, response variance can also be partitioned into sub-dimensions for individual relevant questions. This is similar to that of repeated-measures or within-subjects experimental designs. Within-subjects designs are useful for their efficiency and statistical power but have the potential disadvantage of complex analysis.[40] An important difference is that omnibus analytic methods are not routinely expected to provide granular conclusions, while the MFH in field polygraphy includes an implicit expectation of granular precision. In practical terms, the MFH is the hypothesis of a multiple-issue diagnostic exam for which results are classified as deceptive or truthful using subtotal scores for individual relevant questions. The MFH assumes that test outcomes are optimized by making decisions using the question subtotals instead of the overall test total. Consistent with the known limitations of omnibus analysis methods, field practitioners have adopted standards and policies that preclude conclusions that a person has been deceptive to one or more questions and truthful to one or more questions within a single exam. Nevertheless, the expectation that multi-facet questions will isolate the response variance associated with deception has been difficult to counter even though the published literature contains a number of studies that directly and indirectly address the MFH.

Published evidence at this time does not support the hypothesis that test accuracy with event-specific exams can be optimized through the interpretation of independent response variance at the level of the individual target questions. There is support for the more general hypothesis that both grand total and subtotal scores may discriminate deception and truth-telling at rates better than chance but with considerable imbalance with guilty and innocent persons. Results from these analyses indicate that overall decision

[40]Which is largely mitigated through the use of computerized statistical analysis.

accuracy of event-specific polygraph exams may be optimized by using grand total scores, thus forgoing the assumption that responses will vary independently for individual relevant questions – including when those questions employ different action verbs or describe different behavioral roles in a known or alleged incident. [41]

Accuracy with subtotal scores was found to be generally lower than for grand total scores, and imbalanced with guilty and innocent persons. Published information was insufficient to investigate the degree to which this may be improved through the use of statistically referenced cut-scores and mathematical corrections for multiplicity effects. However, evidence does not preclude the use of multi-facet questions without an assumption of independence. Effect sizes for grand total scores of examinations consisting of a combination or primary, secondary and behavioral role-descriptive questions are equivalent to published information for other single issue exams.

These results are consistent with previous reports (e.g., Senter & Dollins, 2003) that decisions using grand total and subtotal scores may provide different advantages. Overall decision accuracy was greater with grand total scores, with better balanced accuracy for guilty and innocent cases and better specificity to truth-telling compared to decisions using subtotal scores. However, decisions using subtotal scores provided increased test sensitivity to deception and reduced inconclusive rates for guilty examinees, though with disproportionately weaker accuracy with innocent persons. Considering the imbalance of effects with guilty and innocent cases and the fact that criterion accuracy for innocent cases was reduced to less than .5, these results do not support the hypothesis that test questions can be used to determine behavioral role or level of involvement of guilty examinees. Furthermore, these results do not support polygraph field practices that attempt to determine that a person has been deceptive to one or more questions and also truthful to one or more questions within a single examination.

Prior to the advent of numerical evaluation and the use of normative data, the use of test questions describing several different behavioral aspects of a known or alleged crime would have been viewed as an important strategic innovation. It is easy to understand the strategy of presenting stimulus questions with an array of factual information designed to focus the attention and responses of the guilty person - while the innocent person is expected to have no factual or behavioral association. While this strategy seems to have potential face validity, evidence-based professional practices require replicable empirical support for the continued use of a method. Published evidence suggests that the MFH contributes to a small increase in test sensitivity is associated with decisions using subtotal scores accompanied by a larger decrease in test specificity and lower overall accuracy.

Use of the term multi-facet cannot be easily found in testing contexts outside the polygraph profession. One of the earliest references to the term within the polygraph profession was in the terminology reference by Krapohl and Sturm (1997). The concept of a multiple-facet test appears to have emerged within the polygraph profession as a useful way of distinguishing between multiple-issue screening polygraphs and event-specific diagnostic polygraphs that make use of both primary and secondary relevant questions. Both multiple-issue screening polygraphs and multi-facet diagnostic polygraphs have been traditionally scored and interpreted with the assumption that responses to test questions vary independently. Although somewhat useful for its ability to remind us of the difference between investigative polygraphs of known or alleged incidents and screening tests, the term multi-facet has the potential for confounding our knowledge and understanding of the decision-theoretic and statistical concerns that determine test effectiveness – specifically, whether we assume independent or non-independent variance of the relevant questions.

[41]Neither Senter, Dollins, and Krapohl (2008), nor recent studies on the AFMGQT (Nelson & Blalock, 2012; Nelson, Blalock & Handler, 2011; Handler & Nelson, 2012; Nelson, Handler, Morgan, & O'Burke, 2012) attempted to interpret independent response variance or the criterion accuracy of the individual target questions.

Traditional practice has been to interpret these specific-incident exams using the same decision rules as with multi-issue screening exams, premised on an assumption of independent variance. Support for the validity of the multi-facet hypothesis requires that the interpretation of independent response variance leads to increased criterion accuracy. Despite a number of attempts to investigate its contribution to polygraph test effectiveness, there remains insufficient evidence to support the validity of the MFH. Evidence has consistently indicated that multi-facet questions are non-independent.

Studies dating back to the early 1970s have consistently informed us that the polygraph test is effective at discriminating between deception and truth-telling, but less effective at determining individual questions to which a person is lying or truthful. In practical terms this should be taken to mean that examinees can be said to pass or fail the test as a whole, but not the individual questions.[42]

Despite the absence of quantitative empirical support for the multi-facet hypothesis, use of multi-facet test questions may provide qualitative advantages that are not captured by a numerical or statistical scoring model. These unquantified advantages may pertain to the staging of pretest discussion and to confidence that complete and adequate test stimuli were presented to the examinee during testing. Also, field examiners may find multi-facet questions useful for developing a post-test interview strategy. Referring professionals should be cautioned against naïve expectations that multiple roles or behavior can be successfully targeted and parsed with granular precision during a single examination.

In response to anticipated arguments that it is the *utility* of the test that is optimized and optimization of test accuracy is not an objective of the MFH, we caution that utility is not optimized when the test results are more likely to produce an error that misleads an investigator or consumer. In response to numerous anecdotal stories regarding individual examinations during which an examinee exhibited strong response to a particular question with subsequent confirmation, anecdotal stories comprise a miniscule fraction of examinations that are actually conducted Evidence-based polygraph testing necessarily focuses on what happens most of the time.

Field examiners who find it useful to use multi-facet questions to stage the discussion or investigation before, during or after the polygraph test, may wish to do so using the event-specific formats developed at the University of Utah (Handler & Nelson, 2008; Kircher & Raskin, 1988; Nelson 2018b; Raskin, Honts, Nelson & Handler, 2015). These formats consist of three or four relevant stimulus questions regarding a single known or alleged incident. Relevant questions can be expressed as a single issue or can also be used to describe multiple facets of a single known or alleged incident. An example of the use single issue questions is as follows: *Did you rob that bank located at ___ in Austin?, Did you rob that bank located at ___ in Austin last Thursday?*, and, Did you rob that bank at ___ on (date)? (Nelson & Handler, 2008). An example of multiple facet test questions, involving both primary and secondary relevant questions is as follows: *Did you rob that bank at ___ on (date)?, Did you plan with anyone to rob that bank at ___?, and Did you participate in that robbery of that bank?* Regardless of the approach to target selection and question formulation, testing protocols must conform to the published literature if the test data are interpreted with the assumption that response variance is non-independent.

We caution against any expectation of perfect test accuracy, and remind that scientific tests and scientific test results are fundamentally probabilistic. The polygraph, like other tests, is a useful though imperfect tool. Recall that Podlesny and Truslow (1993) wrote:

> Users of results obtained with the present technique should be cautioned

[42]This practical approach may differ for multi-issue screening tests for which the strength of the assumption of independent criterion variance is thought to be greater. More research is needed with multiple issue exams regarding the independence of response variance. [Refer to Nelson, Blalock and Handler (2019) for additional discuss about how categorical results are parsed and reported in different ways for event-specific diagnostic exams and multiple-issue screening polygraphs.]

> that errors in classifying guilty and innocent subjects are not unlikely. The results further suggest that attempts to subcategorize deception using present MGQTs are not advisable. Where MGQT examinations are used to detect deception versus no deception in distributed-crime-role contexts, decision methods based on overall question totals might reduce inconclusives and improve accuracy with innocent subjects in comparison with methods based on individual question scores. (p.795).

A similar caution was provided previously by Raskin et al., (1988), when they wrote:

> A related problem is raised by the finding of higher false positive rates for questions answered truthfully by suspects who were also deceptive to at least one relevant question in the same test. It appears that answering deceptively to at least one relevant question in the test tends to weaken the reactions to the control questions, thereby making it difficult for them to produce reactions that are larger than those to relevant questions that are answered truthfully. Therefore, field polygraph examiners should attempt to devise sets of relevant questions that the suspect can be expected to answer all truthfully or all deceptively. The case information and the importance of each relevant question should be carefully considered in formulating the set of relevant question to be asked, and separate question series should be used whenever it seems likely that the suspect might answer some of the relevant questions truthfully and some of them deceptively.

Despite the pragmatic recommendations of Raskin *et* al. (1988) regarding the formulation of test questions, and of Podlesny and Truslow in interpreting the results, the practical appeal of multiple-issue tests has resulted in the continued use of polygraph tests for which responses test questions are assumed to vary independently. And, this practice continues to occur in both screening and diagnostic testing contexts. While there is obvious convergence of evidence regarding the MFH and the independence of questions withing multi-facet polygraph exams, conducted in the context of a specific incident or allegation, less is known about the variance of multi-issue screening polygraphs conducted in the absence of any known allegation or incident.

Barland et al. (1989) wrote:

> One interesting finding of Experiment 2 was that the examinations did not detect deception at the level of the individual crimes. This result has important implications for examiners who must test on multiple relevant issues, as it suggests that the numerical scores associated with individual relevant issues may be a poor guide in choosing issues for interrogation. This result suggests that when deception is inferred, the interrogator may need to address all of the relevant issues on the examination with the interrogation.

The available evidence at this time indicates that the MFH is false. However, this does not negate field practices that make use of combinations of primary and secondary relevant questions when it is understood that responses to multi-facet test stimuli regarding a single known or alleged incident do not vary independently. Test accuracy is optimized by interpreting responses to multi-facet questions in a manner similar to that of other event-specific examinations (i.e., assuming that the variance of responses to relevant stimuli is non-independent). In practical terms, this may require reevaluating decision policies used to achieve the final categorical result that is interpreted from the numerical data. Accurate PDD test results contribute to guide decisions about the need for subsequent discussion and investigation, but inaccurate PDD results may contribute to inaccurate risk-assessment and risk management decisions.

References

American Polygraph Association (2011). Meta-analytic survey of validated polygraph techniques. *Polygraph, 40(40)*, 203-305. [Electronic version] Retrieved January 6, 2012, from http:// www .polygraph.org.

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

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

Barland, G. H. & Raskin, D.C. (1975). Psychopathy and detection of deception in criminal suspects. *Psychophysiology,(12)*, 224.

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

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.

Cohen, B. H. (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 https://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 2-11-2010). 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 https://www.antipolygraph.org/documents/dodpi-numerical-scoring-08-2006.pdf on 6-28-2007).

Department of Defense Polygraph Institute Research Division Staff. (2001). *Test of a mock theft scenario for use in the psychophysiological detection of deception: IV (DoDPI00-R-0002).* Fort Jackson, SC: Department of Defense Polygraph Institute.

Handler, M. (2006). The Utah PLC. *Polygraph,(35)*, 139-148.

Handler, M. & Nelson, R. (2008). Utah approach to comparison question polygraph testing. *European Polygraph,(2)*, 83-110. Reprinted in Polygraph 38(1) 15-33.

Handler, M. & Nelson, R. (2012) Criterion Validity of the United States Air Force Modified General Question Technique and Three Position Scoring, *Polygraph, 41 (1).*

Handler, M., Nelson, R., Goodson, W. & Hicks, M. (2010). Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies. *Polygraph,(39)*, 200-215.

Honts, C. R., Thurber, S. & Handler, M. (2020). A comprehensive meta-analysis of the comparison question polygraph test. *Applied Cognitive Psychology, 2021*, 1-17.

Horvath F. S. (1988). The utility of control questions and the effects of two control question types in field polygraph techniques. *Journal of Police Science and Administration, 16(3)*, 198-209. Reprinted in Polygraph, 20, 7-25.

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

Horvath, F. S. & Reid, J.E. (1971). The reliability of polygraph examiner diagnosis of truth and deception. *Journal of Criminal Law, Criminology and Police Science,(62)*, 276-281.

Hunter, F. L. & Ash, P. (1973). The accuracy and consistency of polygraph examiners' diagnosis. *Journal of Police Science and Administration,(1)*, 370-375.

Keeler, L. (1930). A method for detecting deception. *American Journal of Police Science, 1*, 38-52. Reprint in Polygraph, 23, (2), 134-144.

Keeler, L. (1933). Scientific methods of crime detection with a demonstration of the polygraph. *Kansas Bar Association Journal, (12)*, 22-31.

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

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., Handler, M. & Sturm, S. (2012). PDD *Terminology Reference.* American Polygraph Association: Nashville, TN.

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

Krapohl, D. J. & Norris, W. F. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph, 29*, 185-194.

Krapohl, D. J. & Stern, B.A. (2003). Principles of Multiple-Issue Polygraph Screening: A model for applicant, post-conviction offender, and counterintelligence testing. *Polygraph, 32*, pp. 201-210.

Krapohl, D., & Sturm, S. (1997). *Terminology Reference for the Science of Psychophysiological Detection of Deception.* Chattanooga, TE: American Polygraph Association.

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

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

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

Nelson, R. (2015). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5)*, 69-82.

Nelson, R. (2018). Practical polygraph: a survey and description of decision rules. *APA Magazine, 51(2)*, 127-133.

Nelson, R. (2018). Credibility assessment using Bayesian credible intervals: a replication study of criterion accuracy using the ESS-M and event-specific polygraphs with four relevant questions. *Polygraph & Forensic Credibility Assessment 47 (1)*, 85-90.

Nelson, R. & Blalock, B. (2012). 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, 42,.*

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).*

Nelson, 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., 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).*

National Research Council (2003). *The Polygraph and Lie Detection.* National Academy of Sciences.

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

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.

Raskin, D. C., Barland, G. H., & Podlesny, J. A. (1978). *Validity and reliability of detection of deception.* Washington, D.C.: U.S. Government Printing Office.

Raskin, D. C., Kircher, J. C., Honts, C. R. and Horowitz, S. W. (1988) *A Study of Validity of Polygraph Examinations in Criminal Investigation*, Grant number 85-IJ-CX-0040. Salt Lake City: Department of Psychology, University of Utah.

Raskin, D. C., Honts, C. R., Nelson, R. & Handler, M (2015). Monte Carlo estimates of the validity of four relevant question polygraph examinations. *Polygraph, 44(1)*, 1-278.

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

Reid, J. E. & Inbau, F.E. (1977). *Truth and deception: The polygraph ('lie detector') technique (2nd ed).* Williams & Wilkins.

Senter, S. M. (2003). Modified general question test decision rule exploration. *Polygraph, 32(4)*, 251-263.

Senter, S. M., & Dollins, A. B. (2003) *New Decision Rule Development: Exploration of a Two-Stage Approach.* Department of Defense Polygraph Institute, Report No. DoDPI01-R-0007. Reprinted in Polygraph, 37(2), 149-164.

Senter, S. M., & Dollins, A. B. (2004). *Comparison of three versus three or five question series contingency rules:* A replication. Department of Defense Polygraph Institute, Report No. DoDPI01-R-0008. Reprinted in Polygraph, 33(4), 223-233.

Slowik, S. M. & Buckley, Joseph P, III (1975). Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure and GSR recordings. *Journal of Police Science and Administration,(3)*, 305-309.

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

US Customs and Border Protection (2010). *U.S. Customs and Border Protection Office of Internal Affairs Credibility Assessment Division Policy and Reference Manual with Appendices from DACA and Federal Standards.* Retrieved from https://antipolygraph.org/documents/cbp-polygraph-handbook-2010-01-07.pdf.

Wicklander, D. E. & Hunter, F.L. (1975). The influence of auxiliary sources of information in polygraph diagnosis. *Journal of Police Science and Administration,(3)*, 405-409.

Appendix A.

Table 3. Sensitivity and specificity of grand total and subtotal scores.

Study	Grand total Scores		Subtotal scores	Subtotal Scores
	Sensitivity (sd) {90% CI}	Specificity (sd) {90% CI}	Sensitivity (sd) {90% CI}	Specificity (sd) {90% CI}
Horvath and Reid, 1971	.850 (.059) {.791 to .909}	.905 (.048) {.857 to .953}	.814 (.064) {.75 to .878}	.827 (.062) {.765 to .889}
Hunter and Ash, 1973	.871 (.055) {.816 to .926}	.857 (.058) {.799 to .915}	.833 (.061) {.772 to .894}	.839 (.060) {.779 to .899}
Slowik and Buckley, 1975	.838 (.061) {.777 to .899}	.905 (.048) {.857 to .953}	.733 (.073) {.660 to .806}	.881 (.053) {.828 to .934}
Wicklander and Hunter, 1975	.942 (.038) {.904 to .98}	.867 (.056) {.811 to .923}	.905 (.048) {.857 to .953}	.808 (.065) {.743 to .873}
Raskin et al., 1988	.651 (.078) {.573 to .729}	.515 (.082) {.433 to .597}	.478 (.082) {.396 to .56}	.316 (.076) {.24 to .392}
Barland et al., 1989	.816 (.064) {.752 to .88}	.417 (.081) {.336 to .498}	.667 (.078) {.589 to .745}	.545 (.082) {.463 to .627}
Podlesny and Truslow, 1993	.694 (.076) {.618 to .770}	.750 (.071) {.679 to .821}	.847 (.059) {.788 to .906}	.375 (.080) {.295 to .455}
Krapohl and Norris, 2000[43]	.604 (.080) {.524 to .684}	.479 (.082) {.397 to .561}	.896 (.050) {.846 to .946}	.083 (.045) {.038 to .128}
Senter, 2003[44]	.530 (.082) {.448 to .612}	.746 (.072) {.674 to .818}	.944 (.038) {.906 to .982}	.212 (.067) {.145 to .279}
Senter and Dollins, 2003 (experiment 1)	.600 (.081) {.519 to .681}	.714 (.074) {.64 to .788}	.766 (.070) {.696 to .836}	.549 (.082) {.467 to .631}
Senter and Dollins, 2003 (experiment 2)	.713 (.074) {.639 to .787}	.671 (.077) {.594 to .748}	.861 (.057) {.804 to .918}	.423 (.081) {.342 to .504}
Weighted means[45]	**.660 (.078) {.582 to .738}**	**.707 (.075) {.632 to .782}**	**.833 (.061) {.771 to .894}**	**.483 (.082) {.401 to .565}**

[43]Data from Krapohl and Norris (2000) was included in the data and analysis from Senter (2003), and are excluded from the weighted mean calculations in Table 2 to avoid redundancy.

[44]The N of individual questions was imputed from the number of reported sample cases based on an assumption that this test format was commonly used with four target questions at the time the sample data were collected.

[45]Confidence intervals were calculated using a variance estimate derived from the binomial approximation to the normal distribution using the number of the number of cases, not the n of the individual questions. This approach is thought to be more conservative, with a larger variance estimate and wider corresponding confidence intervals.

Appendix B.

Table 2. Error rates for decisions based on grand total and subtotal scores.

Study	Grand total Scores		Subtotal scores	Subtotal Scores
	False-negative rate (sd) {90% CI}	False-positive rate (sd) {90% CI}	False-negative rate(sd) {90% CI}	False-positive rate (sd) {90% CI}
Horvath and Reid, 1971	.095 (.029) {.047 to .143}	.150 (.036) {.091 to .209}	.125 (.033) {.071 to .179}	.096 (.029) {.048 to .144}
Hunter and Ash, 1973	.100 (.030) {.051 to .149}	.143 (.035) {.085 to .201}	.115 (.032) {.063 to .167}	.144 (.035) {.086 to .202}
Slowik and Buckley, 1975	.152 (.036) {.093 to .211}	.067 (.025) {.026 to .108}	.182 (.039) {.119 to .245}	.065 (.025) {.024 to .106}
Wicklander and Hunter, 1975	.083 (.028) {.038 to .128}	.050 (.022) {.014 to .086}	.054 (.023) {.017 to .091}	.078 (.027) {.034 to .122}
Raskin et al., 1988	.058 (.023) {.020 to .096}	.148 (.036) {.090 to .206}	.091 (.029) {.044 to .138}	.170 (.038) {.108 to .232}
Barland et al., 1989	.051 (.022) {.015 to .087}	.182 (.039) {.119 to .245}	.188 (.039) {.124 to .252}	.468 (.050) {.386 to .55}
Podlesny and Truslow, 1993	.125 (.033) {.071 to .179}	.042 (.020) {.009 to .075}	.056 (.023) {.018 to .094}	.208 (.041) {.141 to .275}
Krapohl and Norris, 2000[46]	.25 (.043) {.179 to .321}	.104 (.031) {.054 to .154}	.001 (.003) {<.001 to .006}	.521 (.050) {.439 to .603}
Senter, 2003[47]	.037 (.019) {.006 to .068}	.432 (.050) {.351 to .513}	.006 (.008) {<.001 to .019}	.432 (.050) {.351 to .513}
Senter and Dollins, 2003 (experiment 1)	.088 (.028) {.041 to .135}	.027 (.016) {<.001 to .054}	.051 (.022) {.015 to .087}	.139 (.035) {.082 to .196}
Senter and Dollins, 2003 (experiment 2)	.07 (.026) {.028 to .112}	.055 (.023) {.017 to .093}	.049 (.022) {.013 to .085}	.228 (.042) {.159 to .297}
Weighted means[48]	.072 (.035) {.030 to 0.115}	.119 (.053) {.066 to .173}	.050 (.036) {.014 to .086}	.207 (.067) {.140 to .273}

[46]Data from Krapohl and Norris (2000) was included in the data and analysis from Senter (2003), and are excluded from the weighted mean calculations in Table 2 to avoid redundancy.

[47]The number of individual questions was imputed from the number of reported sample cases based on an assumption that this test format was commonly used with four target questions at the time the sample data were collected.

[48]All standard deviations were calculated using binomial approximation to the normal distribution using the number of the number of cases. for both grand total and subtotals. This approach is thought to be more conservative, with a larger variance estimate and wider corresponding confidence intervals.