Meta-Analytic Survey of Criterion Accuracy of Validated Polygraph Techniques

Report Prepared For

The American Polygraph Association Board of Directors

Nate Gordon, President (2010-2011)

by

The Ad-Hoc Committee on Validated Techniques

Mike Gougler, Committee Chair

Raymond Nelson, Principal Investigator

Mark Handler

Donald Krapohl

Pam Shaw

Leonard Bierman

Introduction

Pamela Shaw
President
American Polygraph Association

Over the past few years the APA has carefully undertaken a number of significant initiatives that have affected its members. Prompted by the National Academy of Sciences report of 2003, and further energized by the ongoing review by the forensic science community, APA leaders and members recognized that there were real dangers ahead if we continued on the road we had always taken.

To strategically plan for and ensure our survival in the years ahead, the APA has been implementing initiatives to set in motion needed improvements. In particular, the APA is seeking to increase the level of science in our practices, to standardize our methodologies, to focus on continuous improvement, to upgrade our education, and broaden our vision to cover not only the interests of members, but to include protection of the public, as well. In other words, we are proactively pursuing increased professionalism in polygraphy.

In pursuit of this goal, the APA Strategic Plan has identified several key milestones along the way. They include measures such as development of best practice model polices, mandating continuing education, and making polygraph research articles available on the APA website for easy access by members. One of the key steps is to require APA members to use methods that could be defended by published research beginning in 2012. It is this last step that is the subject of this special edition of *Polygraph*.

The requirement to use validated testing methods is not a new idea, of course. Other fields such as medicine and psychology eventually came to the same conclusion, albeit many years after the fields were established. It has turned out to be a great thing for them. Try to imagine, if you can, what the fields of medicine and psychology would be like if there were no requirement to validate their methods. Validation serves a number of important functions, not the least of which is protecting the public from misuse, incompetence and quackery. The net effect is not just to help the public but to elevate the profession; a winning solution for everyone. It is from this perspective of "enlightened self-interest" that arise the polygraph initiatives of recent years.

As we approached the year 2012 and the validation requirement for our techniques, it became clear that some APA members may not have received instruction from their polygraph schools as to which methods had scientific support. It seemed reasonable that the APA should undertake a literature review to see what evidence was available, and provide that to the membership. Earlier this year, then-President Nate Gordon established an ad hoc committee to accomplish this task. I was proud to serve with that committee. Over the following months the committee worked feverishly to complete a report that could be available to the membership by the end of the year. In the following pages you will see their findings.

As you read the committee report, there are a few caveats that you need to know. First, as the report itself says, this committee report is not APA policy, but rather a literature summary. There are other literature summaries published by APA in the past 20 years that were likewise not APA policy. The report is for information only, and may be helpful to members already using or intending to use the techniques listed.

Second, the Board has come to be aware of pending research studies on screening techniques. The Board will consider new By Law provisions on screening methodologies to encourage members to use the best methods available. Those decisions will be made in early 2012.

And finally, the APA will extend a great deal of effort in its seminars to help members become familiar and competent with techniques that meet the validity requirements.

Many of you have expressed support, encouragement and championed the on-going efforts towards validation and higher standards: for that, I thank you. I would also like to acknowledge and thank the unrelenting and devoted efforts of the committee members for the countless hours they've dedicated to this project. Your efforts are key to our future endeavors.

We are at a great time in polygraph history and we can be proud of the steps we are taking to move our profession forward. We must all grow with the knowledge in our field and the demands within our field to ensure our future success. This report is an essential step in that direction and I am proud to provide this document to you in this special issue of *Polygraph*.

Executive Summary

In 2007 the American Polygraph Association (APA) adopted a Standard of Practice, effective January 1, 2012, that requires APA members to use validated Psychophysiological Detection of Deception (PDD) examination techniques that meet certain levels of criterion accuracy.[1] Those requirements state that event-specific diagnostic examinations used for evidentiary purposes must be conducted with techniques that produce a mean criterion accuracy level of .90 or higher, with an inconclusive rate of .20 or lower. Diagnostic examinations conducted using the paired-testing protocol must produce a mean criterion accuracy level of .86 or higher, with inconclusive rates of .20 or lower. Examinations conducted for investigative purposes must be conducted with techniques that produce a mean criterion accuracy level of .80 or higher, with inconclusive rates of .20 or lower.[2] The goal is to eliminate the use of un-standardized, non-validated or experimental techniques in field settings where decisions may affect individual lives, community safety, professional integrity, and national security.

There exists today a confusing array of test question formats that are at once similar and dissimilar, and for which there are also alternatives in the selection of a method for test data analysis. Equally confusing is the abundance of published research, and the meaning and applicability of that research to the techniques used in field settings. The APA Board of Directors assumed responsibility for organizing this information in the form of a systematic review and meta-analysis of the published scientific literature which describes the criterion validity of presently available polygraph techniques. In the course of doing this, it has at times been necessary to define what appear to be obvious concepts. One such concept is that of validation, which, as it applies to PDD exams, is stipulated by the APA Standards of Practice (Section 3.2.10) to refer to the combination of: 1) a test question format that conforms to valid principles for target selection, question construction, and in-test presentation of the test stimuli, and 2) a validated method for test data analysis as it applies to a specified test question format. Although many factors may affect the overall effectiveness of PDD examinations, these two parts are recognized as fundamental to the criterion accuracy of PDD examinations. The accuracy of all tests is contingent upon these two activities: obtaining a sufficient quantity of diagnostic information, and interpreting the information correctly. The two-fold purpose of this meta-analysis was to advise the APA and its membership about which PDD techniques satisfy the standard practice requirements that take place January 1, 2012, and to answer questions about our present knowledge-base regarding the criterion validity of PDD techniques as they are presently used.

The ad hoc committee to examine the evidence on the criterion accuracy of polygraph techniques was appointed by APA President Nate Gordon during the March 2011 meeting of the Board of Directors. The committee was composed of Past President and Board Director Mike Gougler (committee chair), Past-President and Editor-in-Chief Don Krapohl, President Elect Pam Shaw, Board Director Raymond Nelson, and APA Members Mark Handler and Leonard Bierman.

The committee also took into consideration that there are both financial and proprietary issues attached to the formulation of such a list. The stakeholders represent a diverse group of professionals and interests. The effectiveness of the APA and the

[1] Criterion accuracy refers generally to the degree to which a test result corresponds with what the test is designed to detect. In the field of PDD, criterion accuracy denotes the ability of a combination of testing and scoring techniques to discriminate between truthful and deceptive examinees, and ranges from 0.00 for no validity to 1.00 for perfect validity. Criterion accuracy is one form of validity, and in some research reports it may be referred to as decision accuracy, or just accuracy.

[2] Near the completion of this study the APA Board of Directors enacted a change in standards, endorsing the use of PDD screening techniques for which research indicates an accuracy rate that is significantly greater than chance.

credibility of the polygraph profession required the committee to give precedence to the accuracy and integrity of the research review over the financial and personal interests of any individual developer of PDD testing techniques. The committee's default approach was an inclusionary review process in which any stakeholder could submit supportive data and information for consideration. This did not mean that anything submitted would automatically be included or endorsed as valid, but it did mean that all recommendations would be considered.

The committee began its process with a discussion of the merits and strengths of laboratory and field research. Field studies are important to polygraph research as these studies have the advantage of ecological validity[3] and are therefore assumed to have increased generalizability. However, the generalizability of field studies is compromised to some unknown extent by the selection process which necessarily depends on the availability of often-incomplete confirmation data. Real world confirmation data are selective, neither random nor representative of all data, and confirmed cases more often may have correct PDD results than do unconfirmed cases. As a result, field studies may overestimate PDD decision accuracy to some degree. While field studies are highly useful for studying correlations, they provide imperfect measures of criterion validity.

Laboratory studies are also important to polygraph research as these studies can more easily control and reduce research and sampling biases. Use of experimental and quasi-experimental research designs, along with random sampling and random criterion assignment, can increase the generalizability and repeatability of research results. Because of their ability to control a greater number of variables, laboratory studies are fundamental to our ability to study questions of causality and construct validity. However, the generalizability of laboratory studies is complicated by the fact that these studies may not represent the broad range of variables thought to influence the results of field examinations. They are therefore presumed to have ecological validity that is weaker, to some unknown degree, than that of field studies.

The committee took the position that both field and laboratory studies have advantages and disadvantages, and that neither type alone would be sufficient to study all of the issues of concern to polygraph researchers. Both types of research are of vital importance to the study and development of knowledge in the polygraph profession. Differences between criterion accuracy of field and laboratory studies were found to be statistically small and insignificant by the 2002 report on the polygraph by the US National Research Council. For the purpose of reviewing the current state of validation of existing polygraph techniques, the committee gave field and laboratory studies equal consideration.

The committee compiled a list of studies that satisfied the qualitative and quantitative requirements for inclusion in the review, and for which there existed two or more satisfactory publications that describe generalizable evidence of criterion validity. The committee formulated recommendations regarding which techniques satisfied the requirements of the pending 2012 APA provisions for criterion accuracy. The committee then completed a meta-analysis that bench-marked the findings against the 2012 APA standards for evidentiary, paired testing, and investigative examinations. In addition, statistical tests were completed to check for study integrity, and to search for inconsistencies and outlier results.

[3] Ecological validity addresses how well the experimental settings, processes, subjects and materials match those in real-life conditions. Though it is not the same as external validity, greater ecological validity may provide more confidence that the findings of the study will generalize to other settings.

Inclusion in the research review required that studies in the meta-analysis be published in *Polygraph* or other peer-reviewed scientific publications.[4] Studies were considered for selection if they were published by an academic degree-granting institution that was accredited by an accrediting agency recognized by the US Department of Education or foreign equivalent. In addition, research publications of studies funded by government agencies were also considered for selection. Edited academic texts, including individual chapters, were considered. However, studies available only in self-published books were excluded. Additional qualitative requirements were that selected studies must have employed a recognizable PDD technique for which a published description exists for the test structure, test question sequence, and administration. Studies must also have used a recognizable test data analysis (TDA, chart interpretation) method and included a description of the features, numerical transformations (score assignments), decision rules and normative data or cutscores. In addition, studies must have used instrumentation and component sensors that reflect common field practices. Selected studies must have used a variety of confirmation criteria including examinee confessions, high quality forensic evidence, or substantiated evidence that the crime was not committed.[5] Study results based on samples that were subject to experimental manipulation (e.g., fatigue, intoxication, programmed countermeasure use) were not included.[6]

Quantitative information requirements included some form of reliability statistics for each technique as evidence of the generalizability of the results. Several types of statistical measurements for PDD test reliability were reported in the published literature, and all were accepted for inclusion.[7] The committee also evaluated the reliability, generalizability and representativeness of the sample distributions through multivariate ANOVAs using the deceptive and truthful scores. It was expected that multiple samples drawn from the same underlying population, administered the same PDD technique, and scored with the same TDA method would replicate among the sampling distributions of scores. It was also expected that aggregation of the results of replicated sampling distributions would be more representative and generalizable than the results from any single sampling distribution. In addition to sample size information, a minimum of four statistical values were required for the meta-analysis: test sensitivity and specificity, and the inconclusive rates for guilty and innocent cases.

It was important to the credibility of the committee's findings that studies be accepted or endorsed based on the merits of the included studies. For this reason, access to the published evidence and raw data was made a priority, and the list of validated techniques was constructed according to a systematic review of published research. In addition to a re-analysis of the study data, studies were included when it was possible to calculate a complete dimensional profile of criterion accuracy.

Following the completion of the literature survey, a list of all identified techniques was sent to the school directors or representatives of all APA accredited polygraph schools. School directors or their representatives were invited to provide any

[4] The journal *Polygraph* instituted expert peer review in 2003. Articles published prior to that time were subject only to editorial review. Because Polygraph is an important academic and historic resource, studies published prior to 2003 and without peer review were included in this meta-analysis if they satisfied all of the other qualitative and quantitative requirements for selection.

[5] One included study did not meet this requirement, and consisted only of cases confirmed by confession. Consistent with the known concern about inflated accuracy estimations resulting from over-reliance on confession confirmation for sample case selection, this study reported a near-perfect level of decision accuracy.

[6] Present standards for research and publication by the APA stipulate that principal investigators should not also serve as study participants (i.e., examinees, examiners, or scorers). However, this was not a requirement in the past, and studies were not excluded from the meta-analysis based on this criterion.

[7] One included technique is without published evidence of inter-scorer reliability or agreement.

published studies or citations to published studies for techniques that were not yet identified. One additional technique was suggested for inclusion at that time.[8] In a follow-up mailing, a shorter list was sent to all school directors, or representatives, and other researchers involved in the development or validation of PDD techniques. This list included only those techniques for which published and replicated studies were identified, along with another request to provide any publications or citations for techniques that were not yet included in the survey. Two additional studies were suggested for inclusion at that time.[9] Following the submission of initial results to the APA Board of Directors, and presentation of a preliminary version of this Executive Summary to the APA membership, one additional, recently published, study report was submitted in support of the IZT.[10]

The committee contacted developers of PDD techniques for which there were insufficient published studies for inclusion in the meta-analysis, and volunteered assistance to anyone requesting it. Two studies were completed, have been accepted for publication, and are awaiting printing, for the Backster You-Phase technique. This technique was included in the meta-analysis. Additional studies were also completed for the Air Force Modified General Question Test (AFMGQT), the Federal You-Phase Technique, and the Directed Lie Screening Test (DLST), which have also been accepted for publication. The result of these additional efforts was that the complete array of techniques in common use today was included in the meta-analysis.

Thirty-eight studies satisfied the qualitative and quantitative requirements for inclusion in the meta-analysis. These studies involved 32 different samples, and described the results of 45 different experiments and surveys. These studies included 295 scorers who provided 11,737 scored results of 3,723 examinations, including 6,109 scores of 2,015 confirmed deceptive examinations, 5,628 scores of 1,708 confirmed truthful exams. Some of the cases were scored by multiple scorers and using multiple TDA methods.

Table 1 at the end of this executive summary shows the findings for those methods for which there is published and replicated evidence of criterion validity that satisfies the requirements of the APA Standards of Practice. Five PDD testing and analysis combinations meet APA 2012 requirements for evidentiary testing, five for paired-testing,[11] and four for investigative examinations.[12]

Two PDD techniques produced accuracy rates that were outliers from and inconsistent with the distribution of results from all other techniques. They were the Integrated Zone Comparison Technique (IZCT) and the Matte Quadri-Track Zone Comparison

[8] The Marcy Technique was suggested for inclusion. However, no published studies could be located regarding this technique.

[9] The Gordon et al (2000) study of the Integrated Zone Comparison Technique (IZCT) could not be included due to a lack of adequate statistical information. The primary author informed the committee that the report was completed without him seeing the data, which belong to the intelligence service of a foreign government and therefore unavailable. Data for the Mangan, Armitage and Adams (2008) study of the Matte Quadri-Track Zone Comparison Technique were provided to the committee, and this study was included.

[10] A portion of the data for the Shurani (2011) study of the IZCT was provided to the committee and this study was included.

[11] All PDD techniques that meet the criterion accuracy requirement for paired-testing also meet the standard requirement for investigative testing, and those techniques that meet the standard requirement for evidentiary work also meet the requirements for paired-testing and investigative testing.

[12] All techniques that employed three-position TDA methods consistently exceeded the 2012 boundary requirements for inconclusive rates (20%). Because criterion accuracy rates for techniques with three-position TDA did not differ significantly from seven-position criterion accuracy, field practices that involve an initial analysis with the three-position TDA method may be considered acceptable if inconclusive results are resolved via subsequent analysis with a TDA method that provided both accuracy and inconclusive rates that meet the requirements of the APA 2012 standards.

Technique (MQTZCT). While it is within the realm of possibility that these two techniques are superior to other techniques, studies supporting them proved to have more unresolved methodological issues than others included in this meta-analysis. In addition to the committee's discovery of anomalous sampling distributions, both of these techniques are supported by studies authored by the developers and proprietors, and for which the developer/proprietor functioned as both principal investigator and study participant. From a scientific perspective, even well designed research generated by advocates of a method who have a vested interest in the outcome, and who act as participants and authors of the study report does not have the compelling power of research not so encumbered by these factors. The techniques have been duly included here because they met the more general requirements outlined in the APA Standards of Practice. The committee advises that because of the potential impact on examiner effectiveness that could result from reliance on outlier results, it would be prudent for examiners to exercise an extra measure of caution before accepting data from studies showing extraordinary effects before they are subject to independent confirmation and extended analysis.

A dimensional profile of criterion accuracy was calculated for each PDD technique, including the unweighted average of the proportions of correct decisions for deceptive and truthful cases, excluding inconclusive results, along with the unweighted average of the proportions of inconclusive results.[13] Results were aggregated for techniques that satisfy the APA 2012 requirements for evidentiary testing, paired testing, investigative testing, and for all PDD techniques included in the meta-analysis. Excluding outlier results, comparison question techniques intended for event-specific (single issue) diagnostic testing, in which the criterion variance of multiple relevant questions is assumed to be non-independent,[14] produced an aggregated decision accuracy rate of .890 (.829 - .951), with a combined inconclusive rate of .110 (.047 - .173). Comparison question PDD techniques designed to be interpreted with the assumption of independence of the criterion variance of multiple relevant questions, produced an aggregated decision accuracy rate of .850 (.773 - .926) with a combined inconclusive rate of .125 (.068 - .183). The combination of all validated PDD techniques, excluding outlier results, produced a decision accuracy level of .869 (.798 - .940) with an inconclusive rate of .128 (.068 - .187). Data at the present time are sufficient to support the polygraph as highly accurate, but insufficient to support an assertion that PDD testing can provide perfect or near-perfect accuracy.

Excluding outlier results, multi-variate analysis showed there were no significant one-way differences in decision accuracy for any of the PDD techniques that satisfy the requirements of the APA 2012 standards of practice. Neither were there any significant one-way differences in PDD techniques at the different levels of validation specified in the APA standards of practice, or for PDD techniques interpreted with decision rules based on an assumption of independence when compared with PDD techniques interpreted with decision rules based on an assumption of non-independence. This illustrates that the APA categorical distinctions are arbitrary, not empirically founded, and scientifically meaningless. These data are insufficient to support the notion that any PDD technique is superior to another, and instead suggest that differences in PDD question formats may be less important than

[13] The unweighted average was considered to be a more conservative and realistic calculation of the overall accuracy of all PDD examination techniques. Calculation of the weighted average, or the simple proportion of correct decisions, often results in higher statistical findings that are less robust against differences in base-rates and therefore less generalizable.

[14] Independence, in scientific testing, refers to assumptions about whether external factors that affect the criterion state of each question (i.e. truthfulness about past behavior) is assumed to affect the criterion state of other questions. In PDD testing, the results of multi-facet and multi-issue exams are interpreted with decision rules based on the assumption of independence, while the results of event-specific single-issue examinations are more often interpreted with decision rules based on the assumption of non-independence.

previously assumed. Differences in PDD techniques may be limited to assumptions and procedural differences pertaining to TDA methodologies intended for the investigation of independent or non-independent investigation target questions. This should be subject to further study.

The present evidence supports an argument that PDD testing can provide both test sensitivity to deception and test specificity to truth-telling at rates that are significantly greater than chance when conducted and interpreted with the assumptions of criterion independence as well as non-independence among the test questions. Evidence shows that all PDD techniques included in the meta-analysis provide test sensitivity at rates that are significantly greater than chance. However, the present evidence is insufficient to support that every PDD technique is capable of providing test specificity to truth-telling at rates that are significantly greater than chance. Details can be found in the complete report.

If the present results were to be considered an overestimation of PDD accuracy, the major causes of that over-estimation would be deficiencies in the sampling methodologies. One such factor is over-reliance on case confirmation via examinee confession, which may present the potential for the systematic exclusion of false-positive or false-negative errors for which no confession would be obtained. Another sampling concern would be publication or *file-drawer* bias, in which less favorable results are not submitted for publication and thus not available for inclusion in a meta-analysis or other systematic review. Another potential cause of accuracy overestimation would be the lack of independence between the technique developer, principal investigator and examiner participants in some included studies. If the present results were to be considered an under-estimation of PDD accuracy, the major cause might be argued to be deficiencies in the ecological validity of experimental and survey methodologies of the included studies. The present results are intended only to summarize the presently available publications that satisfy the requirements for inclusion in the meta-analysis. Limitations of the meta-analysis are discussed in the full report.

In closing, no attempt should be made to represent the results of this meta-analytic survey as an enforceable policy or standard. Although the dissemination of a list of validated polygraph techniques, could be viewed by some as a form of de facto APA endorsement of those techniques, the actual role of this meta-analysis is as a thorough summary of the existing PDD literature. Although policies may tend to remain fixed for periods of time, scientific evidence is continuously evolving. The committee offers that questions and discussions of test validity are a matter of science and not mere policy. As such these questions are best answered by scientific evidence. This meta-analysis should be considered an information resource only, and no attempt should be made to represent this list of PDD techniques as the final authority on PDD test validation. Although completed with the goal of creating a comprehensive and inclusive list of validated techniques, it remains possible that other studies and techniques exist but have not been included in this meta-analysis. There exists in the published literature some evidence of validity for PDD techniques that were not able to be included in this meta-analytic survey. Of course, a meta-analysis based on different study selection or inclusion criteria may yield different results. Nothing in this executive summary or the complete report should be construed as preventing the use of any PDD technique for which the criterion accuracy level can be defended with scientific evidence. The information herein is provided to the APA Board to advise its professional membership of the strength of validation of PDD techniques available at this time. This information is intended only to ease the burden on PDD professionals and to help make evidence-based decisions regarding the selection of PDD techniques for use in field settings. It may also assist program administrators, policy makers, and courts to make evidence-based decisions about the informational value of PDD test results in general.

Table 1. Mean (standard deviation) and {95% confidence intervals} for correct decisions (CD) and inconclusive results (INC) for validated PDD techniques. References can be found at the end of the complete report.

Evidentiary Techniques/ TDA Method	Paired Testing Techniques/ TDA Method	Investigative Techniques/ TDA Method
Federal You-Phase / ESS[1] CD = .904 (.032) {.841 to .966} INC = .192 (.033) {.127 to .256}	AFMGQT[4,8] / ESS[5] CD = .875 (.039) {.798 to .953} INC = .170 (.036) {.100 to .241}	AFMGQT[6,8] / 7 position CD = .817 (.042) {.734 to .900} INC = .197 (.030) {.138 to .255}
Event-Specific ZCT / ESS CD = .921 (.028) {.866 to .977} INC = .098 (.030) {.039 to .157}	Backster You-Phase / Backster CD = .862 (.037) {.787 to .932} INC = .196 (.040) {.117 to .275}	CIT7 / Lykken Scoring CD = .823 (.041) {.744 to .903} INC = NA
IZCT / Horizontal[2] CD = .994 (.008) {.978 to .999} INC = .033 (.019) {.001 to .069}	Federal You-Phase / 7 position CD = .883 (.035) {.813 to .952} INC = .168 (.037) {.096 to .241}	DLST (TES)[8] / 7 position CD = .844 (.039) {.768 to .920} INC = .088 (.028) {.034 to .142}
MQTZCT / Matte[3] CD = .994 (.013) {.968 to .999} INC = .029 (.015) {.001 to .058}	Federal ZCT / 7 position CD = .860 (.037) {.801 to .945} INC = .171 (.040) {.113 to .269}	DLST (TES)[8] / ESS CD = .858 (.037) {.786 to .930} INC = .090 (.026) {.039 to .142}
Utah ZCT DLT / Utah CD = .902 (.031) {.841 to .962} INC = .073 (.025) {.023 to .122}	Federal ZCT / 7 pos. evidentiary CD = .880 (.034) {.813 to .948} INC = .085 (.029) {.028 to .141}	-
Utah ZCT PLT / Utah CD = .931 (.026) {.879 to .983} INC = .077 (.028) {.022 to .133}	-	-
Utah ZCT Combined / Utah CD = .930 (.026) {.875to .984} INC = .107 (.028) {.048 to .165}	-	-
Utah ZCT CPC-RCMP Series A / Utah CD = .939 (.038) {.864 to .999} INC = .185 (.041) {.104 to .266}	-	-

[1] Empirical Scoring System.

[2] Generalizability of this outlier result is limited by the fact that no measures of test reliability have been published for this technique. Also, significant differences were found in the sampling distributions of the included studies, suggesting that the samples data are not representative of each other, or that the exams were administered and/or scored differently. One of the studies involved a small sample (N = 12) that was reported in two articles, for which the participating scorer was also the technique developer. One of the publications described the study as a non-blind pilot study. Both reports indicated that one of the six truthful participants was removed from the study after making a false-confession. The reported perfect accuracy rate did not include the false confession. Neither the perfect accuracy nor the .167 false-confession rate are likely to generalize to field settings.

[3] Generalizability of this outlier result is limited by the fact that the developers and investigators have advised the necessity of intensive training available only from experienced practitioners of the technique, and have suggested that the complexity of the technique exceeds that which other professionals can learn from the published resources. The developer reported a near-perfect correlation coefficient of .99 for the numerical scores, suggesting an unprecedented high rate of inter-scorer agreement, which is unexpected given the purported complexity of the method. Additionally, the data initially provided to the committee for replication studies included only those cases for which the scorers arrived at the correct decision, excluding scores from those cases for which the scorers did not achieve the correct decision. Missing scores were later provided to the committee for both the Mangan et al (2008) and Shurani and Chavez (2009) studies. However, the resulting sampling means were different from those reported for both replication studies. Because of these discrepancies, the statistical analysis was not re-calculated with the missing scores, and the reported analysis reflects the sampling distribution means as reported. Sampling means for replication studies should be considered devoid of error or uncontrolled variance.

[4] Two versions exist for the AFMGQT, with minor structural differences between them. There is no evidence to suggest that the performance of one version is superior to the other. Because replicated evidence would be required to reject a null-hypothesis that the differences are meaningless, and because the selected studies include a mixture of both AFMGQT versions, these results are provided as generalizable to both versions. AFMGQT exams are used in both multi-facet event-specific contexts and multi-issue screening contexts. Both multi-facet and multi-issue examinations were interpreted with decision rules based on an assumption of criterion independence among the RQs.

[5] The AFMGQT produced accuracy that is satisfactory for paired testing only when scored with the Empirical Scoring System.

[6] There are two techniques for which there are no published studies but which are structurally nearly identical to the AFMGQT: the LEPET and the Utah MGQT. Validity of the AFMGQT can be generalized to these techniques if scored with the same TDA methods.

[7] Concealed Information Test, also referred to as the Guilty Knowledge Test (GKT) and Peak of Tension test (POT). The data used here were provided in the meta-analysis report of laboratory research by MacLaren (2001).

[8] Studies for these PDD techniques were conducted using decision rules based on the assumption of criterion independence among the testing targets. Accuracy of screening techniques may be further improved by the systematic use of a successive-hurdles approach.

Report of the Ad Hoc Committee on Validated Techniques

Abstract

Meta-analytic methods were used to calculate the effect size of validated psychophysiological detection of deception (PDD) techniques, expressed in terms of criterion accuracy. Monte Carlo methods were used to calculate statistical confidence intervals. Results were summarized for 45 different samples from experiments and surveys, including scored results from 295 scorers who provided 11,737 scored results of 3,723 examinations, including 6,109 scores of 2,015 confirmed deceptive examinations, 5,628 scores of 1,708 confirmed truthful exams. Fourteen different PDD techniques were supported by a minimum of two published studies each that satisfied the qualitative and quantitative requirements for inclusion in the meta-analysis. Results for the individual studies, and for different PDD techniques, were compared using multivariate analytic methods. Two studies produced outlier results that are not accounted for by the available evidence and which are not generalizable. Excluding outliers, there were no significant differences in criterion accuracy between any of the PDD techniques supported by the selected studies. Excluding outlier results, comparison question techniques intended for event-specific (single issue) diagnostic testing, in which the criterion variance of multiple relevant questions is assumed to be non-independent, produced an aggregated decision accuracy rate of .890 (.829 - .951), with a combined inconclusive rate of .110 (.047 - .173). Comparison question PDD techniques designed to be interpreted with the assumption of independence of the criterion variance of multiple relevant questions (multiple-issue and –facet) produced an aggregated decision accuracy rate of .850 (.773 - .926) with a combined inconclusive rate of .125 (.068 - .183). The combination of all validated PDD techniques, excluding outlier results, produced a decision accuracy of .869 (.798 - .940) with an inconclusive rate of .128 (.068 - .187).

Introduction

Is the polygraph scientifically valid? How accurate is the polygraph? The simplicity of these common questions implies that the accuracy of psychophysiological detection of deception (PDD) tests can be described with a simple answer or with a single numerical index. The present approach to answering these and other questions about criterion validity[1] is a meta-analytic review of all available studies of criterion accuracy for all PDD examination techniques. Because the comparison question test (CQT) is the most commonly used and researched of all PDD techniques, this analysis will be primarily directed toward the CQT. There will be a limited discussion of research supporting the use of the Concealed Information Test (CIT).

The origins of all modern CQT formats can be traced to Reid (1947) who showed that some form of comparison question (CQ), intended to evoke a response from a truthful examinee, could improve test accuracy and reduce the occurrence of false-positive errors. CQT formats will often fall into one of two major families of techniques: techniques that emerged as modifications of the technique described by Reid (1947), and techniques that emerged as modifications of the technique described by Backster (1963). These techniques are conducted using differing, though often similar, procedures based on differing assumptions. These different assumptions and procedures can yield differences in test performance or test accuracy. Some techniques are highly theoretical about the exact nature and cause

[1] The terms “accuracy,” “test accuracy,” “criterion accuracy,” and “criterion validity” are used interchangeably and synonymously throughout this document. The term “decision accuracy” is also used to describe criterion validity, but in a more limited sense, referring only to the accuracy of decisions, excluding inconclusive results. In a more complete sense, the term “criterion accuracy” refers to a dimensional set of concerns involving all aspects of test accuracy.

of emotional or cognitive activity and resultant psychophysiological changes. Other techniques have emphasized an evidence-based scientific approach which forgoes unproven hypotheses and complex psychological assumptions about the exact thoughts and emotions of the examinee.

In general, the family of Zone Comparison Test[2] (ZCT) formats that emerged from the work of Backster (1963) has been used most effectively for event-specific diagnostic testing. ZCT questions are formulated to describe the examinee's involvement in a single known or alleged behavioral issue of concern, and are interpreted with decision rules based on an assumption of non-independence[3] of the criterion variance of the test questions. In contrast, the family of Modified General Question Test[4] (MGQT) formats that has emerged from the work of Reid (1947) are intended to describe and evaluate the examinee's involvement in different behavioral roles or different levels of involvement in a known or alleged incident. Although research has supported the CQT as capable of providing accuracy at levels that are significantly greater than chance, previous research (Barland, Honts & Barger 1989; Podlesny & Truslow, 1993; Research Division Staff, 1995a, 1995b) has not supported the effectiveness of polygraph questions at pinpointing the exact behavioral role or level of involvement within a event-specific examination. In addition to their use in multi-facet investigations of known or alleged incidents, MGQT techniques are easily adapted to use in multi-issue screening contexts in which test questions are formulated to describe the examinee's possible involvement in several different behaviors for which there is no known incident or allegation. Both multi-facet and multi-issue MGQT examinations are commonly interpreted with decision rules based on an assumption that the criterion variance of the relevant question (RQ) stimuli is independent. As a matter of field practice, both families of techniques have at times been used under assumption of both independence and non-independence among the RQs.

Other Reviews

Previous systematic reviews have been completed in an attempt to provide objective answers to the question of PDD accuracy and reconcile this with claims of perfection. Abrams (1973) reviewed polygraph validity studies dating back to the early part of the 20th century, and reported an accuracy rate of .980. Later, Abrams (1977) reported the average accuracy of polygraph validity studies to be .910. Still later, Abrams (1989) summarized the accuracy of polygraph tests as .880.

Ansley (1983) reported the results of 1,964 laboratory cases and 1,113 field cases and described a decision accuracy level of .968, excluding inconclusive results. However, accuracy rates were not reported separately for criterion deceptive and truthful cases, and inconclusive rates were not reported. At that time the Relevant-Irrelevant technique was reported as more accurate (.960) than CQT methods (.952). Ansley (1983) reported the accuracy of CIT formats to be .912. Later, Ansley (1990) summarized the results of 10 field examinations, involving 2,042 criminal cases since 1980, reporting an overall accuracy rate of .980 for deceptive cases and .970 for truthful cases, using the decisions of the original examiners. Ansley (1990) also described the results of 11 studies of *blind* evaluations of 922 criminal examinations, reporting accuracy levels of .900, with a reported accuracy rate of .940 for deceptive cases and .890 for truthful cases.

[2] Sometimes referred to as the historically correct expression "Zone Comparison Technique" as well as the "Zone of Comparison Technique."

[3] Independence, in scientific testing, refers to assumptions about whether external factors that affect the criterion state of each question (i.e. truthfulness about past behavior) is assumed affect the criterion state of other questions. In PDD testing, the results of multi-facet and multi-issue exams are interpreted with decision rules based on the assumption of independence, while the results of event-specific single-issue examinations are more often interpreted with decision rules based on the assumption of non-independence.

[4] Also referred to as the Modified General Question Technique.

Honts and Peterson (1997) and Raskin and Honts (2002) reported the accuracy of the polygraph as exceeding .900. It is consistent with the .900 accuracy estimation of Raskin and Podlesny (1979). In contrast, the systematic review completed by the Office of Technology Assessment (OTA, 1983) suggested that laboratory studies had an average unweighted accuracy of .832 with an inconclusive rate of .269, while field studies had an average unweighted accuracy rate of .847 with an inconclusive rate of .042. Crewson (2001) surveyed studies of diagnostic and screening polygraphs[5] in a comparison with medical and psychological tests, and reported diagnostic polygraphs to have an average accuracy rate of .880. Crewson also reported the average accuracy of screening polygraphs as .740. The Crewson information was also reported by Blackstone (2011) who argued that the confusion between diagnostic and screening polygraphs was a reason the polygraph did not enjoy greater support from the law.[6]

The most recent scientific review was completed by the National Research Council (NRC, 2003) who reported accuracy rates, in terms of the area under the curve (AUC) using the receiver operating characteristic (ROC) analysis, and concluded that laboratory studies had an average AUC of .860 while field studies had an average AUC of .890. More recently, Kokish, Levenson, and Blasingame (2005) reported the results of an opinion survey of sex offenders subject to polygraph monitoring as a condition of supervision and treatment. They reported that the offenders expressed a high rate of agreement with the results of the polygraph, over .900, but cautioned that offenders also claimed a false admission rate of approximately .050.

Although valuable in some ways, none of these previous surveys are capable of providing a satisfactory level of guidance regarding the American Polygraph Association (APA) 2012 standard for the use of validated techniques. These previous studies do not include information describing more recent advances in PDD criterion accuracy, and none of these surveys does an adequate job providing a complete dimensional profile of criterion validity of individual PDD examination techniques. More importantly, none of these previous surveys satisfies the need for summary information regarding study replication and the level of reliability and generalizability of study results for individual PDD techniques as used in field practice.

All previous reviews of PDD test accuracy are unsatisfying in their ability to answer the present question regarding the validity, criterion accuracy and reliability of PDD techniques in use today. First, previous reviews do not address test accuracy with an adequate description of the procedural combination of the test question sequence and test data analysis (TDA) method applied in the study, both of which are thought to have an important impact on test effectiveness. Second, and related to the first concern, is that none of the previous reviews made an effort to exclude PDD techniques that are no longer taught at accredited training programs or have fallen out of use in the field. As a result, a number of previous reviews are of

[5] Diagnostic tests are any tests conducted in response to known problems, known symptoms, known incidents, or known allegations. Screening tests are any tests conducted in the absence of a known problem, and are intended to search for possible problems. In practice, diagnostic tests are commonly formulated around a single issue of concern. Screening tests, because of the absence of any known problems, and because of interest in several types of possible problems, are often constructed around multiple issues. The terms multi-issue and mixed-issue are used interchangeably. It is not the number of issues that defines the distinction between diagnostic and screening tests, but the presence or absence of a known problem.

[6] Blackstone (2011) also confuses the distinction between diagnostic and screening polygraphs; first by using the less common terms "forensic" and "utility" instead of the more widely understood terms "diagnostic" and "screening," and then by attempting to portray single-issue screening exams as diagnostic exams. Blackstone further states that multi-facet exams are screening exams and later that multi-facet exams are distinct from multi-issue exams in that multi-issue exams are conducted in the absence of a known issue, indicating that multi-facet exams are a type of diagnostic exam conducted in response to a known problem. In practice, the criterion variance of the RQs of both multi-issue and multi-facet polygraphs is assumed to be independent, and both types of exams are interpreted with decision rules that reflect this assumption.

little practical use to PDD field examiners, program administrators, policy makers and consumers regarding the merits of different PDD techniques.

Present Objectives

Of primary interest to this review are those PDD techniques for which there exists evidence in support of criterion validity at levels required by the standards of practice of the APA, which effective January 1, 2012, require the use of validated techniques. Those requirements state that event-specific diagnostic examinations conducted for evidentiary purposes, for which it is expected that the results may be used as evidence in a judiciary proceeding, should be conducted using techniques that produce a criterion accuracy level of .900 or higher, excluding inconclusives, and with an inconclusive rate of .200 or lower. Diagnostic examinations conducted using the paired-testing protocol can achieve a very high accuracy rate through the combination of results from examinations conducted with techniques that produce a mean criterion accuracy level of .860 or higher, excluding inconclusives, and with inconclusive rates of .200 or lower. Examinations conducted for investigative purposes should be conducted with techniques that produce a mean criterion accuracy level of .800 or higher, excluding inconclusives, and with inconclusive rates of .200 or lower.[7] Validated techniques are further required to be supported by published and replicated scientific studies. To be generalizable, the studies should be based on samples that are representative of the general population.

In addition to specifying requirements for criterion accuracy of PDD examination techniques, the APA has adopted a standard specifying that a validated technique consists of the combination of a test question sequence or format which conforms to valid PDD testing principles, coupled with a valid method of test data analysis. The combination of these two core components is a recognition that a valid test must first obtain a suitable quantity of interpretable and meaningful (i.e., diagnostic) information, after which the information must be interpreted effectively. Neglecting either of these would result in unsatisfactory test performance. Moreover, because the results of a single un-replicated study are regarded as inconclusive in the realm of science, validated techniques are further required to be supported by at least two publications.

Selecting evidenced-based techniques minimizes exposure that would result from the use of un-standardized, un-validated, sub-optimal, or experimental methods. While the present effort was undertaken to provide useful information in this regard, readers are reminded that the report constitutes a literature review of publications available at the time the report was issued. New instrumentation, new validity research and new methods of analysis will become available in the future. In light of continuing advancements in the field, the findings in this report should be considered a reference, and not a policy of the APA.

The ethics of test administration were not addressed in this meta-analysis. Discussions of PDD test accuracy can sometimes digress into discussions of the ethics surrounding the procedures for test administration of probable-lie CQT formats, for which it is considered necessary to psychologically maneuver the examinee in order to achieve a satisfactory level of test specificity to truthfulness and to constrain false-positive errors to minimal levels. This discussion can also lead to unproductive, and indeed avoidant, deflections about increased incremental validity (i.e., "test utility") of

[7] Near the completion of this report the APA Board of Directors proposed a change to the Standards of Practice specific to screening techniques because of the paucity of available research in this area despite the importance of this application to law enforcement and national security. The proposal would permit the use of screening techniques if research indicates an accuracy significantly greater than chance, and recommends the use of a successive hurdles approach to minimize errors. Because the proposal had not been voted prior to the completion of this report, no additional analyses of screening methods using the proposed standards are included here.

decisions made by professional consumers of PDD tests.[8]

Discussions of PDD test accuracy may also encompass the ethical complications of conducting a test for which the examiner and examinee purposes may be dissimilar. For example: the examinee's desire is to generate data and test results to exonerate himself from further scrutiny, while the purpose to the examiner is often to psychologically leverage a confession of guilt or disclosure of information from deceptive examinees. In its most limited and un-scientific use the polygraph can become little more than a prop to enhance the effectiveness of an interview or interrogation, with little or no concern for the test result. It is important when reporting the criterion accuracy of PDD examinations to limit the focus to only those issues pertaining to the level of accuracy of polygraph decisions rather than merely confession rates. In other words, how effective are modern polygraph techniques at correctly classifying examinees as deceptive or truthful?

Test accuracy, in a scientific sense, means several things and can convey information about many important concerns. Foundational among these concerns is the issue of construct validity, which, in its most simplistic representation, refers to the correctness of the underlying constructs, principles, or ideas on which a test is constructed. Simply stated, construct validity refers to whether the PDD test does what it is intended to do. As a practical matter, PDD tests are often referred to as lie-detection tests, therefore, a broad formulation of the question of construct validity would involve whether a PDD test actually tests for or measures lies. Lies are amorphous and therefore cannot be measured *per se*. Deception is a temporal act, and PDD exams, like many other scientific tests, are scored numerically by measuring or observing the examinee's response to the test stimuli.[9] Deception or truthfulness is inferred statistically, by observing or measuring the responses to several iterations of the test stimuli, aggregating the responses, and then using structured decision rules to interpret the result through comparison with normative data.

Construct validity can also refer to the correctness of assumptions about the function of the structural components of the PDD test: whether the questions function as intended. Several studies have investigated the construct validity of various types of test questions. For example: overall truth questions have been shown not to function as intended, Hilliard (1979) and Abrams (1984) both providing insight into the general complications and concerns pertaining to questions of intent. Likewise, symptomatic questions, intended to test for or correct for outside issues, have been shown to not function as intended or reputed (Honts, Amato & Gordon, 2004; Krapohl & Ryan, 2001;), despite some weak evidence of support in an early study (Capps, Knill & Evans, 1993). Technical questions designed to test for, or make use of, esoteric phenomena such as a guilt-complex have been shown to not function effectively (Podlesny, Raskin & Barland, 1976). Similarly, sacrifice questions, regarding the examinee's intent to answer truthfully regarding the RQs, have been shown not to function as intended (Capps, 1991; Horvath, 1994).

[8] The term *incremental validity* is preferred to the older term *utility* because it implies an expectation for empirical evidence of increased decision accuracy or decision effectiveness on the part of consumers of PDD test results, as a result of information gained from the polygraph, and not a mere assumption that all information will prove helpful or useful.

[9] Deception during PDD exams is inferred empirically in that PDD studies have shown that deception and truth-telling can be determined with the CQT at rates that are significantly greater than chance as a function of the differential magnitude of response to relevant and comparison stimuli. Differences in response magnitude are thought to be a function of the salience of the stimuli. Persons who are being deceptive regarding the relevant stimuli are expected to show responses of generally larger magnitude to relevant than comparison stimuli, while persons who are truthful regarding the relevant stimuli are expected to show generally larger responses to comparison stimuli.

Although hypotheses are abundant, scientific studies have been unable to show evidence of construct validity for the array of technical questions, with the exception of one. The CQ is generally capable of producing larger reactions from truthful persons than RQs. While construct validity is an important concern, this meta-analysis addressed criterion validity, the ability to differentiate deception from truth-telling at a practical level, with an emphasis on the identification of PDD techniques for which there exists published and replicated evidence in support of test accuracy. It did not address questions pertaining to construct validity.

Criterion accuracy of CQT methods used in PDD examinations cannot be adequately described by a single numerical value. Instead, criterion validity for CQT PDD examinations is the result of an interaction of several dimensions of concern, including correct and false hits for deceptive decisions, and correct and false hits for truthful decisions. In field practice, test results are interpreted or expressed in terms of the presence or absence of significant reactions indicative of deception. Categorical test results for individual cases are either positive or negative,[10] and the criterion validity of a test is estimated by partitioning the results of sample cases into true positives, false positives, true negatives and false negatives. Attempts to describe criterion accuracy are further complicated by inconclusive results, for which sample results are partitioned into inconclusive results for deceptive and truthful groups. In addition to the challenges of measuring and describing estimates of PDD test accuracy, different polygraph techniques achieve different dimensional *profiles* of criterion accuracy. Some techniques may be intended to provide high test sensitivity at the expense of other dimensions of test accuracy, while other techniques may be designed to seek a balance of test sensitivity and test specificity.

This meta-analytic survey is intended to summarize the present state of existing published scientific evidence of criterion validity of PDD examination techniques, and to provide guidance regarding those techniques which can be expected to reliably provide criterion accuracy that satisfies the APA's requirements for precision at the evidentiary, paired testing, and investigative levels. Therefore, this study is limited to those techniques for which the available evidence *supports* their criterion validity, and does not include PDD techniques with un-replicated research, or techniques for which there is replicated evidence at levels that do not satisfy the requirements of the APA standards.[11]

Method

A literature survey was conducted to identify published studies that provided usable information regarding the criterion accuracy of identified PDD techniques. The results of un-replicated studies are not useful in meta-analytic research, and were therefore not included. As a practical decision this meant that at least two published studies were required for the inclusion of an examination technique in the meta-analysis. Examination techniques were retained in the meta-analysis if published and replicated studies were identified in support of the validity of a technique, and if the aggregated results of the included studies indicated a reliable and generalizable level of accuracy consistent with the requirements of the APA Standards of Practice for evidentiary testing, paired testing, or investigative testing. However, it was not a requirement that *individual* studies produce criterion accuracy at the levels specified by the APA.

[10] Although PDD test results are interpreted, in field practice, for the presence or absence of significant indicators of deception, the results of scientific tests are often discussed in value-neutral language. Positive test results are designated to signify the presence of the issue or concern that is being tested. Negative test results signify the absence of the concern or issue.

[11] Because research is ongoing in all fields of science, and because standards of practice undergo periodic review and necessary modification, readers are reminded that other PDD techniques may satisfy the present requirements of the APA standards. Field examiners, program administrators and quality assurance reviewers are advised to evaluate this information with awareness for new and emerging standards and information.

It was deemed important to be as inclusive as possible with stakeholders (i.e., school directors and developers of PDD techniques) in the selection of studies and techniques for the meta-analysis as it was anticipated that the work of many PDD field examiners and trainers may be affected by the results and recommendations of this meta-analysis. To accomplish this, a long-list of all identifiable PDD techniques was assembled and disseminated to all APA accredited polygraph schools in late March 2011 using the contact information listed at the APA website (www.polygraph.org). School directors or their representatives were invited to advise the committee of any techniques which had not yet been identified, and to provide either citations or copies of studies that could be accessed and reviewed in support of the suggested techniques.[12]

In early June 2011 a short-list was disseminated to all APA accredited polygraph schools, again using the contact information at the APA website, including both techniques and citations for which published and replicated studies were identified. Also, a list was sent describing those techniques for which the basis of publication was inadequate for inclusion in the meta-analysis. School directors or their representatives were again invited to respond and advise the committee of any techniques or any published studies that should be considered for inclusion in the meta-analysis.[13]

Study Selection

Requirements for the selection of individual studies into the meta-analysis were both qualitative and quantitative. Some studies were not designed to function as studies of criterion validity, but were intended to investigate specific research questions, such as the effects of countermeasures on PDD accuracy. Studies designed to examine causality and construct questions may not be useful for answering questions about criterion accuracy, however, studies were included if they provided sufficient information to calculate the criterion accuracy of a survey sample or normal control group that was not subject to experimental manipulation beyond truthfulness or deception.

The APA president and the committee chairperson expressed to the committee members that additional studies should be considered for inclusion in the meta-analysis if there was sufficient time to complete the review and publication process prior to the completion of the meta-analysis. Several studies were subsequently completed and submitted for peer-review and publication. Results from those studies were included once the studies were accepted for publication. All of these "in press" studies were designed as criterion accuracy studies, consistent with the requirements of the meta-analysis.

Qualitative selection requirements. Qualitative requirements for selection and inclusion in the research review were that studies selected must be published in the journal *Polygraph* or other peer-reviewed scientific publication.[14] Studies were also considered for selection if they were published by an academic degree-granting institution that was accredited by an accrediting agency

[12] One technique, developed by Lynn Marcy, was requested to be added to the list for review. However, no published studies could be located regarding that technique. At least one school representative recommended additional research for several PDD techniques.

[13] One study was suggested for inclusion in the meta-analysis at that time, the Gordon et al. (2000) field study on the IZCT, for which the 2010-2011 APA President was the developer and has a proprietary interest. However, the Gordon et al. (2000) included no reliability statistics, no statistical parameters or description of the sampling distributions of deceptive and truthful scores. The committee was advised by the primary author (Personal communication, June 10, 2011) that he had never seen the data or the cases because they belong to the intelligence service of a foreign government. It was determined that this study could not be included due to the lack of published reliability data, and inability to evaluate the study data with that from other studies. Exclusion of this study did not prevent the IZCT from being included in the meta-analysis.

[14] The journal *Polygraph* instituted expert peer review in 2003. Articles published prior to that time were subject only to editorial review. Because Polygraph is an important academic and historic resource, studies published prior to 2003 and without peer review were included in this meta-analysis if they satisfied all of the other qualitative and quantitative requirements for selection.

that is recognized by the United States (US) Department of Education or its foreign equivalent. Also considered for selection were research publications of studies funded by government agencies that underwent external peer review. Edited academic texts and their chapters were also included. Studies not subject to editorial or external review, and studies described only in self-published books were not considered.

Selection into the meta-analysis required that studies were conducted in a manner that allows for the confident generalization of the study results to field settings. These requirements included that studies be conducted using instrument recording and component sensors that reflect field practices (i.e., using two pneumograph sensors, electrodermal sensors, and cardiograph arm-cuff sensors), and a PDD technique for which there exists a published description of the examination technique, including rules and procedures for target selection, question formulation, and test presentation of the test stimuli. In addition, it was required that the method of test data analysis be consistent with field practice and supported by a published description of its features, transformations (scoring rules), decision rules and normative data[15] or cutscores. Results from studies that solely utilized automated algorithmic TDA models were not included in the meta-analysis.

Ground truth criteria must have been independent of the polygraph decision.[16] Because the decision of the original examiner could have been influenced by extra-polygraphic information, blind scores were preferred over original examiner decisions.[17] Basic demographic information was required regarding the examinees and the experience and training of the examiners.

Principal investigators for studies selected for inclusion were required to be blind to the criterion status of the cases and study participants.[18] Although the present APA research standards require that principal investigator not participate in the data collection, this requirement did not exist prior to March 2011, and therefore was not enforced in the selection of studies for the meta-analysis.

Laboratory and field studies. Field studies are important to polygraph research as these studies have the advantage of known ecological validity and are therefore assumed to have increased generalizability in this regard. However, the representativeness and generalizability of field studies are compromised, to some unknown degree, by the inherently non-random case selection process which depends on the availability of confirmation data. While field studies are highly useful for studying correlations, their usefulness is frustrated by the impossibility of controlling enough variables to determine causality and construct validity.

Laboratory studies are also important to polygraph research as these studies can

[15] Most PDD TDA methods do not use normative data, but use traditional cutscores that were determined hypothetically or arbitrarily. Although traditional cutscores have been verified as effective, the lack of normative data means that the level of statistical significance for traditional cutscores remains unknown and these cutscores might be suboptimal.

[16] Confirmation based on confession alone would exclude inconclusive and error cases, and would tend to inflate accuracy calculations. Judicial outcomes as a criterion and are also not independent if polygraph evidence was considered during the judicial proceedings, and could lead to inflated accuracy estimates. One included study (Mangan, Armitage &Adams, 2008) did not meet this requirement, and was based only on sample cases that were confirmed by confession. Not surprisingly, the study resulted in a reported 100% accuracy rate. Verschuere, Meijer, & Merckelbach (2008) argued the results of this study as a methodological artifact and therefore unreliable.

[17] As a practical matter, the inclusion of extra-polygraphic information may be advantageous if it increased the accuracy of field examiners. In the present analysis, answers to questions of criterion validity pertain only to whether or not the PDD exam data contain information that can be scored and interpreted to an accuracy conclusion.

[18] One study, (Gordon et al.,2005), was reported as non-blind in another report by Mohamed et al.(2006) who described the same comparison of fMRI results with those of the PDD examination.

more easily be constructed using random methods that reduce research and sampling bias, and thereby increase the generalizability of resulting information. Laboratory studies also provide a greater ability to control a broader range of variables, and are important to the study of causality and construct validity. However, the generalizability of laboratory studies is thought to be reduced by the fact that these studies are sometimes conducted in circumstances that imperfectly approximate the ecological conditions of field examinations.

For the purpose of reviewing the current state of validation regarding existing polygraph techniques, the ad hoc committee chose to regard field and laboratory studies with equal consideration if they satisfied the qualitative and quantitative requirements for selection into the meta-analysis. Differences between criterion accuracy of field and laboratory studies have historically been statistically insignificant, and it would be unwise to attempt to opine or hypothesize about the meaning of any differences observed in a single comparison. Research studies and reviews have shown a high level of agreement between field and laboratory studies, and the ultimate cause of any differences should be determined through the study of data. Anderson, Lindsay, and Bushman (1999) recently examined a broad range of laboratory-based psychological research on aggressive behavior and concluded the following, "correspondence between lab- and field-based effect sizes of conceptually similar independent and dependent variables was considerable. In brief, the psychological laboratory has generally produced truths, rather than trivialities." (p. 3). In the area of research directly related to the polygraph the NRC (2003) found no significant differences between the results of laboratory and field research. Similarly, Pollina et al. (2004) found no significant differences in classification accuracy of field and laboratory polygraph research.

Quantitative selection requirements. Quantitative requirements for inclusion in the meta-analysis were that studies provide sufficient information to calculate reliability and criterion validity of the technique employed. In order to calculate a complete dimensional profile of criterion accuracy, reported or available data must minimally include the sample size for truthful and deceptive groups, along with correct decisions, inconclusive rates and error rates for the truthful and deceptive cases. Several of the principal investigators were contacted for additional information regarding the sampling distributions. Quantitatively, studies were excluded only due to the lack of adequate information available to calculate the reliability and generalizability of the study results, and the dimensional profile of criterion accuracy.

Reliability. Several different statistical metrics have been used to describe the reliability of PDD examination techniques, including the Pearson product moment correlation of the numerical scores of study participants, Cohen's Kappa statistics describing the chance-corrected level of agreement between two study participants, Fleiss' Kappa statistics describing the level of agreement between three or more participants, and the uncorrected proportion of decision agreement between study participants. None of these reliability metrics was favored over the others, and studies were not excluded for missing reliability data so long as they included any statistical description of the interrater reliability of numerical scores or decisions.[19]

Sampling distributions. Mean and standard deviation parameters were required, or at least minimally able to be calculated from available data, for the deceptive and truthful distributions of scores for all selected studies. It was expected that multiple samples drawn from the same underlying population and scored with the same TDA method would produce sampling distributions that do not differ significantly. It was also expected that replication and aggregation of the results of sampling distributions would produce results that are more representative and generalizable

[19] One included technique did not meet this requirement. None of the published studies on the IZCT have included any statistical evidence of inter-rater reliability.

than any single sampling distribution, hence one of the reasons for requiring at least two studies of a technique.

Some studies were published with incomplete descriptions of the sampling distributions. Therefore, it was necessary to obtain raw scores from some of the principal investigators in order to calculate these missing statistics.[20] Most data were still available, and investigators were willing to provide additional information as requested.[21] It was expected that the sample distributions for each PDD technique would not differ at statistically significant levels if the samples were obtained from examinees who were representative of the same underlying population, the PDD technique was administered in a similar manner for each study, and the data were scored and interpreted with a similar application of the rules and procedures for test data analysis.

The absence of significant differences in sampling distributions would be interpreted as indicative that the sample distributions are representative of each other. This would increase our confidence regarding the reliability with which the samples are representative of, and generalizable to, other testing populations. Significant differences would be interpreted as indicative of samples drawn from different populations, or to differences in PDD test administration or the application of the rules and procedures for test data analysis. Confidence in the reliability and reproducibility would be reduced under these circumstances.

Criterion accuracy

Study information for criterion validity was regarded as sufficient for inclusion in the meta-analysis if a study provided enough information to calculate the complete dimensional profile of criterion accuracy, including: test sensitivity and specificity (i.e., test accuracy for deceptive and truthful groups excluding inconclusive results), inconclusive rates for the deceptive and truthful groups, positive predictive value (PPV) (i.e., the proportion of true positives to all positive results) and negative predictive value (NPV) (i.e., the proportion of true negatives to all negative results), and the proportion of correct decisions excluding inconclusive results among deceptive and truthful cases, labeled unweighted accuracy. It was not necessary for a study to provide all of these dimensional descriptions, and studies were included if it was possible for the committee to calculate these statistics from the available information. The complete dimensional profile could be calculated from a minimum of five values: decision accuracy for truthful and deceptive groups, with or without inconclusive results, along with the inconclusive rates for the deceptive and truthful groups and the number of study participants assigned to each group.

The reduction of these important criterion dimensions to a single number cannot be accomplished without neglecting a substantial portion of important information. Another complication was that some measures of accuracy are non-resistant to difference in base-rates or prior probabilities. For example: PPV and NPV are vulnerable to differences in base-rates and inconclusive rates, so these criterion dimensions are less useful for comparison of the accuracy of different techniques for which the studies may be conducted using samples with differing base-rates. The unweighted average of correct decisions, and the unweighted average of the inconclusive rates for deceptive and truthful cases was determined to provide the most usable and generalizable criterion information when comparing studies with potentially different base-rates.

[20] Statistical descriptions of sampling distributions are now commonly required for publication in scientific journals, including the journal *Polygraph*. Editors and reviewers of scientific publications did not always require this information in the past because the importance of future meta-analytic research was not always anticipated.

[21] Statistical parameters or raw scores could not be obtained for three studies, regarding two techniques, which were conducted by the US Department of Defense. Although committee members were aware that the studies had been subjected to thorough and adequate review by scientists at the Department of Defense, the absence of the data was inconvenient. Reliability statistics were provided in the Department of Defense study reports, and it was decided to retain these studies in the meta-analysis. These studies were later replicated independently.

Moderator variables

Results of included studies and PDD techniques were coded and grouped for conformance with the APA 2012 standards for evidentiary testing, paired-testing, and investigative testing. PDD techniques were also coded for whether they are intended to be interpreted with the assumption of independent or non-independent criterion variance among the RQ stimuli. No other moderators or mediators were coded for this study.

Data analysis

All samples were regarded as biased and imperfect representations of the populations from which they are drawn. This meant that some differences would be expected to be observed between the sample distribution parameters and the population distributions parameters if it were possible to obtain data from the entire population. However, representative samples would be expected to deviate from the population in ways that are not statistically significant. Samples from different studies, if all are representative of the population, would also be expected to differ in ways that are not statistically significant.

Multivariate ANOVAs were used to compare the sampling distribution parameters of each study to the sampling distributions of other replication studies for each PDD examination technique. Monte Carlo methods were used to calculate standard errors and statistical confidence intervals for the criterion accuracy profiles of each of the studies included in the meta-analysis. Data were aggregated for those techniques that are intended to be interpreted with the assumption of independence among the RQ stimuli, and those that are interpreted with the assumption of non-independence.

Results of the meta-analysis were not graded or weighted for study quality, other than the quantitative selection criteria that were previously described for inclusion in the meta-analysis. In other words, all studies were considered equally if they satisfied the qualitative requirements for inclusion in the meta-analysis. Study results were weighted by sample size and number of participating scorers, in that studies based on larger samples and studies that involved a greater number of scorers were given proportionally more weight in the meta-analysis. Because some samples were collected with multiple scorers who scored only a subset of the entire sample, weighting values are equivalent to the number of scored results for each study and each PDD technique.

Research questions addressed by this meta-analysis are: 1) which PDD examination techniques have published and replicated evidence of validity that satisfies the APA 2012 standards of practice requirement for decision accuracy and inconclusive rates, 2) what is the overall accuracy of validated PDD techniques interpreted with the assumption of independence among the RQ stimuli, 3) what is the accuracy level of PDD techniques interpreted with the assumption of non-independence, among the RQ stimuli, 4) are there significant differences or outliers among any of the validated PDD techniques, and 5) are there any outlier results that are not accounted for by the presently available evidence.

Results

Alpha was set at .05 for all statistical comparisons.

Validated PDD Techniques

Thirty-eight studies satisfied the criteria for inclusion in the meta-analysis. These studies were based on 32 different samples of confirmed cases, from which 45 different samples of scores were obtained. These studies involved 295 scorers who provided 11,737 scored results of 3,723 examinations, including 6,109 scores of 2,015 confirmed deceptive examinations, 5,628 scores of 1,708 confirmed truthful exams. Some of the samples were used in different studies, some were scored using different TDA models, and some samples were scored by more than one scorer. Appendix A shows a list of the included studies and sample sizes. Criterion accuracy reported for each of the included studies can be seen in Appendix B. Reliability statistics for the included studies are shown in Appendix C. Appendix D shows the mean and standard deviations for the sampling distributions of deceptive and truthful scores for the included studies. Fourteen PDD techniques were identified as being supported by published and replicated

studies that met the qualitative and quantitative selection requirements for this meta-analysis. Presented alphabetically, they were:

AFMGQT / Seven-position TDA

The United States Air Force Modified General Question Technique (AFMGQT)[22] is a modern variant of the family of CQT formats that have emerged as modifications of the original General Question Technique (Reid, 1947) and Zone Comparison Technique (Backster, 1963). The AFMGQT can be used effectively with two, three or four RQs. AFMGQT examinations are used in both multi-facet event-specific diagnostic contexts and multi-issue screening contexts (e.g., public safety employee selection, government security screening, post-conviction screening programs, etc.). AFMGQT exams conducted in both multi-facet and multi-issue contexts are interpreted with decision rules based on an assumption that the criterion variance of the RQs is independent. Three studies describe the criterion accuracy of the AFMGQT when scored with the seven-position TDA model.

Senter, Waller and Krapohl (2008), using blind seven-position numerical scores of 33 programmed deceptive and 36 programmed truthful examinees who were tested using the AFMGQT following their participation in a mock roadside bombing scenario, reported an unweighted decision accuracy of .849, with an inconclusive rate of .015.

Nelson and Handler (In press) used Monte Carlo methods to study the criterion accuracy of seven-position numerical scores of AFMGQT exams with two, three and four RQs. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy was reported as .814, along with an unweighted inconclusive rate of .280.

Nelson, Handler, Morgan and O'Burke (In press) obtained blind numerical scores from three experienced examiners employed by the Iraqi government who used the seven-position TDA model to evaluate a confirmed case sample of AFMGQT (N = 22) exams that were selected from the U.S. Department of Defense confirmed case archive. Eleven of the cases were confirmed as deceptive; the remaining 11 were confirmed as truthful. A total of 66 examination scores were obtained, and unweighted decision accuracy was reported as .761, along with an unweighted inconclusive rate of .242.

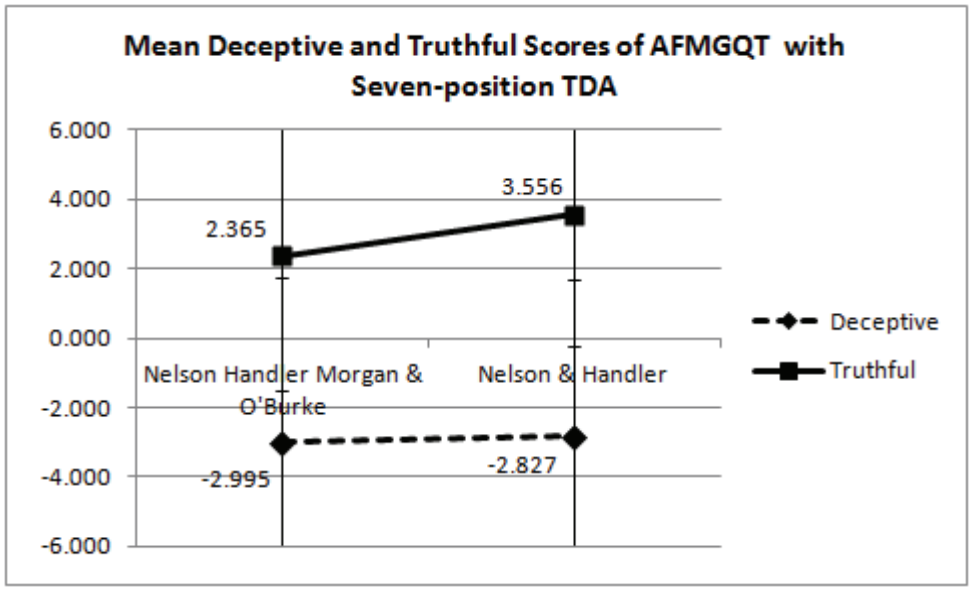
Figure 1 shows a mean and standard deviation plot of the subtotal scores[23] of the sampling distributions of the three AFMGQT seven-position studies. No mean and standard deviation data were available for the AFMGQT study completed by Senter, Waller and Krapohl (2008). A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [$F(1,68) = 0.263$, ($p = .610$)], nor was the main effect for sampling distribution [$F(1,68) = 0.013$, ($p = .910$)].

The combined decision accuracy level of these AFMGQT seven-position TDA studies, weighted for sample size and number of scorers, was .822 with a combined inconclusive rate of .191. Reliability for seven-position scores of AFMGQT exams was reported by Senter, Waller and Krapohl (2008) as kappa statistic of .750. The proportion of overall decision agreement, excluding inconclusive results, for all studies was .965.

[22] Two versions exist for the AFMGQT: version 1 and version 2. Differences between these two techniques are based on unstudied assumptions regarding test structure, and the effect of these differences has not been thoroughly studied. There is no compelling hypothesis suggesting the performance of one version would be different or superior to another. Evidence available at the present time suggests that both versions perform adequately and no significant differences have been identified. Therefore, these results are suggested as generalizable to versions 1 and 2 of the AFMGQT, and both of which are represented in the included studies.

[23] Subtotal scores were used in this analysis because the AFMGQT is scored with decision rules using only subtotal scores which assume criterion independence among the RQs.

Figure 1. Mean deceptive and truthful subtotal scores for AFMGQT seven-position studies.



AFMGQT / ESS

Three studies describe the criterion accuracy of AFMGQT exams when scored using the ESS.

Nelson and Blalock (In press) transformed seven-position AFMGQT scores from the Senter, Waller and Krapohl (2008) laboratory study to ESS scores, including 33 results for confirmed deceptive cases and 36 results for confirmed truthful cases. Unweighted decision accuracy was .839, with an inconclusive rate of .152.

Nelson, Blalock and Handler (2011) obtained blind ESS scores from two inexperienced examiners and one experienced examiner who used the ESS to evaluate a sample of confirmed AFMGQT exams (N = 22), including 11 exams that were confirmed as deceptive and 11 exams that were confirmed as truthful. A total of 66 examination scores were obtained, and unweighted decision accuracy was .883, with an inconclusive rate of .183.

Nelson, Handler and Senter (In press) used Monte Carlo methods to study the criterion accuracy of ESS scores of AFMGQT exams with two, three and four RQs. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy was reported as .876, with an unweighted inconclusive rate of .178.

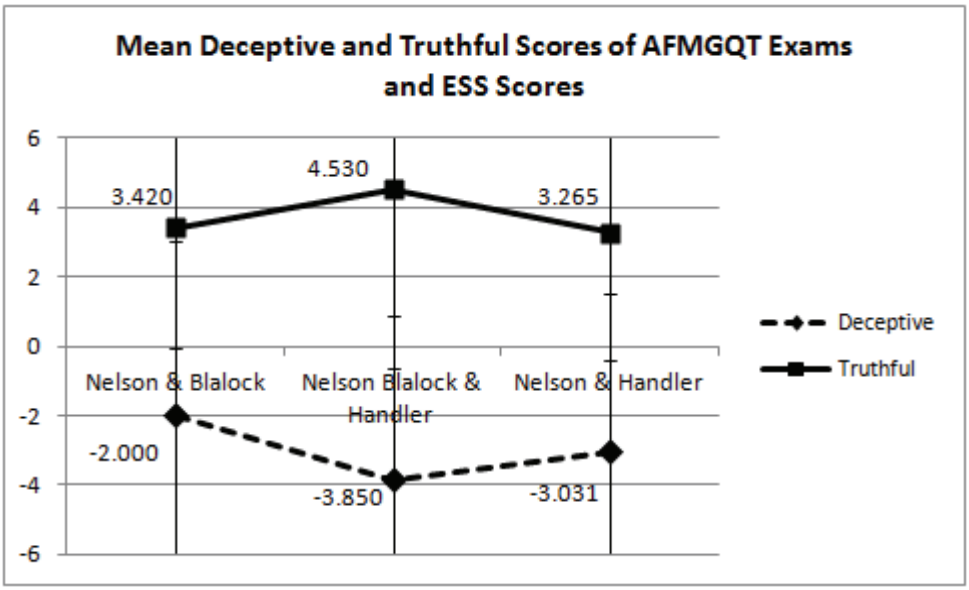
Figure 2 shows a mean and standard deviation plot of the scores of the sampling distributions of the three AFMGQT ESS studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,123) = 2.467, (p = .119)], nor was the main effect for sampling distribution [F (1,123) = 0.009, (p = .925)].

The combined decision accuracy level of these AFMGQT ESS studies, weighted for sample size and number of scorers, was .875 with a combined inconclusive rate of .170. Reliability for ESS scores of AFMGQT exams, reported by Nelson, Blalock and Handler (2011) as the bootstrap mean of pair-wise correlation coefficients, was .930.

Backster You-Phase

The Backster You-Phase technique is an event-specific diagnostic technique, based on Backster's Zone Comparison concept. This technique is scored using TDA rules developed by Cleve Backster and taught almost exclusively at the Backster School of Lie

Figure 2. Mean deceptive and truthful subtotal scores for AFMGQT ESS studies.



Detection (2011). Both generic ZCT (Department of Defense, 2006; Honts, Raskin & Kircher, 1987) and boutique ZCT variants (Gordon et al., 2000; Matte & Reuss, 1989) have been developed from the Backster You-Phase technique. Scores from two recent studies were aggregated to calculate the criterion accuracy profile of Backster You-Phase examinations.

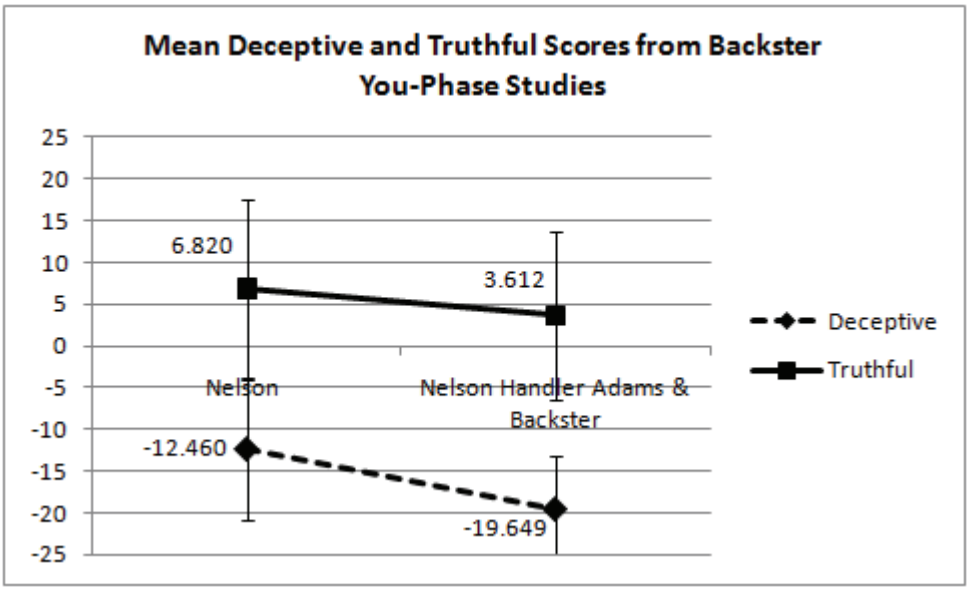
Nelson (In press) used seeding parameters calculated from the composite distributions of two non-selected studies[24] to seed a Monte Carlo space of 100 simulated Backster You-Phase exams. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy for the Nelson (In press) Monte Carlo study of You-Phase exams was .927, with an inconclusive rate of .321.

Nelson, Handler, Adams and Backster (In press) surveyed the results of seven examiners who provided 144 blind numerical scores for a sample (N = 22) of 11 confirmed deceptive and 11 confirmed truthful You-Phase examinations. These seven scorers had a range of experience from less than one year to more than 30 years. Results from the blind-scores survey produced an unweighted decision accuracy rate of .825 with an inconclusive rate of .117. Figure 3 shows a mean and standard deviation plot for the scores of deceptive and truthful cases. A two-way unbalanced[25] ANOVA showed that the interaction of sampling distribution and

[24] Honts, Hodes and Raskin (1985) used the Backster You-Phase technique in a countermeasure study for which the traditional arm cuff was replaced with an alternative cardio sensor. Meiron, Krapohl and Ashkenazi (2008) used the Backster You-Phase technique in a study of the Either-Or Rule, using a highly selected sample from which the results of problematic examinations were not included, resulting in a sample that is assumed to be systematically devoid of error variance. Although neither of these studies was usable alone, the parameters that describe the composite distributions of deceptive and truthful scores was assumed to be a more generalizable representation of error or uncontrolled variance along with diagnostic variance for scores from the Backster You-Phase exams.

[25] Unbalanced ANOVAs, using the harmonic mean of the sample Ns, was used throughout this study when necessitated by differences in sample sizes. As a result, the total degrees of freedom in the ANOVA summary may not reflect the sum of all samples in the same way as a balanced ANOVA design. Unbalanced ANOVA designs can be expected to provide slightly less statistical power than balanced ANOVA designs.

Figure 3. Mean and standard deviation plot for Backster numerical scores of confirmed You-Phase exams.



criterion status was not significant F (1,68) = 0.869, (p = .355)], nor was the main effect for sampling distribution [F (1,68) = 0.164, (p = .686)].

The combined results of the two published studies of Backster You-Phase exams, weighted for the sample size and number of scorers, produced a decision accuracy rate of .863 and an inconclusive rate of .196. Reliability of Backster numerical scores of You-Phase exams, reported by Nelson et al. (In press) as a bootstrap mean of pairwise correlation coefficients, was .567.

Concealed Information Test (CIT)

The CIT, also known as a Guilty Knowledge Test (GKT; Lykken, 1959) and related to the Peak of Tension (POT) technique (Ansley, 1992), is an event-specific diagnostic technique that can be used to investigate whether an examinee possesses knowledge or information that would be available only to investigators and a guilty or involved suspect. Like the CQT, the CIT/GKT is based on the principle of salience, including emotion, cognition and behavioral conditioning as the psychological basis of physiological response (Senter, Weatherman, Krapohl & Horvath, 2010). Also, the CIT/GKT is conducted using instrumentation that is similar to CQT methods, including electrodermal sensors, and may include cardiograph and pneumograph sensors. However, the CIT/GKT does not include comparison questions and is not a CQT method. Therefore, the CIT/GKT has not been subject to the same ethical concerns as probable-lie CQT methods regarding manipulating the examinee as a feature of test administration.[26] Hypothetical explanations for psychophysiological mechanisms underlying the CIT/GKT have not been limited to emotion and fear as the sole basis of response. Also, the CIT/GKT has remained free from scientific criticisms involving the role of

[26] Similarly, directed-lie methods have been shown to work as well as probable-lie methods, and are less subject to the ethical complications regarding manipulating the test subject (Bell, Kircher & Bernhardt, 2008; Blalock, Nelson, Handler & Shaw, 2011; Honts & Reavy, 2009).

examinee confessions as both an objective of the test and as a form of verification of the test results.[27]

MacLaren (2001) published the results of a meta-analysis of 50 samples in 22 studies involving the CIT/GKT. Thirty-nine of those samples involved 1,070 examinees of which 666 had engaged in a behavioral act for which they had concealed information along with 404 examinees who had no involvement or knowledge of the behavioral details. Eleven samples involved 177 examinees who possessed concealed knowledge of, but did not actually engage in, the behavioral act under investigation.

Using the scoring protocol and test methodology described by Lykken (1959), results reported by MacLaren produced a test sensitivity level of .815 for behaviorally involved examinees, along with a test specificity level of .832 for un-involved examinees who had no concealed knowledge. Unweighted decision accuracy was .823. However, when results included those examinees who possessed concealed information but were behaviorally un-involved, the test sensitivity level was .759, and the unweighted decision accuracy rate was .795.

Directed-lie Screening Test / Seven-position TDA

The Directed-lie Screening Test (DLST) is based on the Test for Espionage and Sabotage (TES) that was developed by the U.S. Department of Defense.[28] As indicated by its name, the DLST technique uses directed-lie CQs and is designed for screening exams that are conducted in the absence of any known problem. Screening tests are often constructed and interpreted with multiple issues for which the criterion variance of the CQs is assumed to be independent. Two studies, involving the DLST/TES and seven-position TDA, have been published by the U.S. Department of Defense. Research reports were available, though insufficient to meet the study inclusion criteria, due to the absence of sampling standard deviations in the study reports. Also, data for the Department of Defense studies were not available for this committee to review. However, two DLST/TES replication studies were recently completed and the data from those studies was included in this meta-analysis.

The first published study of the DLST/TES (Research Division Staff, 1995a) was a laboratory experiment involving a mock espionage scenario. Three experienced federally trained examiners scored 94 examinations, involving 26 programmed deceptive examinees and 68 programmed truthful examinees. Results from this study produced an unweighted decision accuracy level of .788, with an inconclusive rate of .155.[29]

In the second published study of the DLST/TES (Research Division Staff, 1995b) 10 experienced federally trained examiners provided scores for 30 deceptive and 55 truthful laboratory examinations involving a mock espionage scenario. Results from this study produced an unweighted decision

[27] Overemphasis on confession confirmation and non-independent criterion has led to criticisms of contamination and overestimation of CQT accuracy.

[28] The name Directed Lie Screening Test is used in contexts for which the investigation targets differ from espionage and sabotage.

[29] Although all information was included in the published reports, some false-positive errors and inconclusive results were excluded from previously reported statistics for this study. False-positive results were removed from the reported results when the examinee made post-test admissions that would have been viewed as substantive in field settings, causing the results to be viewed as not erroneous. Inconclusive results were removed from the previously reported results because DLST/TES procedures which require the immediate re-examination of inconclusive results. It was not possible for the blind scorers to repeat inconclusive examinations, and the investigators elected to describe DLST/TES accuracy without inconclusive results that could not be subject to re-examination. All false-positive and inconclusive results were included in the present results because this was considered a more conservative estimate of criterion accuracy for this technique.

accuracy level of .888, with an inconclusive rate of .009.[30]

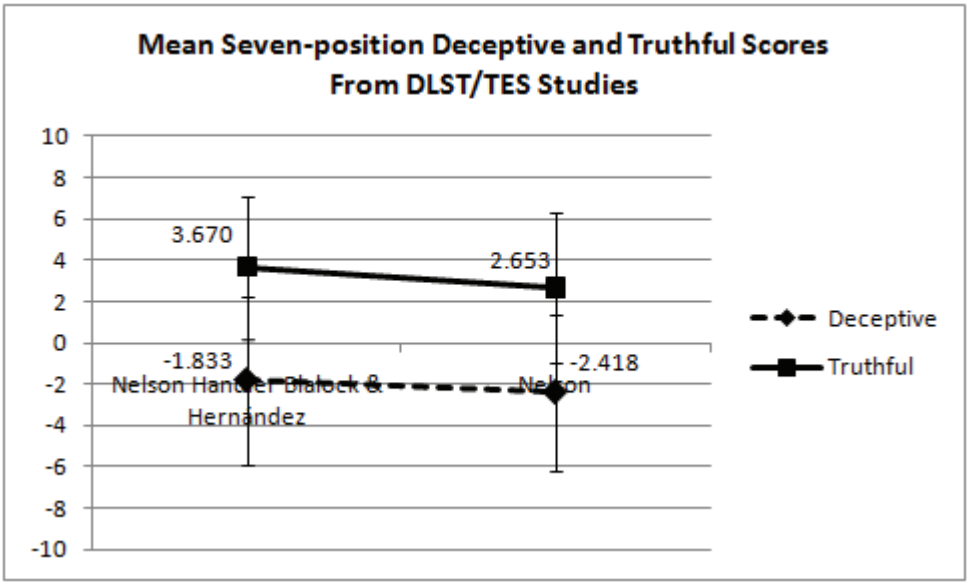
Nelson (In press) used Monte Carlo methods to study the criterion accuracy of the DLST/TES, and reported an unweighted decision accuracy level of .874, with an inconclusive rate of .096. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases.

Nelson, Handler, Blalock and Hernández (In press) studied the DLST/TES in a mock espionage scenario with examiner trainees from the Iraqi government. Two scorers, including one experienced federally trained examiner who is also an APA primary instructor, and one international examiner who is an APA member from México, provided blind seven-position scores for 25 programmed deceptive and 24 programmed truthful examinees. Fifty scores were obtained for programmed deceptive examinees and 48 scores were obtained for programmed innocent examinees. The unweighted decision accuracy level was .831 with an inconclusive rate of .092.

Figure 4 shows a mean and standard deviation plot for the seven-position deceptive and truthful scores of DLST/TES exams. No mean and standard deviation data were available for the TES studies completed by the US Department of Defense. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,128) = 0.109, (p = .742)], nor was the main effect for sampling distribution [F (1,128) = 0.023, (p = .880)].

The combined results of the four published studies of the DLST/TES with seven-position TDA, weighted for the sample size and the number of scorers, produced a decision accuracy rate of .844, and an inconclusive rate of .088. Reliability of DLST/TES studies completed by the U.S. Department of Defense, calculated via Cohen's Kappa, was reported as .760. The average proportion of pairwise decision agreement for all DLST/TES seven-position studies was .806.

Figure 4. Mean and standard deviation plot for seven-position scores of DLST/TES exams.



[30] Two-false positive errors were removed from the previously reported accuracy estimations due to post-test admissions that were deemed by the principal investigator to have been likely to be considered substantive and not erroneous in field settings. One inconclusive result was removed from the previous accuracy estimations. Calculations in this report include all error and inconclusive results.

Directed-lie Screening Test / ESS

Four studies describe the criterion accuracy of the DLST/TES when scored with the ESS.

Nelson and Handler (In press) used Monte Carlo methods to study the criterion accuracy of DLST/TES exams scored with the ESS. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy was reported as .831 with an inconclusive rate of .104.

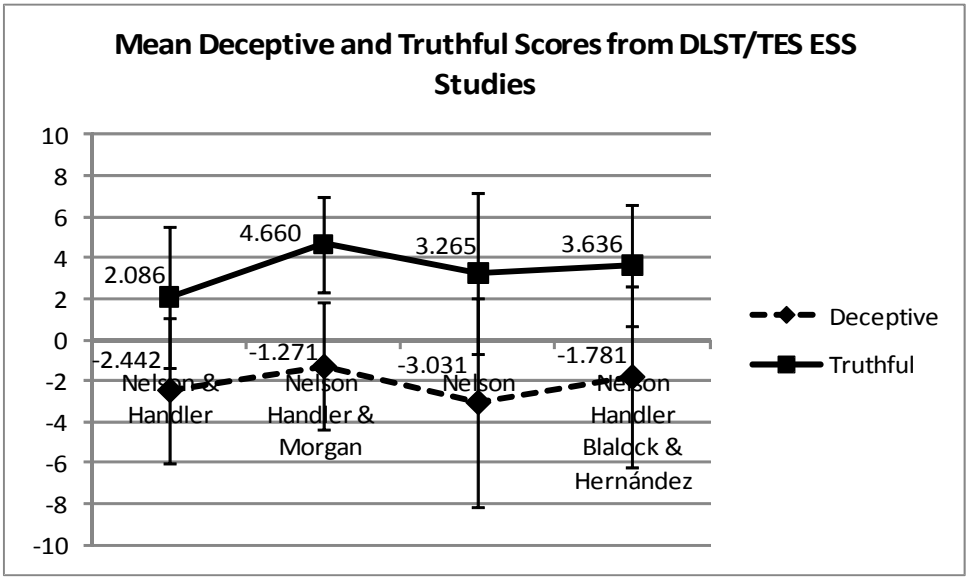
Nelson, Handler and Morgan (In press) used a mock espionage scenario to study the criterion accuracy of ESS scores of DLST/TES exams conducted by seven inexperienced examiner trainees on eight non-naive examinees who were fully conversant with the operation and scoring of PDD examinations including the DLST/TES. Blind scores were obtained for 25 programmed deceptive exams and 24 programmed truthful exams. Unweighted decision accuracy was .854 with an inconclusive rate of .088.

Nelson (In press), using a different Monte Carlo method, compared ESS scores of DLST/TES examinations to seven-position and three-position scores. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy of ESS scores was reported as .871, with an inconclusive rate of .048.

Nelson, Handler, Blalock and Hernández (In press) reported the results of blind ESS scores of DLST/TES examinations. Seven-position scores from two scorers, including one experienced federally trained examiner who is also an APA primary instructor, and one international examiner who is an APA member from México, were transformed to ESS scores, including 50 blind scores for 25 programmed guilty and 48 blind scores for 24 programmed innocent examinees. Unweighted decision accuracy was .859, and the unweighted inconclusive rate was .123.

Figure 5 shows a mean and standard deviation plot of the scores from the four DLST/TES ESS studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,289) = 2.396, (p = .123)], nor was the main effect for sampling distribution [F (3,289) = 0.156, (p = .925)].

Figure 5. Mean and standard deviation plot for ESS scores of DLST/TES exams.



The combined decision accuracy level of these studies, weighted for sample size and number of scorers, was .858 with a combined inconclusive rate of .090. Reliability of ESS scores of DLST/TES examinations, reported as the bootstrap mean of the proportion of the pairwise decision agreement, excluding inconclusive results, was .911 for the Nelson, Handler and Morgan (In press) study, and .769 for the Nelson, Handler, Blalock and Hernández (In press) study. The average rate of pairwise decision agreement for these studies was .840.

Federal You-Phase / Seven-position TDA

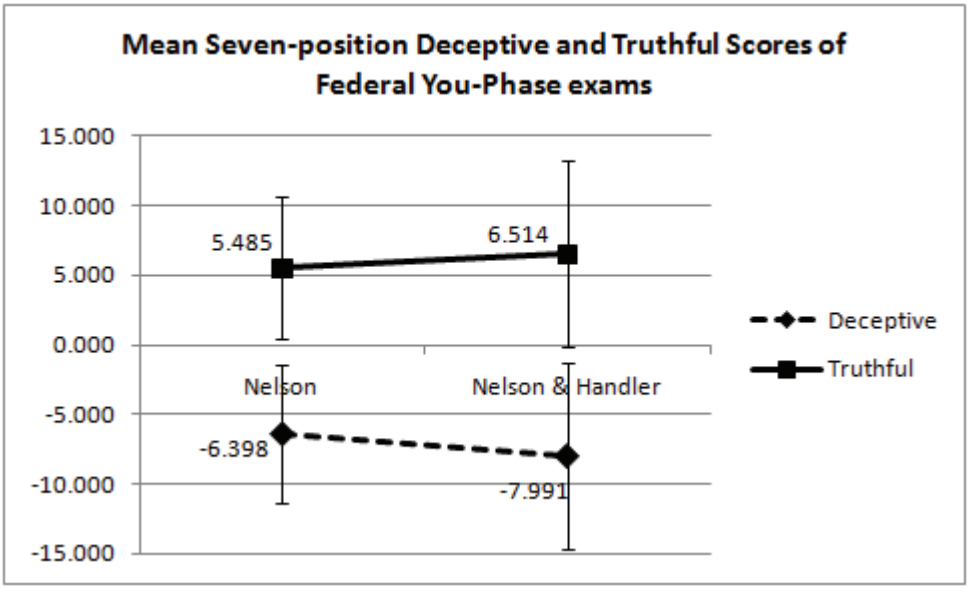
The Federal You-Phase technique[31] is an event-specific diagnostic technique constructed with two RQs. Two studies describe the criterion accuracy of the Federal You-Phase technique when scored using the seven-position TDA model.

Nelson (2011) used Monte Carlo methods to calculate the criterion accuracy of Federal You-Phase exams when scored with the seven-position TDA model. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy was .870, and the unweighted inconclusive rate was .301.

Nelson, Handler, Blalock and Cushman (In press) obtained blind scores from eight inexperienced and two experienced scorers who used the seven-position TDA model to provide 220 scores for a sample of Federal You-Phase examinations (N = 22) selected from the confirmed case archive of the U.S. Department of Defense. Eleven of the cases were confirmed as deceptive, and 11 were confirmed as truthful. Unweighted decision accuracy was .885, with an unweighted inconclusive rate of .108.

Figure 6 shows a mean and standard deviation plot of the scores from the Federal You-Phase seven-position criterion studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,68) = 0.628, (p = .431)], nor was the main effect for sampling distribution [F (1,68) = 0.001, (p = .977)].

Figure 6. Mean and standard deviation plot for seven-position scores of Federal You-Phase exams.



[31] Sometimes referred to as the Bi-Zone technique.

The combined decision accuracy level of these studies of Federal You-Phase exams scored with seven-position scoring, weighted for sample size and number of scorers, was .883 with a combined inconclusive rate of .168. Reliability of seven-position scores of Federal You-Phase exams, reported as the bootstrap mean of the proportion of the pairwise decision agreement, excluding inconclusive results, was .852.

Federal You-Phase/ESS

Two studies describe the criterion accuracy of Federal You-Phase examinations when scored with the ESS.

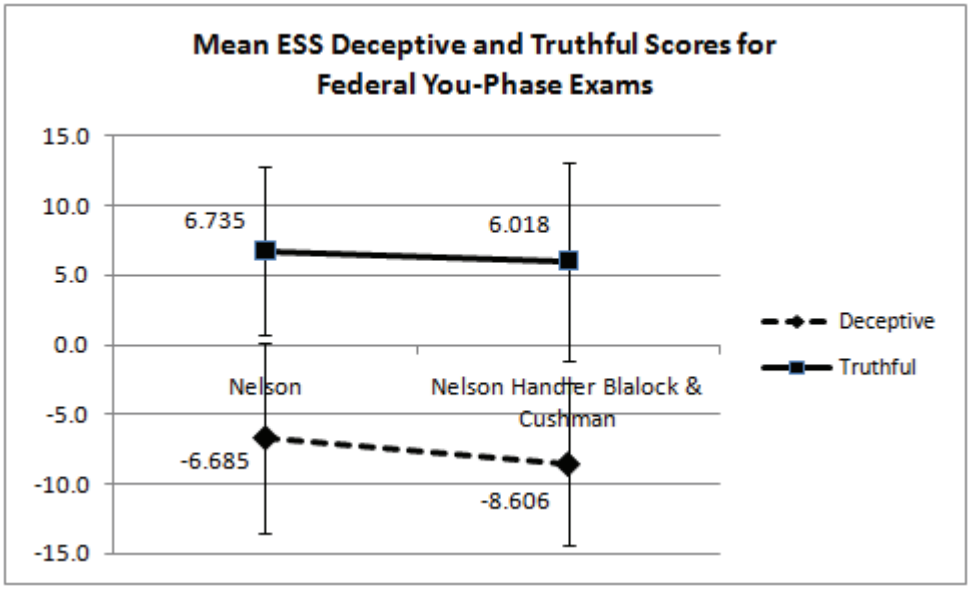
Nelson (2011) used Monte Carlo methods to calculate the criterion accuracy of Federal You-Phase exams when scored with the ESS. The Monte Carlo space consisted of 50 criterion truthful cases and 50 criterion deceptive cases. Unweighted decision accuracy was .897 and the unweighted inconclusive rate was .096.

Nelson, Handler, Blalock and Cushman (In press) reported the criterion accuracy of Federal You-Phase exams using ESS scores that were obtained by transforming 220 blind seven-position scores obtained from eight inexperienced and two experienced scorers who evaluated a sample of Federal You-Phase examinations (N = 22) selected from the confirmed case archive of the U.S. Department of Defense. Eleven of the cases were confirmed as deceptive, and 11 were confirmed as truthful. Unweighted decision accuracy was .906, and the unweighted inconclusive rate was .235.

Figure 7 shows a mean and standard deviation plot of the ESS scores of the Federal You-Phase studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,68) = 0.155, (p = .695)], nor was the main effect for sampling distribution [F (1,68) = 0.021, (p = .886)].

The combined decision accuracy level of these studies of Federal You-Phase exams, weighted for sample size and number of scorers, was .904 with a combined inconclusive rate of .192, when scored with the ESS TDA method. Reliability, reported as the bootstrap proportion of pair-wise decision agreement excluding inconclusive results, was .897.

Figure 7. Mean deceptive and truthful scores for Federal You-Phase / ESS studies.



Federal ZCT / Seven-position TDA

The Federal ZCT is an event-specific diagnostic technique constructed with three RQs. Three studies describe the criterion accuracy of the Federal ZCT when scored using the seven-position TDA model.

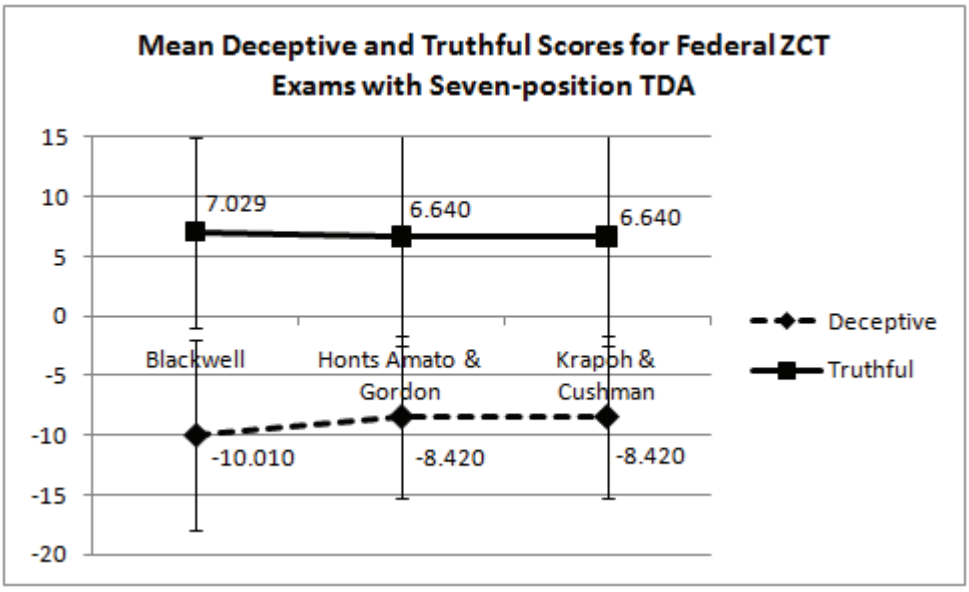
Blackwell (1998) described the criterion accuracy of 100 confirmed Federal ZCT exams, of which 65 examinations were confirmed as deceptive while 35 exams were confirmed as truthful. A total of 195 scored results were obtained for criterion deceptive exams, and 105 scored results were obtained for criterion truthful exams. Three experienced federally trained examiners scored all of the cases using the seven-position TDA method.[32] Unweighted decision accuracy was .793, and the unweighted inconclusive rate was .159.

Krapohl and Cushman (2006), reported the criterion accuracy of Federal ZCT exams scored by a cohort of 10 experienced examiners, each of whom who scored 50 confirmed deceptive field examinations and 50 confirmed truthful exams selected from the U.S. Department of Defense confirmed case archive. A total of 1,000 scored results were obtained. Unweighted decision accuracy was .852, and the unweighted inconclusive rate was .198.

Honts, Amato and Gordon (2004), reported the criterion accuracy of Federal ZCT exams that were evaluated by three scorers who used the federal seven-position TDA model. A total of 72 scores were obtained for 24 criterion deceptive exams, and 72 scores were obtained for 24 criterion truthful exams. Unweighted decision accuracy was .958, and the unweighted inconclusive rate was .042.

Figure 8 shows a mean and standard deviation plot of the seven-position scores of the Federal You-Phase studies. A two-way ANOVA showed that the interaction of sampling distribution was not significant, [F (1,204) = 0.706, (p = .402)], nor was the main effect for sampling distribution [F (1,204) = 0.004, (p = .951)].

Figure 8. Mean deceptive and truthful scores for Federal ZCT exams with seven-position TDA.



[32] The older, pre-2006, Federal TDA model employed more features than the presently used evidence-based Federal TDA model. However, Kircher et al. (2005) reported that experienced examiners tend to violate rules that do not work and tend to emphasize procedures that do work. Therefore it is possible that the scores of these examiners reflect current training and field practices more closely than might be initially assumed.

The combined decision accuracy level of these seven-position TDA studies of Federal ZCT exams, weighted for sample size and number of scorers, was .860 with a combined inconclusive rate of .171. Reliability of seven-position scores of Federal ZCT exams, reported as the Fleiss' kappa statistic for categorical decisions of multiple raters was .570, and the pairwise proportion of decision agreement excluding inconclusive results was .800.

Federal ZCT / Seven-position TDA with evidentiary decision rules

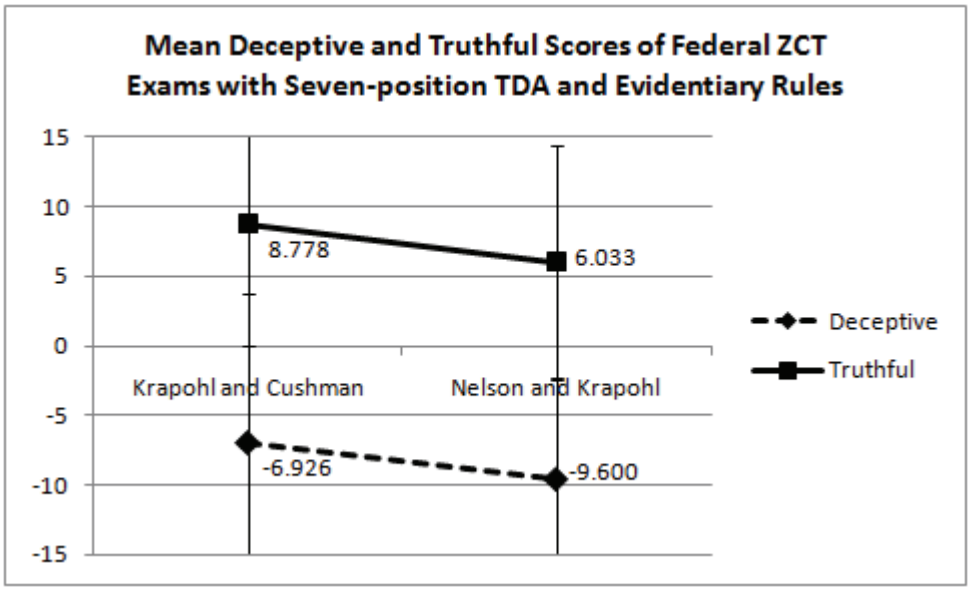
Two studies describe the criterion accuracy of the Federal You-Phase technique when scored using the seven-position TDA model and evidentiary decision rules.

Krapohl and Cushman (2006) described the criterion accuracy of 100 Federal ZCT exams that were scored by 10 experienced examiners using the seven-position TDA model and evidentiary decision rules.[33] A total of 1,000 scored results were obtained. Examinations were selected from the U.S. Department of Defense confirmed case archive. Fifty of the examinations were confirmed as deceptive, and 50 of the exams were confirmed as truthful. Unweighted decision accuracy was .872, and the unweighted inconclusive rate was .073.

Nelson and Krapohl (2011) reported the criterion accuracy of 60 Federal ZCT exams that were evaluated by six experienced federally trained scorers. Thirty of the examinations were confirmed as deceptive and 30 exams were confirmed as truthful. Each scorer evaluated a random subset of 10 exams. Results were evaluated using the Federal seven-position TDA model and evidentiary decision rules. Unweighted decision accuracy was .870, and the unweighted inconclusive rate was .100.

Figure 9 shows a mean and standard deviation plot of the seven-position scores of the Federal ZCT studies, using evidentiary rules. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (1,146) = 0.001, (p = .981)], nor was the main effect for sampling distribution [F (1,146) = 0.046, (p = .830)].

Figure 9. Mean deceptive and truthful scores for Federal ZCT exams with seven-position TDA and evidentiary decision rules.



[33] Krapohl (2005) and Krapohl and Cushman (2006) showed that evidentiary decision rules can substantially reduce inconclusive rates without a corresponding loss of overall decision accuracy.

The combined decision accuracy level of these seven-position TDA studies of Federal ZCT exams with the seven-position TDA method and evidentiary decision rules, weighted for sample size and number of scorers, was .872 with a combined inconclusive rate of .075. Reliability, calculated as the bootstrap average of pairwise decision agreement excluding inconclusive results, was .870.

Integrated Zone Comparison Technique

The Integrated Zone Comparison Technique (IZCT) (Gordon et al., 2000) is a proprietary event-specific diagnostic technique scored with the Horizontal Scoring System (Gordon, 1999).[34] Two studies describe the criterion accuracy of this technique.

Gordon et al. (2005; also described in Mohamed et al., 2006) reported the results of a pilot study involving six guilty and five innocent subjects[35] who participated in a laboratory scenario involving a mock shooting incident. Decision accuracy was reported as 1.000, with an unweighted inconclusive rate of .100.

Shurani and Chaves (2010) reported the results of a survey of 84 field examinations conducted with the IZCT, including 44 scores for confirmed deceptive examinees and 40 scores for confirmed truthful examinees. All examinations were reportedly verified by confessions, with extrapolygraphic evidence extant for some exams. Unweighted decision accuracy was .988, with an unweighted inconclusive rate of .061. No reliability statistics were reported for this study, and the committee was unable to calculate interrater reliability from the available data.

Shurani (2011) reported the results of a field study involving three examiners from Costa Rica who used the IZCT along with an additional experimental technique. The sample consisted of 73 cases for which all possible suspects were tested. Forty-eight cases were confirmed, resulting in N = 188 examinations, that were conducted using the IZCT with three and four RQs.[36] Two inconclusive results were removed from the reported results. No information was reported regarding the number of exams conducted with three or four RQs. However, data were provided to the committee for 84 examinations reportedly conducted using three RQs, including scores for 36 deceptive cases and 48 truthful cases. Scores for the remaining 104 exams were not made available. No sampling mean or standard deviations were reported, and the committee was unable to compare the means of the sample data provided to the committee with any published information. Results of this study were reported with perfect accuracy and zero inconclusive findings. No reliability statistics were reported for this study, and the committee was unable to calculate interrater reliability from the available data.

Figure 10 shows a mean and standard deviation plot of the sampling distributions of the IZCT studies. A two-way ANOVA showed a significant interaction between the sampling distribution and case status [$F(1,173) = 533.771$, $(p < .001)$]. Post-hoc one-way ANOVAs showed that the sampling differences for deceptive cases was not significant. However, the difference in truthful scores for the three samples was significant [$F(2,33) = 21.402$, $(p = .014)$]. Truthful scores were significantly greater for the Shurani and Chaves (2010) and Shurani (2011) studies compared to the Gordon et al. (2005) study.

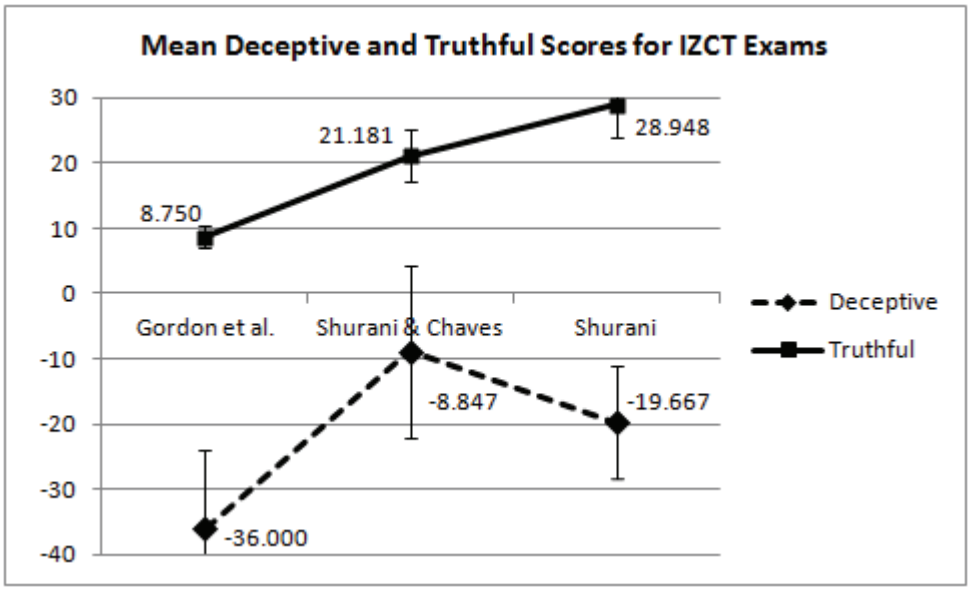
[34] A rank order scoring system based on unique developer-devised measurement features.

[35] The original pilot study design included six innocent subjects, however one truthful subject made a false-confession to the examiner (Gordon, personal communication July 6, 2011) who was also the primary author of the Gordon et al. (2005) study and developer of the IZCT. Inclusion of the false-positive (false-confession) error case would have resulted in less than perfect accuracy.

[36] No published description exists for the use of the IZCT with four RQs. Because the IZCT is scored using a rank order paradigm, inclusion of additional RQs without the inclusion of an equivalent number of additional CQs can be expected to differentially affect the rank-sum scores of relevant and CQs. No published studies described or investigated these statistical complexities.

It was later learned that the Gordon et al. (2005) study was conducted using single issue IZCT exams while the Shurani and Chaves (2010) sample cases were conducted using multi-facet IZCT exams. It is not clear whether this difference accounts for the significant interaction and differences observed in these sampling distributions.[37] A two-way ANOVA comparison, scores x sampling distribution, of the sampling distributions from the Shurani and Chavez (2010) and Shurani (2011) samples revealed a significant interaction [F (1,164) = 43.140, (p < .001)], suggesting that the scores of deceptive and truthful cases were expressed or interpreted differently in the Shurani and Chavez (2010) and Shurani (2011) study samples. One-way differences were not significant. Scores for the Shurani (2011) study were further from zero than the scores for the Shurani and Chaves (2011) study for both deceptive and truthful cases.

Figure 10. Mean deceptive and truthful scores for IZCT samples.



The combined decision accuracy level of these IZCT studies, weighted for sample size and number of scorers, was .994 with a combined inconclusive rate of .033.

No reliability statistics were reported for any of the IZCT studies, and the committee was unable to calculate interrater reliability from the available data.

Matte Quadri-track Zone Comparison Technique

The Matte Quadri-track Zone Comparison Technique (MQTZCT) (Matte & Reuss, 1989) is a proprietary event-specific, single-issue diagnostic technique scored using a modification of the Backster numerical system. Three studies describe the criterion accuracy of the MQTZCT.

[37] In a practical sense, differences in assumptions about independence and non-independence among the test questions will result in the use of different decision rules, and these differences may have had a biasing effect on case confirmation and sample selection for these field studies. Rank order scores for all RQs are always relative to all other relevant and comparison test stimuli. Rank order scores are therefore inherently non-independent, and the mathematical justification for the application of a rank-order scoring model to multi-facet exams, for which the decision rules are based on the assumption of independence, is not clear. This non-trivial statistical and decision theoretical complication has not been adequately discussed or studied.

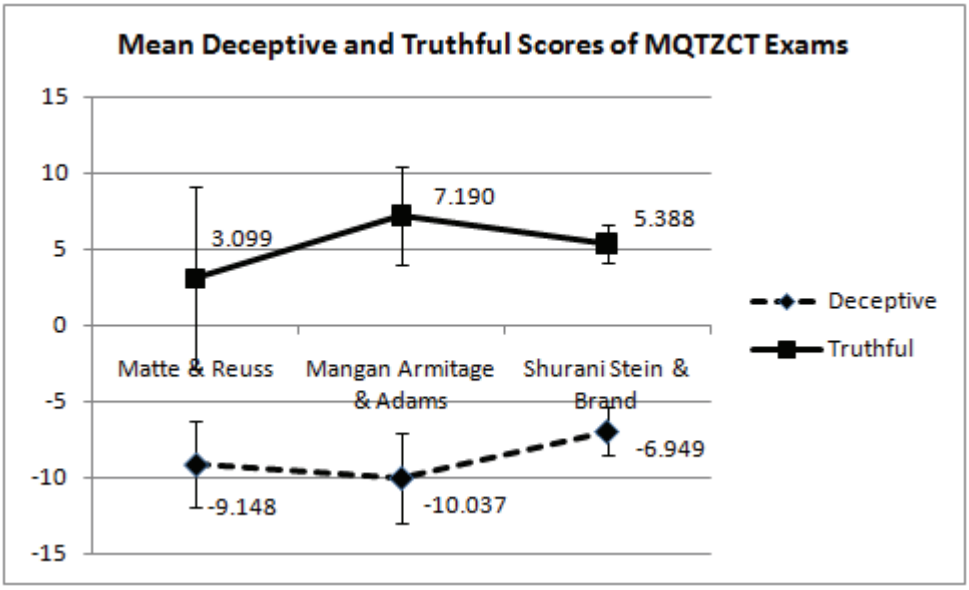
Matte and Reuss (1989) reported the results of 64 deceptive and 58 truthful cases that were confirmed through combinations of confession and other evidence. Unweighted decision accuracy was reported as a perfect 1.000, with an unweighted inconclusive rate of .059.

Mangan, Armitage and Adams (2008) reported the criterion accuracy of a survey of 91 deceptive cases and 45 truthful cases that were confirmed via examinee confession. Decision accuracy was again reported as a perfect 1.000, with an unweighted inconclusive rate of .011.

Shurani, Stein and Brand (2009) reported the criterion accuracy of a survey of 28 deceptive and 29 truthful cases that were confirmed by confession along with additional evidence for some cases. Decision accuracy was reported as .964, with zero inconclusive results.

Figure 11 shows a mean and standard deviation plot of the subtotal scores[38] of the sampling distributions of the three MQTZCT studies.[39] A two-way ANOVA revealed a significant interaction between the sampling distribution and case status [F (1,261) = 361.605, (p < .001)]. Although the different studies appeared to handle deceptive and truthful cases with different effectiveness, post-hoc one-way ANOVAs showed that the differences in scores were not significant for deceptive cases [F (2,141) = 0.389, (p = 0.678) or for truthful cases [F (2,122) = 0.264, (p < .768)].

Figure 11. Mean deceptive and truthful per-chart scores for MQTZCT samples.



[38] Scores for MQTZCT exams are reported as the subtotal per chart, obtained by summing all numerical scores within each chart. Subtotals described elsewhere in this report involve the between-chart RQ subtotals, obtained by summing the numerical scores for each RQ for all charts.

[39] Data initially provided to the ad hoc committee for the Mangan, Armitage and Adams (2008) and Shurani and Chaves (2009) studies included only those scores for which the scorers achieved the correct result, and did not include scores for inconclusive or erroneous results. Missing scores were later provided to the committee for both the Mangan, Armitage and Adams (2008) and Shurani, Stein and Brand (2009) studies. However, the resulting sampling distributions were different from those reported for both studies. Because of these troublesome discrepancies, the statistical analysis was not re-calculated with the missing scores, and the reported analysis reflects the mean scores as reported by Mangan, Armitage and Adams (2008) and Shurani and Chavez (2009). The result of this confound is that sampling distributions, as reported, should be considered systematically devoid of error or unexplained variance, and therefore not generalizable.

The combined decision accuracy level of these MQTZCT studies, weighted for sample size and number of scorers, was .994 with a combined inconclusive rate of .029. Reliability for MQTZCT exams was reported by Matte and Reuss (1989) as .990.[40]

Utah ZCT – Probable Lie Test

The Utah ZCT Probable Lie Test,[41] also referred to as the Utah Probable Lie Test (PLT), (Handler 2006; Handler & Nelson, 2008) and the Utah numerical scoring system (Bell et al., 1999; Handler & Nelson, 2008) were developed by researchers at the University of Utah, as a modification of the Backster ZCT (Backster, 1963). Two studies describe the criterion accuracy of the Utah PLT.

Honts, Raskin and Kircher (1987) reported the results of 10 programmed deceptive and 10 programmed truthful examinees in a study of polygraph countermeasures.[42] Unweighted decision accuracy of blind numerical scores was .889, with an inconclusive rate of .150.[43]

Kircher and Raskin (1988) reported the results from two scorers, both of whom scored 50 programmed deceptive and 50 programmed truthful examinees in a laboratory study. A total of 200 scored results were obtained. Unweighted decision accuracy of blind numerical scores was .935, with an inconclusive rate of .070.

Figure 12 shows a mean and standard deviation plot of the scores of the sampling distributions of the included Utah PLC studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [$F(1,63) = 1.682$, ($p = .200$)], nor was the main effect for sampling distribution [$F(1,63) = 0.108$, ($p = .743$)].

The combined decision accuracy level of these Utah PLT studies, weighted for sample size and number of scorers, was .931 with a combined inconclusive rate of .077. Reliability for Utah PLC exams, expressed as the average of kappa statistics for the two studies was .730, with a pairwise rate of overall decision agreement, excluding inconclusive results, of .975.

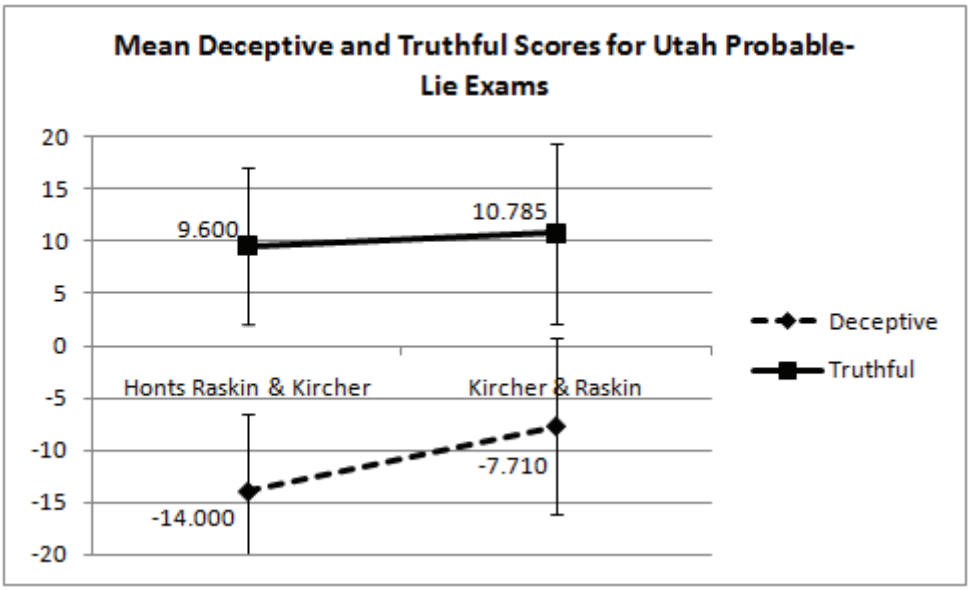
[40] This statistic was published in the Matte and Reuss (1989) reprint of the dissertation published in the journal *Polygraph*, but cannot be located in the original dissertations study for the no longer extant Columbia Pacific University.

[41] Developers of the Utah technique appear to have given little concern to the name of the test question format, and this format has also been referred to as the Utah 3-question version and the Utah PLT. Polygraph field examiners have used the term Utah ZCT because of the obvious similarities with other ZCT variants. The term Utah ZCT is used in this document to aide in the recognition of the procedural and practical similarities between this technique and other three-question ZCT formats intended for single-issue event-specific testing.

[42] Only the non-countermeasure control group cases are included in this analysis.

[43] Honts, Raskin and Kircher (1987) reported mean scores but were not required by editorial and publication standards to report standard deviations for the sampling distributions of deceptive and truthful and deceptive scores at the time of publication. Because data were no longer available to calculate these missing statistics, a blunt estimate of the pooled standard deviation was calculated from the reported t-value for the level of significance of the difference between truthful and deceptive scores.

Figure 12. Mean deceptive and truthful total scores for Utah PLT studies.



Utah ZCT – Directed Lie Test

The Utah ZCT Directed Lie Test (DLT) is a variant of the Utah PLT, using directed-lie CQs in place of probable-lie questions. Two studies describe the criterion accuracy of Utah DLC exams.

Honts and Raskin (1988) reported the criterion accuracy of Utah DLC exams of 25 criminal suspects, including 12 deceptive and 13 truthful persons, whose examination were later confirmed by confession, evidence, the confession of an alternative suspect, or the retraction of an allegation. Unweighted decision accuracy of blind numerical scores was .958, with an inconclusive rate of .077.[44]

Horowitz, Kircher, Honts and Raskin (1997) reported the results of 15 programmed deceptive and 15 programmed truthful examinees who participated in a laboratory experiment. Unweighted decision accuracy of blind numerical scores was .856, with an inconclusive rate of .067.[45]

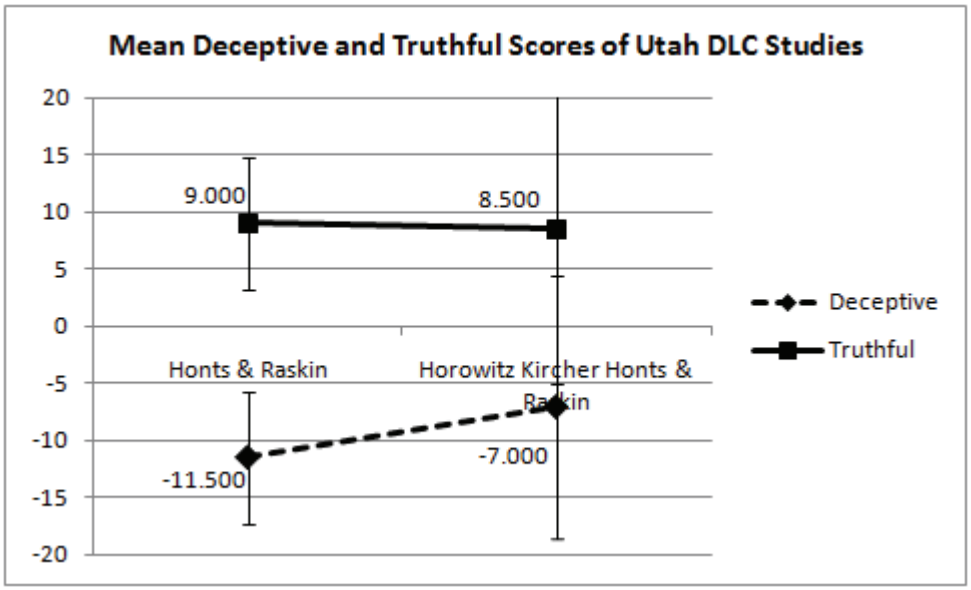
Figure 13 shows a mean and standard deviation plot of the scores of the sampling distributions of the included Utah DLT studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was not significant [F (4,51) = 0.705, (p = .592)], nor was the main effect for sampling distribution [F (2,51) = 0.009, (p = .991)].

The combined decision accuracy level of these Utah DLT studies, weighted for sample size and number of scorers, was .902 with a combined inconclusive rate of .073. Reliability for Utah DLC exams, expressed as the average of Pearson correlation coefficients for the included studies, was .930.

[44] Honts and Raskin (1988) reported mean scores but were not required by editorial and publication standards to report standard deviations for the sampling distributions of deceptive and truthful and deceptive scores at the time of publication. Because data were no longer available to calculate these missing statistics, a blunt estimate of the pooled standard deviation was calculated from the reported F-ratio for the level of significance of the difference between truthful and deceptive scores.

[45] Mean and standard deviation statistics were measured to the nearest 1/2 point from Figure 1 in Horowitz, Kircher, Honts and Raskin (1997) study report.

Figure 13. Mean deceptive and truthful total scores for Utah DLT studies.



Utah ZCT – Canadian Police College/Royal Canadian Mounted Police Version

The Canadian Police College (CPC) and the Royal Canadian Mounted Police (RCMP) have developed a variant of the Utah PLT, referred to as the RCMP Zone or the CPC Series A exam. Three studies describe the criterion accuracy of the Utah RCMP Series A exam.

Honts, Hodes and Raskin (1985) reported the criterion accuracy of Utah PLT exams using the test question sequence of the RCMP Series A exam, including 19 deceptive and 19 truthful cases. Unweighted decision accuracy of blind numerical scores was .833, with an inconclusive rate of .237.

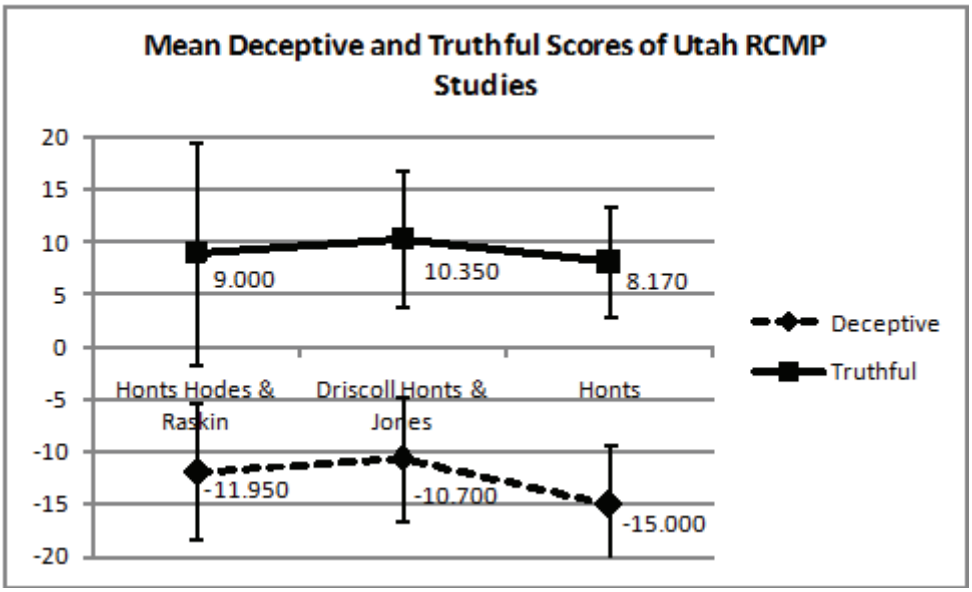
Driscoll, Honts and Jones (1987) reported the criterion accuracy Utah PLT exams, using the results of 20 programmed deceptive and 20 programmed truthful examinees who were recruited from a group counseling program at a Veterans Center, using the test question sequence the RCMP Series A exam. Decision accuracy was reported as 1.000, with an unweighted inconclusive rate of .100.

Honts (1996) reported the results of a survey of criterion accuracy of field examinations conducted by the Canadian law enforcement officers using the RCMP version of the Utah PLT. Twenty-one of the cases were confirmed as deceptive, and 11 of the cases were confirmed as truthful. Unweighted decision accuracy of blind numerical scores was .969, with an inconclusive rate of .210.

Figure 14 shows a mean and standard deviation plot of the scores of the sampling distributions of the included Utah CPC-RCMP studies. A two-way ANOVA showed that neither the interaction of sampling distributions and criterion status [F (1,99) = 0.562, (p = .455)], nor the main effect for sampling distribution [F (1,99) = 0.109, (p = .742)] were statistically significant.

The combined decision accuracy level of these Utah CPC-RCMP studies, weighted for sample size and number of scorers, was .939 with a combined inconclusive rate of .183. Reliability for Utah RCMP exams was reported by Honts (1996) as Kappa = .480 for categorical decision agreement adjusted for chance agreement. The average pairwise Pearson correlation coefficient for numerical scores of the included studies was .940, and the average proportion of decision agreement, excluding inconclusive results, was .883.

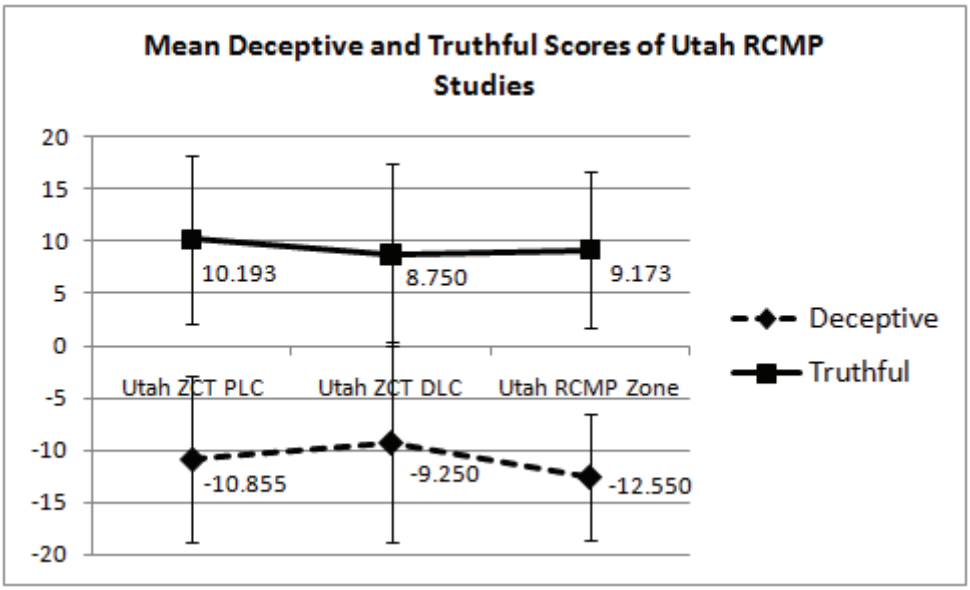
Figure 14. Mean deceptive and truthful total scores for Utah CPC-RCMP studies.



Utah ZCT – Combined PLT, DLT and RCMP Studies

Figure 15 shows a mean and standard deviation plot for the three variants of the Utah PLT. A two-way ANOVA showed that the interaction between the test variant and criterion status was not significant [F (1,246) = 2.553, (p = .111)], nor was the main effect for sampling distribution [F (2,246) = 0.02, (p = .980)]. Because the interaction was approaching a significant level, one-way post-hoc ANOVAs were also completed. Differences between the sampling distributions were not significant for the deceptive scores [F (2,100) = 0.042, (p = .959)] or for the truthful scores [F (2,100) = 0.008, (p = .992)].

Figure 15. Mean deceptive and truthful total scores for three variants of the Utah ZCT.



Unweighted decision accuracy for seven included studies pertaining to the three variants of the Utah technique, weighted for sample size and number of scorers, was .930, with an unweighted inconclusive rate of .107. Reliability statistics were averaged for all included Utah ZCT studies, and produced an average reliability statistic of kappa = .647. The average rate of decision agreement excluding inconclusive results was .958, and the average Pearson correlation coefficient for numerical scores was .913.

Event-Specific ZCT / ESS

The ESS is an evidence-based TDA model that includes normative data for ZCT examinations and other PDD techniques. Because ESS transformations are non-parametric, ESS scores are sensitive to differences in response magnitude yet robust against differences in the linearity of response magnitude.

Nelson et al. (2011) reported a summary of five previous criterion accuracy studies of ESS scores of ZCT examinations, including results reported by Nelson, Krapohl and Handler (2008), Blalock, Cushman and Nelson (2009), Nelson, Blalock, Oelrich and Cushman (2011), Handler, Nelson, Goodson and Hicks (2010) and Nelson and Krapohl (2011). These studies included 5,192 scored results from 140 scorers who evaluated 732 individual examinations. Those results consisted of 2,671 scored results of 384 confirmed deceptive examinations, and 2,521 scored results of 348 confirmed truthful exams. Examinations included both Federal ZCT and Utah ZCT exams. Unweighted decision accuracy of these scores, excluding inconclusive results was .921, and the unweighted inconclusive rate was .098.

Nelson, Krapohl and Handler (2008) reported the criterion accuracy of ESS scores of seven inexperienced examiner trainees who used the ESS to evaluate a sample of 100 exams selected from the U.S. Department of Defense confirmed case archive. Fifty of the examinations were confirmed as deceptive, and 50 of the exams were confirmed as truthful. A total of 700 scored results were obtained. Unweighted decision accuracy of blind numerical scores was .872, with an inconclusive rate of .103.

Blalock, Cushman and Nelson (2009), in a replication study, reported the criterion accuracy of a group of nine examiner trainees who used the ESS to evaluate a sample of 100 exams selected from the U.S. Department of Defense confirmed case archive. Fifty of the examinations were confirmed as deceptive, and 50 of the exams were confirmed as truthful. A total of 900 scored results were obtained. Unweighted decision accuracy of blind numerical scores was .870, with an inconclusive rate of .138.

Nelson, Blalock, Oelrich and Cushman (2011), reported the results of a reliability study involving 25 experienced examiners who used the ESS to evaluate a sample of 10 examinations selected from the U.S. Department of Defense confirmed case archive. Six of the cases were confirmed as deceptive, and four cases were confirmed as truthful. A total of 250 scored results were obtained. The pairwise proportion of decision agreement was .950, and the unweighted average of correct decisions excluding inconclusive results was .958. The unweighted inconclusive rate was .102.

Handler, Nelson, Goodson and Hicks (2010) reported the criterion accuracy of 19 examiner trainees from the México Policía Federal, who used the ESS to evaluate 100 examinations selected from the U.S. Department of Defense confirmed case archive. Fifty of the examinations were confirmed as deceptive, and 50 of the exams were confirmed as truthful. A total of 1,900 scored results were obtained. Unweighted decision accuracy of blind numerical scores was .901, with an inconclusive rate of .040.

Nelson and Krapohl (2011) reported the criterion accuracy of transformed ESS scores from six experienced federally trained examiners who evaluated a sample of 60 examinations selected from the U.S. Department of Defense confirmed case archive. Each examiner scored 10 cases. Thirty of the examinations were confirmed as deceptive, and 30 of the exams were confirmed as truthful. Unweighted decision accuracy of blind numerical scores was .913, with an unweighted inconclusive rate of .020.

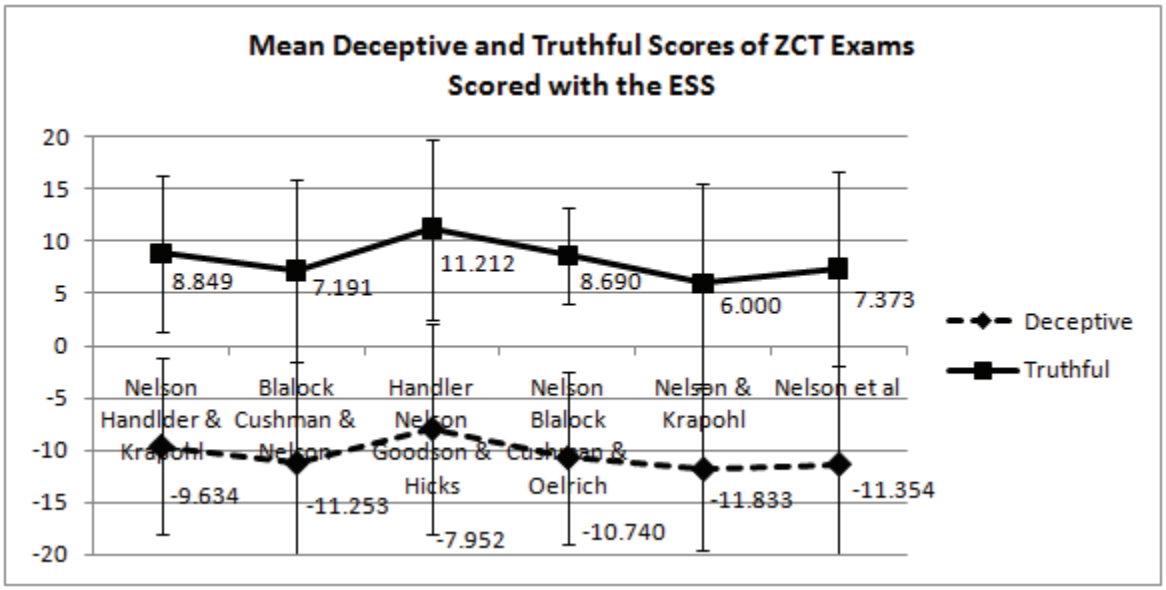
Remaining scores of the Nelson et al. (2011) results consisted of 1,382 scored

results of 572 individual examinations. These results consisted of 741 scored results of 304 confirmed deceptive examinations, and 641 scored results of 268 confirmed truthful exams. Data from the Krapohl and Cushman (2006) study were scored using an automated version of the ESS, including 50 confirmed deceptive examinations and 50 confirmed truthful exams selected from the U.S. Department of Defense confirmed case archive. These exams were also scored by a cohort of 11 examiner trainees from the Colombian Army Counterintelligence Unit, who used the ESS in pairs of two and three examiners to score 10 cases each. Data from the holdout sample used by Krapohl and McManus (1999) were also scored using an automated version of the ESS, including 30 confirmed deceptive examinations and 30 confirmed truthful exams. This holdout sample was also evaluated by a cohort of 35 scorers from Romania, consisting of 15 international polygraph examiners and 20 researchers, psychologists and graduate students, from the University of Iasi in Romania, who used the ESS while working in teams to score subsets of 10 cases each. The holdout sample was also scored by a cohort of 12 examiner trainees from the Panama National Police who worked in teams to score subsets of 10 cases each. In addition, seven examiner trainees from police agencies in the state of Ohio used the ESS to score subsets of 10 cases each from the holdout sample. One subset of 10 cases was scored by two of the Ohio police trainees.

Numerical scores from the Kircher, Kristjiansson, Gardner and Webb (2005) study (N = 80) were transformed to ESS scores, including 40 scores of confirmed deceptive exams and 40 scores of confirmed truthful exams. Data from the OSS development sample (Krapohl, 2002; Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2008) were evaluated using an automated model of the ESS, including 149 scores for confirmed deceptive exams and 143 scores for confirmed truthful exams. Seven-position scores from two experts who participated in the Kircher and Raskin (1988) study were transformed to ESS scores, including 100 scores for 50 examinations of programmed deceptive examinees, and 100 scores for 50 exams conducted on programmed truthful examinees. Finally, seven-position scores from three expert scorers who evaluated the cases for the Blackwell (1998) study were transformed to ESS scores, including 195 scores for 65 confirmed deceptive examinations, and 105 scores for 35 confirmed truthful exams.

Figure 16 shows a mean and standard deviation plot of the scores of the sampling distributions of the included ZCT ESS studies. A two-way ANOVA showed that the interaction

Figure 16. Mean deceptive and truthful total scores for ZCT ESS studies.



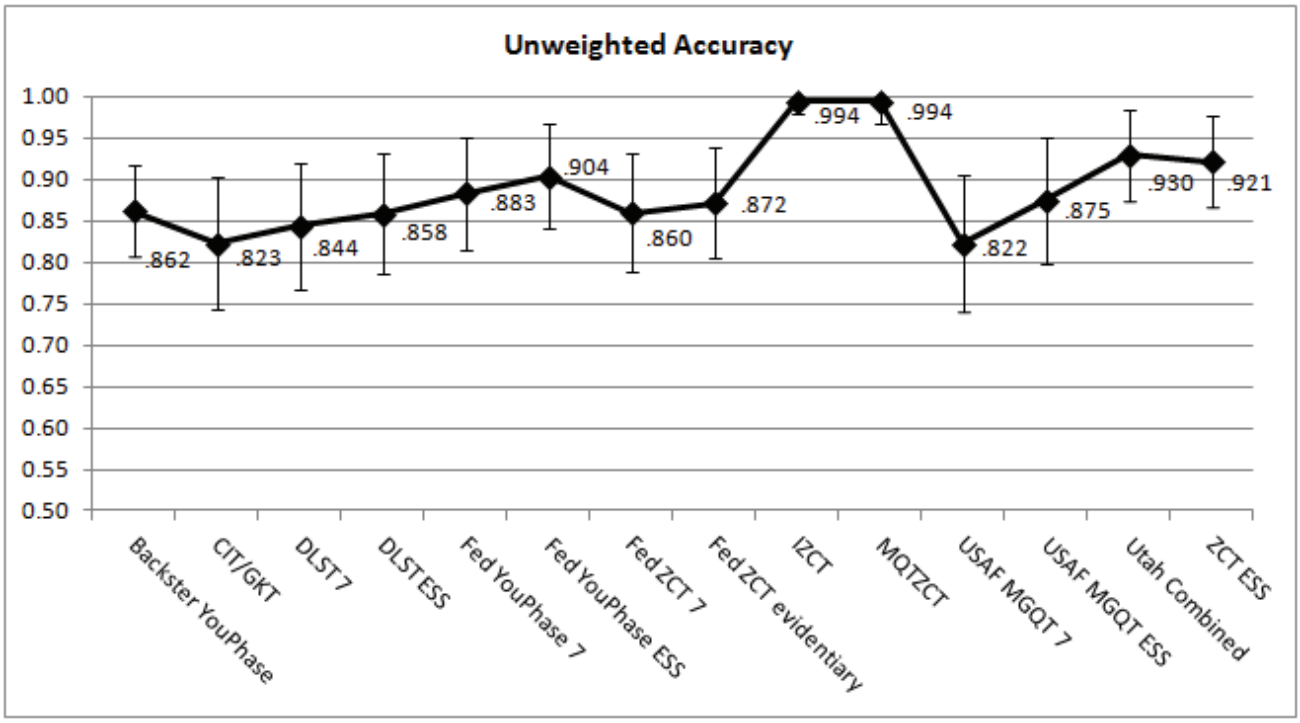
of sampling distribution and criterion status was not significant [F (1,215) = 0.205, (p = .651)], nor was the main effect for sampling distribution [F (5,215) = 0.164, (p = .976)].

The combined decision accuracy level of the included ZCT ESS studies, weighted for sample size and number of scorers, was .922 with a combined inconclusive rate of .098, as reported by Nelson et al. (2011). Reliability statistics for ZCT ESS studies were average to produce a pairwise proportion of decision agreement, excluding inconclusive results, of .950, with average Kappa = .585.

Criterion Accuracy For All Validated Techniques

A one-way ANOVA for unweighted decision accuracy showed that differences in unweighted decision accuracy for these 14 PDD techniques were significant [F (13,5119) = 2.753, (p < .001)]. One-way ANOVAs for case status showed that differences in correct decisions were significant for both criterion deceptive cases [F (13,2494) = 1.982, (p = .019)] and criterion truthful cases [F (13,2542) = 2.764, (p <. .001)]. Figure 17 shows the mean and confidence intervals for the unweighted accuracy of 14 techniques included in the meta-analysis.

Figure 17. Mean and confidence intervals for unweighted decision accuracy of 14 PDD techniques.



A series of ANOVA contrasts showed that differences were significant only for two PDD techniques, the IZCT and the MQTZCT. Exclusion of these two PDD techniques resulted in no significant differences [F (11,4859) = 0.949, (p = .491)] in the unweighted accuracy for the remaining 12 PDD techniques. One-sample t-tests further confirmed the outlier status of the results of these two techniques: t = 212.268 (p < .001) for both the IZCT and the MQTZCT.[46] A series of leave-one-out t-tests revealed that none of

[46] t-values are the same for both of these techniques because the unweighted mean of weighted sampling means is the same for both techniques (.994).

the other techniques produced outlier results when compared to the results of all other techniques.

APA 2012 criterion validity standards

Table 1 (also shown in the Executive Summary) shows a list of the 14 PDD techniques that satisfied the requirements for inclusion in this meta-analysis at criterion accuracy levels specified in the APA 2012 standard requirements for evidentiary testing, paired-testing, and investigative testing. Also shown in Table 1 are the unweighted decision accuracy and inconclusive rates for each PDD technique. Additional details concerning the sampling distributions and a complete dimensional profile of criterion accuracy for each of these techniques and all included studies can be found in Appendix E.

The combination of all validated PDD techniques, excluding outlier results, produced a decision accuracy of .869 (.036) without inconclusive results. The 95% confidence range was from .798 to .940. The mean inconclusive rate, excluding outlier results, was .128 (.030) with a 95% confidence range of .068 to .187. Aggregated reliability statistics produces a mean Kappa statistic of .642 (.102) with a 95% confidence range of .443 to .842. The mean rate of inter-rater decision agreement, excluding outlier results and excluding inconclusive results, was .901 (.082) with a 95% confidence range from .741 to .999. The mean Pearson correlation coefficient for numerical scores, excluding outlier results, was .876 (.116) with a 95% confidence range of .649 to .999. Table 2 shows the aggregated criterion accuracy profile of all validated CQT PDD techniques, weighted for the sample size and number of scorers.[47] Also shown in Table 2 is the criterion accuracy profile including outlier results.

[47] A majority of examinations in field polygraph programs are conducted using PDD techniques that are interpreted with an assumption of criterion independence among the RQs. However, a majority of PDD criterion validity research has been conducted using PDD techniques that are interpreted with an assumption of non-independence. Non-independent examination techniques have greater statistical discrimination power and greater accuracy than independent exam techniques. For this reason, the unweighted average was considered to be a more conservative and generalizable estimate of the overall accuracy of all PDD examination techniques. This was calculated as the unweighted average of the weighted aggregation of independent techniques and the weighted aggregation of non-independent techniques. Calculation of the weighted average would result in an overestimation of accuracy. The unweighted average of PPV and NPV might be a more optimistic and flattering under some conditions, but can be expected to be less generalizable to field circumstances when base rates are unknown or different than the base rates in the study samples.

Table 1. Mean (standard deviation) and {95% confidence intervals} for correct decisions (CD) and inconclusive results (INC) for validated PDD techniques.

Evidentiary Techniques/ TDA Method	Paired Testing Techniques/ TDA Method	Investigative Techniques/ TDA Method
Federal You-Phase / ESS[1] CD = .904 (.032) {.841 to .966} INC = .192 (.033) {.127 to .256}	**AFMGQT[4,8] / ESS[5]** CD = .875 (.039) {.798 to .953} INC = .170 (.036) {.100 to .241}	**AFMGQT[6,8] / 7 position** CD = .817 (.042) {.734 to .900} INC = .197 (.030) {.138 to .255}
Event-Specific ZCT / ESS CD = .921 (.028) {.866 to .977} INC = .098 (.030) {.039 to .157}	**Backster You-Phase / Backster** CD = .862 (.037) {.787 to .932} INC = .196 (.040) {.117 to .275}	**CIT[7] / Lykken Scoring** CD = .823 (.041) {.744 to .903} INC = NA
IZCT / Horizontal[2] CD = .994 (.008) {.978 to .999} INC = .033 (.019) {.001 to .069}	**Federal You-Phase / 7 position** CD = .883 (.035) {.813 to .952} INC = .168 (.037) {.096 to .241}	**DLST (TES)[8] / 7 position** CD = .844 (.039) {.768 to .920} INC = .088 (.028) {.034 to .142}
MQTZCT / Matte[3] CD = .994 (.013) {.968 to .999} INC = .029 (.015) {.001 to .058}	**Federal ZCT / 7 position** CD = .860 (.037) {.801 to .945} INC = .171 (.040) {.113 to .269}	**DLST (TES)[8] / ESS** CD = .858 (.037) {.786 to .930} INC = .090 (.026) {.039 to .142}
Utah ZCT DLT / Utah CD = .902 (.031) {.841 to .962} INC = .073 (.025) {.023 to .122}	**Federal ZCT / 7 pos. evidentiary** CD = .880 (.034) {.813 to .948} INC = .085 (.029) {.028 to .141}	-
Utah ZCT PLT / Utah CD = .931 (.026) {.879 to .983} INC = .077 (.028) {.022 to .133}	-	-
Utah ZCT Combined / Utah CD = .930 (.026) {.875to .984} INC = .107 (.028) {.048 to .165}	-	-
Utah ZCT CPC-RCMP Series A / Utah CD = .939 (.038) {.864 to .999} INC = .185 (.041) {.104 to .266}	-	-

[1] Empirical Scoring System.

[2] Generalizability of this outlier result is limited by the fact that no measures of test reliability have been published for this technique. Also, significant differences were found in the sampling distributions of the included studies, suggesting that the samples data are not representative of each other, or that the exams were administered and/or scored differently. One of the studies involved a small sample (N = 12) that was reported in two articles, for which the participating scorer was also the technique developer. One of the publications described the study as a non-blind pilot study. Both reports indicated that one of the six truthful participants was removed from the study after making a false-confession. The reported perfect accuracy rate did not include the false confession. Neither the perfect accuracy nor the .167 false-confession rate are likely to generalize to field settings.

[3] Generalizability of this outlier result is limited by the fact that the developers and investigators have advised the necessity of intensive training available only from experienced practitioners of the technique, and have suggested that the complexity of the technique exceeds that which other professionals can learn from the published resources. The developer reported a near-perfect correlation coefficient of .99 for the numerical scores, suggesting an unprecedented high rate of inter-scorer agreement, which is unexpected given the purported complexity of the method. Additionally, the data initially provided to the committee for replication studies included only those cases for which the scorers arrived at the correct decision, excluding scores from those cases for which the scorers did not achieve the correct decision. Missing scores were later provided to the committee for both the Mangan et al (2008) and Shurani and Chavez (2009) studies. However, the resulting sampling means were different from those reported for both replication studies. Because of these discrepancies, the statistical analysis was not re-calculated with the missing scores, and the reported analysis reflects the sampling distribution means as reported. Sampling means for replication studies should be considered devoid of error or uncontrolled variance.

[4] Two versions exist for the AFMGQT, with minor structural differences between them. There is no evidence that the performance of one version is superior to the other. Because replicated evidence would be required to reject a null-hypothesis that the differences are meaningless, and because the selected studies include a mixture of both AFMGQT versions, these results are provided as a generalizable to both versions. AFMGQT exams are used in both multi-facet event-specific contexts and multi-issue screening contexts. Both multi-facet and multi-issue examinations were interpreted with decision rules based on an assumption of criterion independence among the RQs.

[5] The AFMGQT produced accuracy that is satisfactory for paired testing only when scored with the Empirical Scoring System.

[6] There are two techniques for which there are no published studies but which are structurally nearly identical to the AFMGQT: the LEPET and the Utah MGQT. Validity of the AFMGQT can be generalized to these techniques if scored with the same TDA methods.

[7] Concealed Information Test, also referred to as the Guilty Knowledge Test (GKT) and Peak of Tension test (POT). The data used here were provided in the meta-analysis report of laboratory research by MacLaren (2001).

[8] Studies for these PDD techniques were conducted using decision rules based on the assumption of criterion independence among the testing targets. Accuracy of screening techniques may be further improved by the systematic use of a successive-hurdles approach.

Table 2. Mean (standard deviation) and {95% confidence Interval} for criterion accuracy profiles for all validated PDD techniques combined.

	Excluding outlier results	All included studies
Number of PDD Techniques	12	14
Number of Studies	39	45
N Deceptive	2,067	2,336
N Truthful	1,802	2,031
Total N	3,869	4,367
Number Scorers	280	295
N of Deceptive Scores	5,840	6,109
N of Truthful Scores	5,399	5,628
Total Scores	11,239	11,737
Percent Correct	.869 (.036) {.798 to .940}	.887 (.033) {.823 to .951}
Inconclusive	.128 (.030) {.068 to .187}	.114 (.028) {.058 to .170}
Sensitivity	.812 (.056) {.702 to .923}	.835 (.051) {.734 to .936}
Specificity	.717 (.061) {.597 to .838}	.751 (.058) {.638 to .864}
FN Errors	.083 (.038) {.008 to .157}	.072 (.035) {.004 to .141}
FP Errors	.144 (.049) {.048 to .239}	.123 (.043) {.039 to .208}
D Inc	.105 (.042) {.022 to .187}	.092 (.037) {.020 to .165}
T Inc	.151 (.042) {.068 to .234}	.136 (.041) {.056 to .216}
PPV	.854 (.049) {.757 to .950}	.874 (.044) {.789 to .960}
NPV	.899 (.047) {.807 to .990}	.911 (.043) {.827 to .995}
D Correct	.909 (.042) {.826 to .992}	.921 (.039) {.844 to .997}
T Correct	.829 (.056) {.721 to .938}	.854 (.049) {.757 to .950}

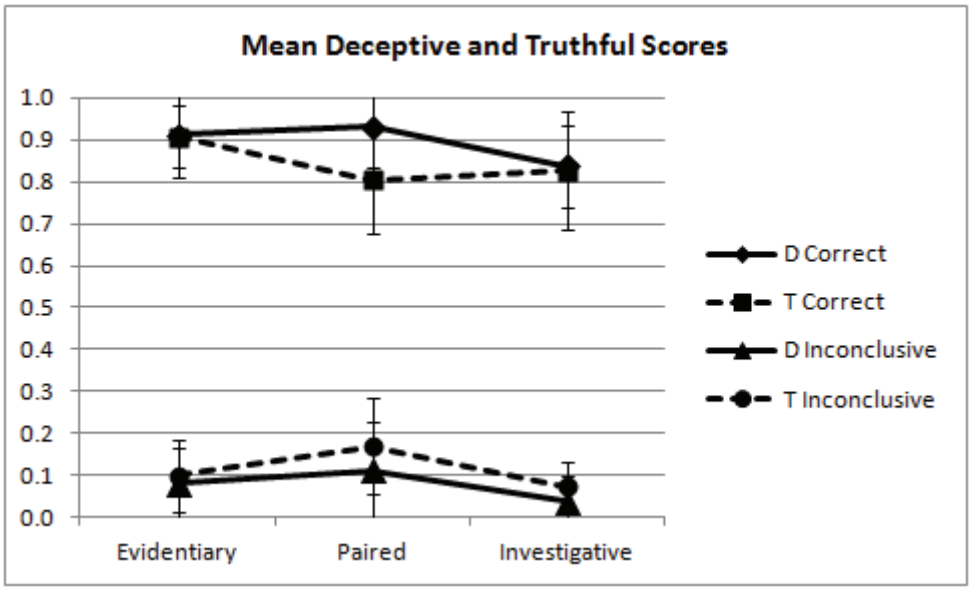
Table 3 shows the criterion accuracy profile of the weighted aggregation of PDD techniques at the evidentiary, paired testing and investigative levels according to the APA 2012 standards. Also shown in Table 3 is the criterion accuracy profile for evidentiary techniques without the results of the two outlier techniques.

Table 3. Criterion accuracy profiles for evidentiary, paired-testing, and investigative techniques.

	Evidentiary Techniques	Evidentiary w/o Outlier Studies	Paired-testing Techniques	Investigative Techniques
Number of Techniques	5	3	5	4
Number of Studies	21	15	12	12
N Deceptive	861	592	435	1,040
N Truthful	776	547	408	847
Total N	1,637	1,139	843	1,887
Number Scorers	174	159	56	65
N of Deceptive Scores	3,297	3,028	1,700	1,112
N of Truthful Scores	3,098	2,869	1,613	917
Total Scores	6,395	5,897	3,313	2,029
Unweighted Average Accuracy	.910 (.027) {.857 to .963}	.903 (.028) {.847 to .958}	.867 (.036) {.796 to .938}	.844 (.039) {.767 to .920}
Unweighted Average Inconclusives	.090 (.029) {.032 to .147}	.095 (.030) {.035 to .154}	.142 (.036) {.071 to .213}	.114 (.028) {.060 to .168}
Sensitivity	.843 (.050) {.745 to .941}	.832 (.053) {.729 to .935}	.828 (.051) {.728 to .928}	.802 (.047) {.710 to .893}
Specificity	.826 (.054) {.721 to .931}	.816 (.055) {.708 to .923}	.670 (.071) {.531 to .809}	.771 (.073) {.627 to .915}
FN Errors	.082 (.033) {.018 to .147}	.089 (.034) {.021 to .156}	.060 (.032) {.001 to .123}	.158 (.042) {.076 to .240}
FP Errors	.083 (.035) {.014 to .152}	.090 (.037) {.018 to .162}	.159 (.052) {.056 to .261}	.159 (.070) {.022 to .296}
D Inc	.080 (.038) {.005 to .155}	.086 (.041) {.004 to .167}	.112 (.043) {.028 to .195}	.038 (.020) {.001 to .077}
T Inc	.099 (.044) {.014 to .185}	.104 (.044) {.017 to .191}	.170 (.058) {.056 to .284}	.073 (.015) {.043 to .102}
PPV	.915 (.034) {.848 to .982}	.908 (.037) {.836 to .979}	.847 (.050) {.749 to .945}	.860 (.037) {.788 to .933}
NPV	.904 (.042) {.823 to .986}	.898 (.043) {.814 to .982}	.920 (.046) {.829 to .999}	.812 (.082) {.651 to .973}
D Correct	.911 (.037) {.839 to .983}	.904 (.039) {.828 to .98}	.932 (.036) {.862 to .999}	.837 (.045) {.749 to .924}
T Correct	.908 (.038) {.833 to .983}	.901 (.040) {.822 to .980}	.804 (.064) {.678 to .930}	.827 (.072) {.686 to .968}

Figure 18 shows the mean and statistical confidence intervals for correct decisions and inconclusive results for three levels of criterion validity described by the APA 2012 standards. Two-way ANOVAs, including outlier results, showed a significant interaction between criterion status and validation category for correct decisions [F (1,10896) = 7433.144, (p < .001)] and for inconclusive results [F (1,10896) = 3562.384, (p < .001)], suggesting that techniques at these different categorical levels may handle deceptive and truthful cases with different effectiveness. However, post-hoc one-way ANOVAs showed there was no significant difference in the proportion of correct decisions inconclusive results, or errors for deceptive or truthful cases.

Figure 18. Mean and confidence interval plot for APA validation categories.



Evidentiary testing techniques

Five PDD techniques were reported to produce both sufficiently high levels of diagnostic accuracy and low inconclusive rates that satisfy the APA 2012 standard requirements for evidentiary testing. Scores from 21 surveys and experiments were summarized to describe the criterion validity of these evidentiary techniques. Studies that support these evidentiary techniques included 174 scorers who provided 7,407 numerical scores for 1,637 confirmed exams, including 3,821 numerical scores for 861 confirmed deceptive examinations and 3,586 numerical scores for 776 confirmed truthful examinations.

Table 4 shows the criterion accuracy profiles for the five PDD techniques that satisfy the APA 2012 requirements for evidentiary/diagnostic testing. Also shown in Table 4 are the number of included studies for each PDD technique, the total number of scored results, reliability along with the mean and standard deviations of the average deceptive and truthful scores of the included studies. Mean test sensitivity, test specificity, and unweighted accuracy have been reported at levels that are statistically significantly greater than chance (50%) for each of these five PDD techniques.

Table 4. Criterion accuracy profiles for evidentiary/diagnostic PDD techniques.					
Technique	Federal You-Phase	IZCT*	MQTZCT*	Utah PLT (combined)	ZCT ESS
TDA Method	ESS	Horizontal	Matte	Utah	ESS
Number of Studies	*2*	*3*	*3*	*7*	*6*
N Deceptive	61	86	183	147	384
N Truthful	61	93	139	138	348
Total N	122	179	319	285	732
Number Scorers	*11*	*8*	*7*	*8*	*140*
N of Deceptive Scores	160	86	183	197	2671
N of Truthful Scores	160	93	136	188	2521
Total Scores	*320*	*179*	*319*	*385*	*5192*
Mean D	-7.512	-21.505	-8.711	-10.885	-10.46
StDev D	6.184	12.606	2.489	7.878	8.949
Mean T	6.146	19.626	5.226	9.372	8.219
StDev T	6.217	4.232	3.479	8.066	8.051
Reliability - Kappa	-	-†	-	*.650*	*0.59*
Reliability - Agreement	*.900*	-†	-	*.960*	*.950*
Reliability - Correlation	-	-†	*.990*‡	*.910*	-
Unweighted Average Accuracy	.904 (.032) {.841 to .966}	.994 (.008) {.978 to .999}	.994 (.013)* {.968 to .999}	.930 (.028) {.875 to .984}	.921 (.028) {.866 to .977}
Unweighted Average Inconclusives	.192 (.033) {.127 to .256}	.033 (.019) {.001 to .069}	.029 (.015) {.001 to .058}	.107 (.030) {.048 to .165}	.098 (.030) {.039 to .157}
Sensitivity	.845 (.052) {.742 to .948}	.977 (.020) {.937 to .999}	.967 (.021) {.926 to .999}	.853 (.049) {.757 to .948}	.817 (.056) {.706 to .927}
Specificity	.757 (.064) {.633 to .882}	.946 (.035) {.878 to .999}	.963 (.033) {.899 to .999}	.809 (.056) {.699 to .918}	.846 (.051) {.747 to .946}
FN Errors	.034 (.026) {.001 to .085}	.012 (.015) {.001 to .041}	.011 (.021) {.001 to .052}	.051 (.031) {.001 to .112}	.077 (.037) {.004 to .151}
FP Errors	.138 (.050) {.039 to .236}	.001 (.005) {.001 to .01}	.001 (.015) {.001 to .03}	.074 (.038) {.001 to .148}	.064 (.034) {.001 to .130}
D INC	.128 (.046) {.037 to .219}	.012 (.014) {.001 to .040}	.022 (.001) {.022 to .022}	.096 (.040) {.017 to .176}	.106 (.044) {.020 to .192}
T INC	.255 (.044) {.170 to .341}	.054 (.035) {.001 to .122}	.037 (.029) {.001 to .094}	.117 (.046) {.027 to .207}	.089 (.042) {.008 to .171}
PPV	.860 (.050) {.761 to .958}	.999 (.004) {.991 to .999}	.999 (.015) {.970 to .999}	.923 (.039) {.847 to .999}	.931 (.038) {.857 to .999}
NPV	.957 (.033) {.892 to .999}	.989 (.019) {.952 to .999}	.985 (.021) {.944 to .999}	.938 (.036) {.867 to .999}	.912 (.042) {.830 to .993}
D Correct	.961 (.029) {.903 to .999}	.988 (.015) {.959 to .999}	.989 (.021) {.948 to .999}	.944 (.034) {.877 to .999}	.913 (.042) {.831 to .996}
T Correct	.846 (.056) {.736 to .956}	.999 (.005) {.989 to .999}	.999 (.015) {.969 to .999}	.916 (.043) {.832 to .999}	.929 (.037) {.857 to .999}

* Outlier results that differ significantly from the normal range of the other techniques.

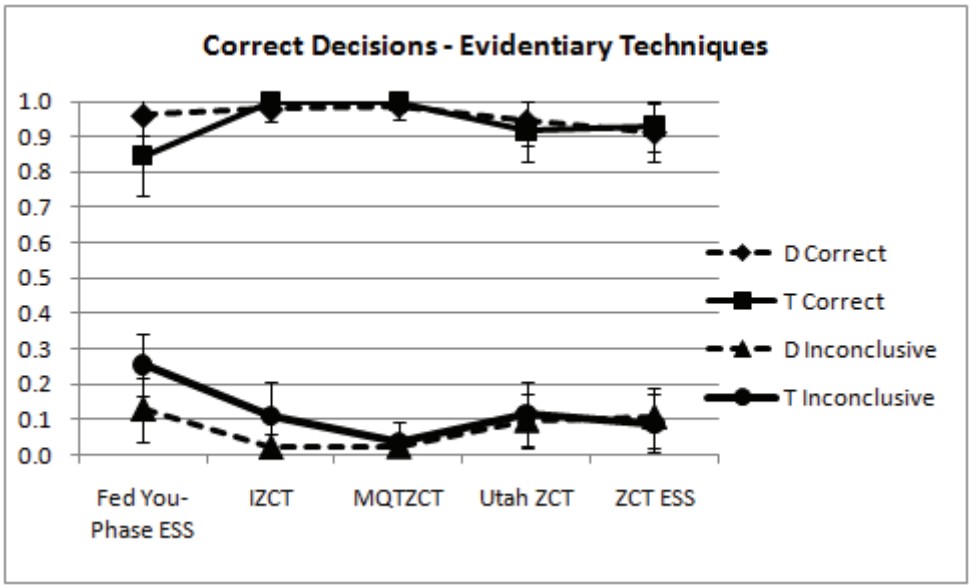
† No reliability data has been published for any of the studies on the IZCT.

‡ A correlation coefficient of .990 is an extraordinary and remarkable finding in any field of research, and suggests an extremely low rate of disagreement between the numerical scores of blind evaluators using the MQTZCT. This statistic cannot be found in the Matte and Reuss (1989) dissertation paper for the now defunct Columbia Pacific University, but was published in the included Matte and Reuss (1989) reprint in *Polygraph*. Despite this extremely high correlation of numerical scores from different scorers, developers and researchers of the MQTZCT have expressed repeated cautions regarding the lack of generalizability of MQTZCT results without intensive proprietary training.

A two-by-five way ANOVA, criterion status x technique, for correct decisions showed a significant interaction between technique and case status [F (1,7397) = 8944.964, (p < .001)], indicating that these five different evidentiary techniques handled deceptive and truthful cases differently.

Post-hoc one-way ANOVAs showed that differences in the rate of correct decisions for criterion truthful cases were significant [F (4,828) = 3.118 (p = .015)], while differences in the rate of correct decisions for deceptive cases were not significant. A series of ANOVA contrasts showed that differences were significant only for the two outlier techniques. There were no significant differences when the outliers were not included. Figure 19 shows a mean and confidence interval plot for correct decisions and inconclusive results of the five evidentiary techniques.

Figure 19. Means and confidence intervals for evidentiary techniques.



Paired-testing techniques

Five techniques were identified as providing a sufficient level of accuracy to satisfy the APA requirement for paired-testing.[48] Scores from 12 surveys and experiments were summarized to describe the criterion validity of these paired-testing techniques. Studies that support these paired-testing techniques included 56 scorers who provided 3,313 numerical scores for 843 confirmed exams, including 1,700 numerical scores for 435 confirmed deceptive examinations and 1,613 numerical scores for 408 confirmed truthful examinations.

Table 5 shows the criterion accuracy profiles for the five PDD techniques that satisfy the APA 2012 requirements for paired testing. Also shown in Table 5 are the number of included studies for each PDD technique, the total number of scored results, reliability along with the mean and standard deviations of the average deceptive and truthful scores of the included studies.

[48] All PDD techniques that meet the APA 2012 standard requirement for evidentiary testing also meet the requirements for paired-testing and investigative testing. Those PDD techniques that that meet the criterion accuracy requirement for paired-testing are also sufficiently valid for investigative testing.

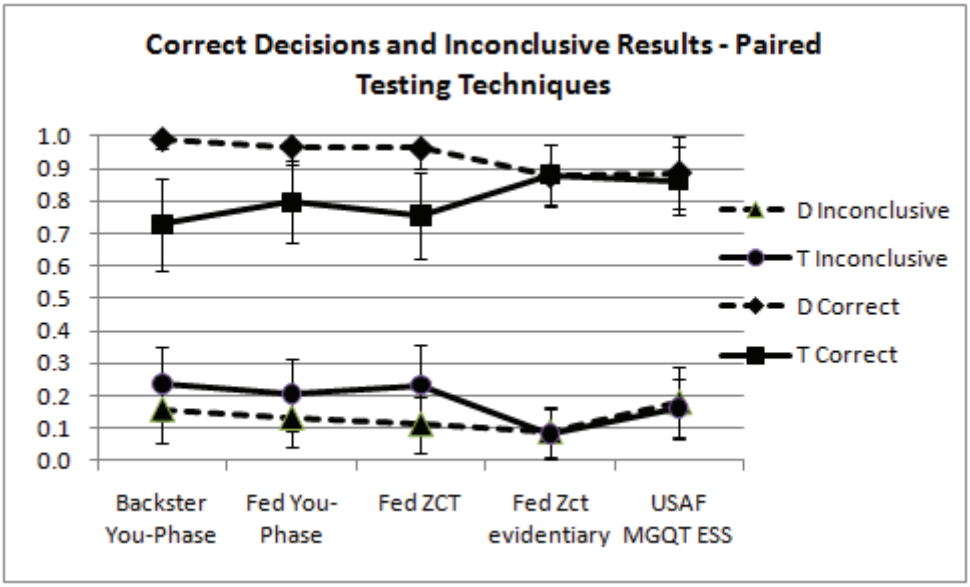
Although test sensitivity and unweighted decision accuracy is significantly greater than chance for all five of these paired testing techniques, three of these techniques produced test specificity levels that were not significantly greater than chance (Backster You-Phase/Backster, Federal You-Phase/7-position, & Federal ZCT/7-position).

Table 5. Criterion accuracy profiles for paired-testing techniques.

Technique	Backster You-Phase	Federal You-Phase	Federal ZCT	Federal ZCT	AFMGQT
TDA Method	Backster	7-position	7-position	7-position evidentiary	ESS
Number of Studies	*2*	*2*	*3*	*2*	*3*
N Deceptive	61	61	139	80	94
N Truthful	61	61	109	80	97
Total N	122	122	248	160	191
Number Scorers	*8*	*11*	*16*	*16*	*5*
N of Deceptive Scores	127	160	767	530	116
N of Truthful Scores	127	160	677	530	119
Total Scores	*254*	*320*	*1,444*	*1,060*	*235*
Mean D	-16.055	-7.195	-8.577	-8.263	-2.960
StDev D	7.417	5.824	9.018	9.032	4.765
Mean T	5.216	5.999	7.466	7.852	3.738
StDev T	10.291	5.893	8.472	9.721	4.104
Reliability - Kappa	-	-	*.570*	-	-
Reliability - Agreement	-	*.850*	*.800*	*.870*	*1*
Reliability - Correlation	*.567*	-	-	-	*.930*
Unweighted Average Accuracy	.862 (.037) {.787 to .932}	.883 (.035) {.813 to .952}	.860 (.037) {.788 to .931}	.880 (.034) {.813 to .948}	.875 (.039) {.798 to .953}
Unweighted Average Inconclusives	.196 (.040) {.117 to .275}	.168 (.037) {.096 to .241}	.171 (.040) {.093 to .249}	.085 (.029) {.028 to .141}	.170 (.036) {.100 to .241}
Sensitivity	.836 (.052) {.734 to .938}	.841 (.050) {.742 to .939}	.858 (.051) {.759 to .957}	.804 (.054) {.697 to .911}	.729 (.065) {.603 to .856}
Specificity	.556 (.070) {.418 to .694}	.632 (.069) {.497 to .768}	.581 (.073) {.438 to .723}	.809 (.057) {.698 to .920}	.700 (.063) {.577 to .823}
FN Errors	.007 (.012) {-.016 to .03}	.028 (.023) {.001 to .073}	.033 (.029) {.001 to .090}	.110 (.044) {.024 to .197}	.092 (.046) {.002 to .182}
FP Errors	.207 (.058) {.091 to .322}	.161 (.051) {.061 to .261}	.188 (.051) {.089 to .287}	.109 (.044) {.022 to .196}	.112 (.047) {.02 to .204}
D INC	.156 (.051) {.055 to .257}	.131 (.046) {.041 to .221}	.110 (.044) {.023 to .196}	.087 (.039) {.010 to .163}	.178 (.056) {.068 to .289}
T INC	.236 (.059) {.119 to .354}	.205 (.057) {.093 to .318}	.232 (.064) {.106 to .358}	.083 (.040) {.003 to .162}	.162 (.047) {.071 to .254}
PPV	.801 (.055) {.693 to .909}	.840 (.053) {.736 to .943}	.838 (.053) {.734 to .943}	.880 (.048) {.786 to .974}	.864 (.058) {.751 to .977}
NPV	.987 (.021) {.945 to .999}	.958 (.035) {.889 to .999}	.940 (.046) {.851 to .999}	.880 (.048) {.786 to .974}	.887 (.052) {.785 to .989}
D Correct	.991 (.014) {.963 to .999}	.968 (.027) {.916 to .999}	.963 (.033) {.898 to .999}	.879 (.048) {.786 to .973}	.888 (.057) {.777 to .999}
T Correct	.728 (.073) {.584 to .873}	.797 (.064) {.672 to .923}	.756 (.067) {.625 to .887}	.881 (.048) {.786 to .976}	.862 (.053) {.758 to .967}

Figure 20 shows a mean and confidence interval plot for correct decisions and inconclusive results of the five paired testing techniques. A two-by-five way ANOVA, criterion status x technique, for correct decisions showed a significant interaction [F (1,3303) = 5891.333, (p < .001)], indicating that these five paired testing techniques produced different rates of correct decisions for deceptive and truthful cases. Post-hoc one-way ANOVAs showed no significant differences in the rate of correct decisions for criterion deceptive cases or criterion truthful cases.

Figure 20. Means and confidence intervals for paired-testing techniques.



A two-by-five way ANOVA, criterion status x technique, for inconclusive decisions showed a significant interaction [F (1,3303) = 5891.333, (p < .001)], indicating that these five paired-testing techniques produced different rates of inconclusive results for deceptive and truthful cases. Post-hoc one-way ANOVAs showed that differences in inconclusive rates were not significant for criterion deceptive or criterion truthful cases.

Investigative testing techniques

Four PDD techniques produced criterion accuracy that satisfies the APA requirement for investigative testing. Scores from 12 surveys and experiments were summarized to describe the criterion validity of these investigative techniques.[49] Studies that support these investigative techniques included 65 scorers who provided 2,029 numerical scores for 1,887 confirmed exams, including 1,112 numerical scores for 1,040 confirmed deceptive examinations and 917 numerical scores for 847 confirmed truthful examinations.

Table 6 shows the criterion accuracy profiles for the four PDD techniques that satisfy the APA 2012 requirements for investigative testing. Also shown in Table 6 are the number of included studies for each PDD technique, the total number of scored results, reliability along with the mean and standard deviations of the average deceptive and truthful scores of the included studies. Unweighted decision accuracy and test sensitivity has been reported as significantly greater than chance for all four of these investigative techniques. Three of these investigative techniques, the CIT, and

[49] One of the included studies was a meta-analysis that summarized the results of laboratory studies using the CIT.

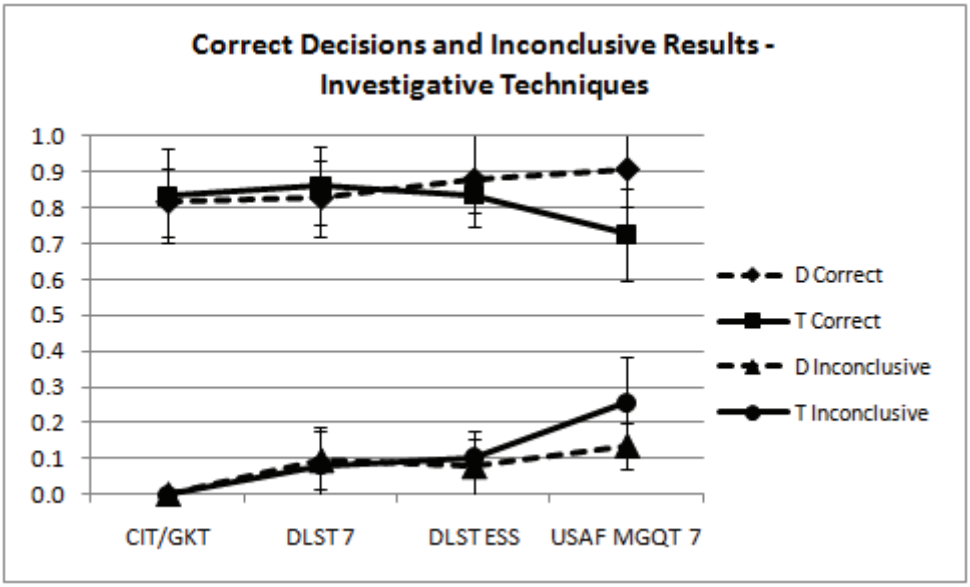
DLST/TES scored with both the seven-position and ESS models, produced test specificity that was significantly greater than chance. Test specificity for one investigative technique, the AFMGQT scored with the seven-position model, was not significantly greater than chance.

Table 6. Criterion accuracy for investigative techniques.

Technique	CIT/GKT	DLST/TES	DLST/TES	AFMGQT
TDA Method	Lykken	7-position	ESS	7-position
Number of Studies	*39*	*4*	*4*	*3*
N Deceptive	666	131	149	94
N Truthful	404	197	149	97
Total N	1070	328	298	191
Number Scorers	*39*	*16*	*5*	*5*
N of Deceptive Scores	*666*	*156*	*174*	*116*
N of Truthful Scores	*404*	*221*	*173*	*119*
Total Scores	*1070*	*377*	*347*	*235*
Mean D	-	-2.126	-2.131	-2.607
StDev D	-	3.959	3.801	4.754
Mean T	-	3.162	3.412	3.114
StDev T	-	3.531	3.153	3.705
Reliability - Kappa	-	*.760*	-	*.750*
Reliability - Agreement	-	*.806*	*.840*	*.965*
Reliability - Correlation	-	-	-	*.940*
Unweighted Average Accuracy	.823 (.041) {.744 to .903}	.844 (.039) {.768 to .920}	.858 (.037) {.786 to .930}	.817 (.042) {.734 to .900}
Unweighted Average Inconclusives	.001 (.001) {.001 to .001}	.088 (.028) {.034 to .142}	.090 (.026) {.039 to .142}	.197 (.030) {.138 to .255}
Sensitivity	.815 (.048) {.721 to .910}	.748 (.062) {.626 to .869}	.809 (.069) {.674 to .945}	.783 (.058) {.669 to .896}
Specificity	.832 (.067) {.700 to .963}	.792 (.060) {.674 to .909}	.751 (.031) {.691 to .811}	.538 (.068) {.405 to .672}
FN Errors	.185 (.048) {.090 to .279}	.156 (.050) {.058 to .255}	.112 (.057) {.001 to .224}	.079 (.050) {.001 to .177}
FP Errors	.168 (.067) {.037 to .300}	.127 (.052) {.026 to .229}	.146 (.027) {.093 to .2}	.203 (.057) {.090 to .315}
D INC	.001 (.001) {.001 to .001}	.096 (.041) {.016 to .175}	.078 (.052) {.001 to .180}	.137 (.033) {.071 to .202}
T INC	.001 (.001) {.001 to .001}	.081 (.037) {.008 to .153}	.102 (.014) {.075 to .130}	.257 (.049) {.160 to .354}
PPV	.889 (.037) {.816 to .961}	.806 (.055) {.698 to .914}	.848 (.041) {.767 to .928}	.79 (.059) {.675 to .905}
NPV	.732 (.076) {.583 to .881}	.878 (.054) {.772 to .983}	.870 (.052) {.768 to .971}	.874 (.062) {.753 to .996}
D Correct	.815 (.048) {.721 to .910}	.827 (.055) {.719 to .935}	.878 (.067) {.746 to .999}	.908 (.053) {.804 to .999}
T Correct	.832 (.067) {.700 to .963}	.861 (.055) {.753 to .969}	.837 (.027) {.783 to .891}	.726 (.066) {.597 to .856}

Figure 21 shows a mean and confidence interval plot for correct decisions and inconclusive results of the five paired testing techniques. A two-by-four way ANOVA, criterion status x technique, for correct decisions showed a significant interaction [F (1,2021) = 1320.745, (p <.001)], indicating that these four investigative techniques differed in their abilities to correctly classify deceptive and truthful cases. However, post-hoc one-way ANOVAs showed no significant differences in the rate of correct decisions for criterion deceptive or criterion truthful cases.

Figure 21. Means and confidence intervals for investigative techniques.



A two-by-three way ANOVA, criterion status x technique, for inconclusive decisions showed a significant interaction [F (1,953) = 404.177, (p < .001)], indicating that these three investigative techniques produced different rates of inconclusive results for deceptive and truthful cases. Post-hoc one-way ANOVAs showed that differences in inconclusive rates were not significant for deceptive cases, but were significant for truthful cases [F (2,478) = 3.418, (p = .034)]. CIT/GKT results do not include an inconclusive category, and this technique was not included in the two-way analysis for inconclusive results.

Independent and non-independent PDD techniques

Table 7 shows the criterion accuracy profile of four PDD techniques that are interpreted with decision rules based on an assumption of independent criterion variance among the RQs, along with the criterion accuracy profile of PDD techniques that are interpreted with decision rules based on an assumption of non-independence. Scores were summarized from 14 surveys and experiments involving PDD techniques that are interpreted with an assumption of independent criterion variance among the RQs. These studies included 31 scorers who provided 1,194 numerical scores for 1,008 confirmed exams, including 562 numerical scores for 468 confirmed deceptive examinations and 632 numerical scores for 540 confirmed truthful examinations. Excluding outlier results, scores from 24 surveys and experiments were summarized to describe the criterion validity of PDD techniques for which the results are interpreted with decision rules based on an assumption of non-independence of the criterion variance of the RQs. These studies included 210 scorers who provided 8,975 numerical scores for 1,791 confirmed exams, including 4,612 numerical scores for 933 confirmed deceptive examinations and 4,363 numerical scores for 858 confirmed truthful examinations.

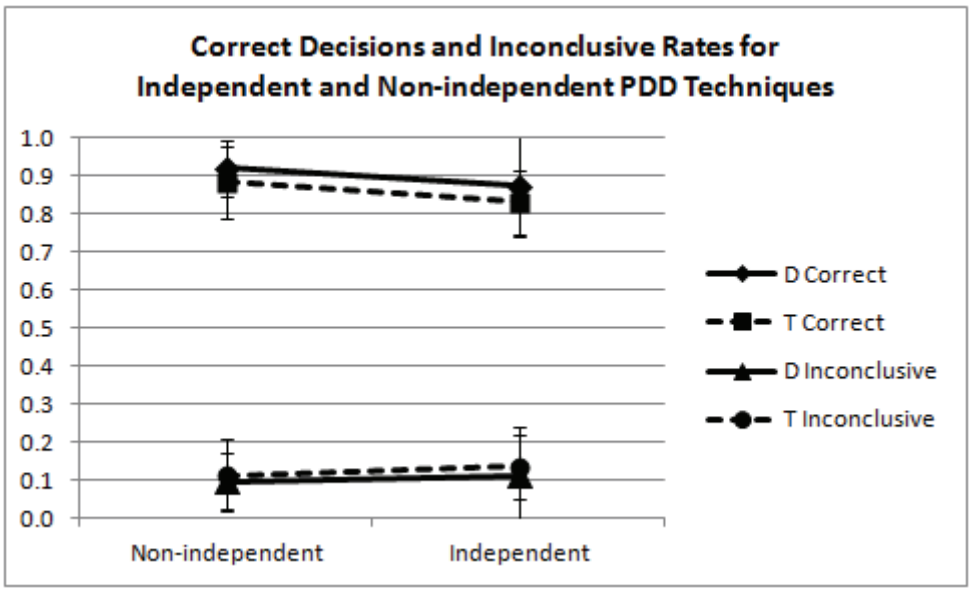
Excluding outlier results, comparison question techniques intended for event-specific (single issue) diagnostic testing, in which the criterion variance of multiple relevant questions is assumed to be non-independent, produced an aggregated decision accuracy rate of .890 (.829 - .951), with a combined inconclusive rate of .110 (.047 - .173). Comparison question PDD techniques designed to be interpreted with the assumption of independence of the criterion variance of multiple relevant questions produced an aggregated decision accuracy rate of .850 (.773 - .926) with a combined inconclusive rate of .125 (.068 - .183). The unweighted average of accuracy for independent and non-independent PDD techniques, excluding outlier results, produced a decision accuracy level of .869 (.798 - .940) with an inconclusive rate of .128 (.068 - .187), as shown in Table 7.

Table 7. Criterion accuracy profile of independent and non-independent PDD techniques.

	Criterion Independent PDD Techniques	Non-independent PDD Techniques	Non-independent Techniques with Outlier Results
Number of Techniques	4	7	9
Number of Studies	14	24	30
N Deceptive	468	933	1,202
N Truthful	540	858	1,087
Total N	1,008	1,791	2,289
Number Scorers	31	210	225
N of Deceptive Scores	562	4,612	4,881
N of Truthful Scores	632	4,363	4,592
Total Scores	1,194	8,975	9,473
Percent Correct	.850 (.039) {.773 to .926}	.890 (.031) {.829 to .951}	.896 (.030) {.837 to .955}
Inconclusive	.125 (.029) {.068 to .183}	.110 (.032) {.047 to .173}	.106 (.031) {.044 to .167}
Sensitivity	.771 (.072) {.630 to .911}	.833 (.052) {.731 to .934}	.840 (.050) {.743 to .938}
Specificity	.719 (.047) {.626 to .811}	.765 (.061) {.646 to .884}	.775 (.059) {.658 to .891}
FN Errors	.113 (.058) {.001 to .226}	.078 (.033) {.013 to .143}	.074 (.032) {.011 to .138}
FP Errors	.144 (.039) {.066 to .221}	.115 (.042) {.032 to .197}	.109 (.041) {.029 to .189}
D Inc	.112 (.051) {.013 to .212}	.093 (.041) {.012 to .174}	.089 (.039) {.011 to .166}
T Inc	.136 (.031) {.076 to .196}	.127 (.049) {.030 to .223}	.122 (.049) {.027 to .218}
PPV	.828 (.059) {.712 to .943}	.886 (.041) {.806 to .967}	.893 (.039) {.816 to .969}
NPV	.878 (.049) {.782 to .973}	.906 (.044) {.820 to .993}	.910 (.043) {.826 to .995}
D Correct	.873 (.066) {.744 to .999}	.915 (.037) {.842 to .988}	.919 (.036) {.849 to .989}
T Correct	.831 (.043) {.746 to .915}	.866 (.049) {.770 to .962}	.873 (.047) {.780 to .965}

Figure 22 shows the mean and statistical confidence intervals for correct decisions and inconclusive rates for PDD techniques interpreted with decision rules based on assumptions of independent and non-independent criterion variance, excluding outlier results. Two-way ANOVAs, criterion state x independence, excluding outlier results, showed a significant interaction for correct decisions [F (1,10165) = 2656.637, (p < .001)] and inconclusive results [F (1,10165) = 806.839, (p < .001)]. However, post-hoc ANOVAs showed there was no significant one-way difference in the proportion of correct decisions or inconclusive results for criterion deceptive cases or criterion truthful cases.

Figure 22. Mean and confidence interval plot for criterion independent and non-independent (excluding outlier results) PDD techniques.



Discussion

Fourteen PDD techniques (shown in Figure 17) meet the requirements of the APA 2012 standards for test validation. These techniques are supported by published descriptions of the protocol for test administration and test data analysis, using instrumentation representative of that used in field practice, and by published and replicated empirical support for the criterion accuracy of a published method for test data analysis.

Five PDD techniques have published evidence of validity that meets the APA 2012 requirements for evidentiary testing, including unweighted decision accuracy over .900 along with inconclusive rates under .200. These five PDD techniques are, in alphabetical order; the Federal You-Phase technique scored with the Empirical Scoring System (ESS), the IZCT, the MQTZCT, the Utah ZCT (including PLT, DLT, and CPC-RCMP variants) scored with the Utah numerical scoring system, and any variant of an event-specific three question ZCT scored with the ESS. Statistical analysis revealed two statistical outliers, the IZCT and the MQTZCT. Two-way ANOVAs indicated that there were no significant differences between the other evidentiary techniques when the outlier results were not included.

Five other PDD techniques were found to produce criterion accuracy that meets the APA 2012 standard requirements for paired-testing, with unweighted decision accuracy over .860 along with inconclusive rates under .200. These PDD techniques are, in alphabetical order, the AFMGQT when scored with the ESS, the Backster You-Phase technique scored with the Backster numerical scoring system, the Federal You-Phase

technique scored with the Federal seven-position TDA model, and the Federal ZCT scored with the Federal seven-position TDA model, and the Federal ZCT scored with the seven-position TDA model and interpreted with evidentiary decision rules. Although this level of validation is intended to serve the needs for criterion accuracy in paired-testing situations, the majority of PDD examinations are not intended for paired testing or evidentiary use in courtroom settings. It is therefore inevitable that many field examinations may be conducted with the PDD techniques in this list though not intended for use in courtroom settings. Although a significant interaction was observed between criterion status and PDD techniques, indicating that different PDD techniques may provide subtle differences in accuracy with criterion deceptive and criterion truthful cases, none of the one-way main effects was significant for decision accuracy or inconclusive results among the deceptive or truthful cases. The present evidence does not support a conclusion that any of these techniques provides accuracy that is different from the other techniques, and instead suggests this group of PDD techniques provides overall criterion accuracy of similar effectiveness.

Four additional PDD techniques were found to satisfy the APA 2012 standard requirements for investigative testing. These four PDD techniques are, in alphabetical order, the CIT/GKT, the DLST/TES scored with the seven-position TDA method, the DLST/TES scored with the ESS, and the AFMGQT[50] when scored with the seven-position TDA method. Although there may be subtle differences in the accuracy of these techniques with criterion deceptive and criterion truthful cases, there were no significant main effect differences for decision accuracy or inconclusive results among the deceptive or truthful cases. These results suggest that this group of PDD techniques provides overall criterion accuracy of similar effectiveness, and the present evidence does not support a conclusion that any of these techniques has accuracy different from the other techniques.

Outlier results

Two outliers were identified: the IZCT and the MQTZCT. Research for both of these techniques reported near-perfect accuracy, and these results were found to be statistical outliers to the distribution of results predicted by all other studies on all other techniques, including the other evidentiary techniques in which these two studies are grouped. These two techniques rely on support from the most problematic research of all studies included in the meta-analysis.

One of the two studies included in support of the IZCT (Gordon et al., 2005; also described by Mohamed et al., 2006) is a very small study described in one publication as a non-blind pilot study. The use of pilot studies to answer questions about criterion accuracy is troublesome. Additionally, both reports indicated that one of the 12 participants in the Gordon et al. (2005) study, a programmed innocent participant, made a false-confession to the examiner, also the primary author, during the pre-test interview. That participant was removed from the experiment, which illuminates the non-blind study design. A false confession in field PDD programs would not be immediately distinguishable from an authentic confession. In field polygraph programs a pre-test confession would be viewed as a practical and successful form of resolution of the matter under investigation. Authentic confessions are regarded as PDD successes, and it is therefore necessary to regard false-confessions as problems. In a field situation, it would only be later, when additional evidence is available, that the confession would be identified as an error and would be viewed as problematic.

Inclusion of this error into the study results would have resulted in a false-positive (i.e., false-confession) rate of .167 and less than perfect test accuracy. Instead, the results from the Gordon et al. (2005) study were provided without the false confession,

[50] Because the Utah MGQT and the LEPET are structurally virtually identical to the AFMGQT, and use the same scoring regimen, it is reasonable to generalize the AFMGQT validation findings to these two techniques.

along with a reported decision accuracy rate of 1.000. It is possible that neither the reported decision accuracy rate of 1.000 nor the false confession rate of .167 is representative of IZCT performance in field settings. An argument could be offered that since this was a non-blind pilot study, which was not designed to serve as a criterion accuracy study, removing errors from the reported study result was justified. Pilot studies like this help guide decisions about the funding and design of more rigorous research into areas such as fMRI or other methods for lie detection. However, the selective exclusion of unfavorable data from a study of criterion accuracy requires strong justification.

An additional concern regarding the evidence supporting the IZCT is the fact that the sampling distributions from the three included studies differ significantly. Significant differences are the result of several possible conditions, including: 1) the samples were selected from different populations, 2) the IZCT was administered differently to the different study samples, or 3) the study samples were scored and interpreted with a different application of the TDA rules. It is also possible that the observed significant differences are the result of a highly selective sampling methodology, in which examinations are included based on the examiner's or investigators judgment of good or confident results, such as would occur in the context of a direct admission regarding the investigation targets. Over-reliance on confession confirmation could have the effect of systematically excluding both false-negative and false-positive error cases, for which no confession would be likely to be obtained.

Regardless of the reason, deceptive and truthful scores were expressed in significantly different ways in the three different studies on the IZCT. The meaning of these significant differences to this meta-analysis is that the included studies appear to be based on samples that are not representative of each other, and it is unknown whether one or more of the studies is not representative of the population of examinees.

A third problematic concern with the IZCT is that none of the published studies included any reliability statistics and calculations of interrater reliability could not be completed from the available data. The absence of reliability statistics does not allow estimates of the generalizability of the study results to results that would be obtained from other examiners or other scorers. Coupled with the significant interaction effects between sampling distribution and case status, the present evidence is insufficient to support the notion that other practitioners would obtain scores or results similar to those reported in the published studies.

Studies supporting the generalizability of the MQTZCT, the other statistical outlier, are limited by some interesting and unique factors. First, the developer of the MQTZCT and previous authors themselves seem to have cautioned against the generalizability of this technique by emphasizing the need for intensive and specialized training available only from practitioners of the method. Indeed they have asserted that the complexity of the technique, and its related psychological hypotheses, are such that other trained PDD examiners should not reasonably expect to learn or properly execute the MQTZCT based on information available in the published sources. An emphasis on strict compliance with a complex and precise systems of many rules gives the impression that the technique should be regarded as fragile, non-robust, and easily disrupted by even slight departures from stipulated procedures.

A second, equally important concern involves the fact that a significant interaction was found between sampling distribution and case status. Although one-way differences were not significant within the deceptive or truthful groups, the significant interaction effect indicates that the scores of criterion deceptive and criterion truthful cases are expressed or interpreted in different ways within the sampling distributions of the three included studies on the MQTZCT. In other words, the data are not congruent even among the studies used to support the MQTZCT. This significant interaction suggests the possibility that the included studies are based on samples that are not representative of each other. It is unknown whether one or more of the studies is not representative of the population of all examinees, reducing our confidence in the potential for generalizability of the reported results.

A third concern involving the MQTZCT is that the reported reliability coefficient of .990 was published in the Matte and Reuss (1989) reprint of the dissertation published in the journal *Polygraph*, but cannot be located in the original dissertation study for the no longer extant Columbia Pacific University. This is both unfortunate and concerning because the unprecedented high rate of inter-scorer agreement is unexpected given the purported complexity of the method.

A final confound to the generalizability of the results of the included studies on the MQTZCT is that the data provided to the committee initially included numerical scores for only those cases for which the scorers achieved the correct result. Data available to the ad-hoc committee did not initially include numerical scores for those cases for which the scorers achieved erroneous or inconclusive results. Missing scores were later provided to the committee for both the Mangan, Armitage and Adams (2008) and Shurani, Stein and Brand (2009) studies. However, the resulting sampling distribution means, calculated with the missing scores, were different from those reported for both studies. Because of these discrepancies, the statistical analysis was not re-calculated with the missing scores, and this analysis reflects the mean scores as reported by Mangan, Armitage and Adams (2008), and by Shurani and Chavez (2009). Field data are always a combination of diagnostic (i.e., controlled or explained) variance and error variance (i.e., uncontrolled or unexplained variance). The sampling means reported in the Mangan, Armitage and Adams (2008) and Shurani, Stein and Brand (2009) studies are systematically devoid of error variance. Given that a significant interaction effect was observed between sampling distribution and case status, the present evidence is insufficient to support the generalizability of the reported study results.

Possible mediators of these outlier results include the possibility that these techniques are simply superior to others. The role of proprietary, personal and financial interests, including business relationships between technique developers and principal investigators, cannot be overlooked, however, and the serious methodological and empirical confounds surrounding the supporting research undermines confidence in the study results and reported accuracy of these techniques. From a scientific perspective, even well designed research generated by advocates of a method who have a vested interest in the outcome, and who act as participants and authors of the study report, does not have the compelling power of research not so encumbered by these potentially compromising factors.[51]

Regardless of what factors contribute to these exceptional results, the confounds associated with the supporting studies undermines confidence that they represent the true accuracies. Expectations that these outlier results will generalize to field settings should be delayed until more complete independent replication studies and extended analysis are completed.

Criterion accuracy

Excluding outliers, the aggregated unweighted accuracy[52] of all PDD techniques was .869 (.798 - .940), with an unweighted inconclusive rate of .128 (.068 - .187). All 14 PDD techniques included in this meta-analysis produced unweighted decision accuracy levels that were significantly greater than chance. Excluding outliers, there were no significant one-way differences in the

[51] Questions may arise as to why these studies and techniques were included in the meta-analysis after identifying so many serious confounds. The techniques were ultimately included in the meta-analysis because they met the more general requirements outlined in the APA Standards of Practice. It was also determined that the meta-analysis is more complete, and therefore more helpful and informative to interested readers, with the inclusion of these studies and techniques.

[52] The unweighted average was considered to be a more conservative and realistic calculation of the overall accuracy of all PDD examination techniques. Calculation of the weighted average, or the simple proportion of correct decisions, often results in higher statistical findings that are less robust against differences in base-rates and therefore less generalizable.

unweighted decision accuracy of any of the 14 PDD techniques, and no significant one-way differences in correct decisions, inconclusive results, or errors for criterion deceptive or criterion truthful cases. Neither were there any significant differences in the aggregated criterion accuracy of PDD techniques at the evidentiary, paired-testing, and investigative levels. Some practical differences were observed in the criterion accuracy profiles of these techniques. All five techniques included at the evidentiary level produced statistically significant effect sizes for both test sensitivity to deception and test specificity to truth-telling.

At the paired testing level all five techniques also produced test sensitivity to deception that was significantly greater than chance, though only two of these techniques, the Federal ZCT scored with the seven-position evidentiary rules, and the AFMGQT scored with the ESS, produced test specificity to truth-telling that was significantly greater than chance. Specificity to truth-telling was not significantly greater than chance for the Backster You-Phase technique with Backster scoring, Federal ZCT with seven-position scoring or Federal You-Phase with seven-position scoring.

For investigative techniques, all four techniques produced test sensitivity to deception that was significantly greater than chance. Specificity to truth-telling was significantly greater than chance for the CIT, and the DLST/TES format when scored with both the seven-position and ESS models, but was not significantly greater than chance for the AFMGQT when scored with the seven-position model.

Excluding outlier results, published and replicated empirical evidence for seven CQT formats, intended for event-specific diagnostic testing for which the results are interpreted using decision rules based on the assumption of non-independence of the criterion variance of the RQs produced an aggregated unweighted accuracy rate of .890 (.829 - .951) along with an inconclusive rate of .110 (.047 - .173). These techniques are, in alphabetical order, the AFMGQT when scored with the ESS, the Backster You-Phase technique scored with the Backster numerical scoring system, the Federal You-Phase technique scored with the Federal seven-position TDA model, the Federal You-Phase technique scored with the ESS, the Federal ZCT scored with the Federal seven-position TDA model, the Federal ZCT scored with the seven-position TDA model and interpreted with evidentiary decision rules, the Utah ZCT (including PLT, DLT, and CPC-RCMP variants) scored with the Utah numerical scoring system, and any variant of an event-specific three-question ZCT scored with the ESS.

Published and replicated empirical evidence exists for four PDD techniques that are interpreted with decision rules based on the assumption of independent criterion variance among the RQs. These techniques produced an aggregated unweighted accuracy level of .850 (.773 - .926) with an inconclusive rate of .125 (.068 - .183). In alphabetical order, these techniques are: the DLST/TES scored with the seven-position TDA method, the DLST/TES scored with the ESS, the AFMGQT when scored with the seven-position TDA method, and the AFMGQT when scored with the ESS.

Despite these observed and practical differences, excluding outlier results, no significant differences were found for decision accuracy or inconclusive results among the PDD techniques that satisfy the requirements of the APA 2012 standards. Similarly no significant differences were found for decision accuracy or inconclusive results for PDD techniques interpreted with the assumption of independence or non-independence among the RQs.

Not all techniques reviewed possessed sufficient empirical support to meet the APA standards for inclusion. Some named PDD techniques were found to lack any published evidence of support that could be used to calculate the sampling distributions, reliability and criterion accuracy profiles needed for inclusion in this meta-analysis; these are listed in Appendix F. Appendix G provides a summary of published studies that could not be included in the meta-analysis. Appendix H contains a description of techniques for which there exists a single un-replicated study that met the requirements for inclusion in this meta-analysis. These techniques could not be included in the meta-analysis as the APA Standard requires a minimum of two

published studies. Appendix I lists those PDD techniques found to have published and replicated evidence of support, but the reported criterion accuracy did not satisfy the validity requirements of the APA 2012 standards.

PDD techniques that make use of the three-position TDA model are not included in the meta-analysis and are therefore not included in Table 1. The criterion accuracy profiles of PDD techniques that make use of the three-position TDA model are shown in Appendix I. The unweighted decision accuracies were significant for all of the techniques based on three-position TDA methods, but not equal for the deceptive and truthful cases. All techniques that employed three-position TDA methods consistently exceeded the 2012 limit for inconclusive decisions (20%). Because criterion accuracy rates for techniques with three-position TDA did not differ significantly from seven-position criterion accuracy, an initial analysis with the three-position TDA method may be considered acceptable if inconclusive results are resolved via subsequent analysis with a TDA method that provides both accuracy and inconclusive rates that meet the requirements of the APA 2012 standards.

Some readers will note that two versions exist for the AFMGQT with minor structural differences between them.[53] There is no evidence to suggest that the performance of one version is superior to the other. Considering that rigorous and replicated evidence would be required to reject a null-hypothesis that the differences are meaningless, and considering that the included studies include a mixture of both AFMGQT versions, these results are provided as generalizable to both versions of the AFMGQT.

Two widely used and recognizable techniques, the LEPET and the Utah MGQT (four-question version of the Utah technique), were not included in the meta-analysis because no published studies could be located in support of these techniques. However, both of these PDD techniques are structurally nearly identical to the AFMGQT. We can find no reason why validation data for AFMGQT cannot be generalized to these techniques if scored with the same TDA methods.

Comparison With Previous Systematic Reviews

We did not test the level of significance of the difference between the present accuracy estimations with those of the OTA (1983), though it can be easily seen that the .847 mean accuracy rate of field studies is outside the 95% confidence interval (.865 to .977) for PDD techniques that meet the APA 2012 requirements for evidentiary testing. This observed difference is most likely due to the exclusion of results from studies that do not conform to recognizable field practices, and to the exclusion of results of PDD techniques that do not produce satisfactory results according to the APA 2012 standards of practice. There is little justification for use in field practice, and therefore little justification for inclusion into accuracy estimations, of PDD techniques that have been supplanted by more effective methods. Inclusion of arcane or substandard methods into accuracy estimation would be the equivalent of attempting to answer an automobile industry question regarding corporate fuel economy while including all makes and models from the 1960s and 1970s gas-guzzling era into calculations of present-day economy. Techniques which produce substandard and unsatisfactory criterion accuracy were therefore excluded from the meta-analysis.

These results are consistent with the results of Honts and Peterson (1997), Raskin and Podlesny (1979), Abrams (1977; 1989), and Ansley's (1990) findings regarding blind evaluation of PDD test data. Results of this meta-analysis are also consistent with the results of the more recent National Research Council (2003) who reported an accuracy rate of laboratory studies as .860 along with an aggregated rate of .890 for field studies, using studies that met their selection criteria. Because the present analysis includes only

[53] The AFMGQT is used in both multi-facet investigations of known incidents and multi-issue screening contexts. Both types of exams, muti-facet and multi-issue, are interpreted with decision rules based on the assumption of independent criterion variance among the RQs.

techniques as they are documented and used in field settings, we suggest that the present results provide a more helpful and practical answer to PDD professionals, program managers, and professional consumers of PDD results who are faced with the need to make evidence-based decisions regarding the selection and field use of presently available PDD techniques.

These results are more conservative than those reported by Ansley (1983; 1990) and those of Abrams (1973), which warrants further discussion. It is unlikely that the PDD test has become less accurate during the last three decades. A more realistic possibility is that the samples included in the early literature reviews by Ansley were more vulnerable to overestimation of test accuracy as a result of sample selection methodology. Ansley (1990) stated that court decisions and evidence are sometimes unreliable, and expressed a preference for confession confirmation of PDD examination results. Over-emphasis on confession confirmation includes the potential for unintended systematic exclusion of false-negative and false-positive errors, both of which conditions are unlikely to lead to a confession, and therefore the confirmation criterion. Confessions themselves are the result of a non-random decision to pursue further discussion with and disclosure from the examinee. If the decision to pursue a confession is based in part on the results of a polygraph exam, then confirmation via confession is non-independent from the test result and therefore self-fulfilling.

The impact of these methodological issues could be sampling distributions that inflate PDD test accuracy.[54] This phenomenon may not be limited to confirmation via confession; all field samples that are selected through the availability and quality of confirmation data are potentially non-random and non-representative. This same concern, regarding the non-independence of confirmation data, applies to investigation results and judicial outcomes that are based in part on the information resulting from a polygraph exam. It is possible that this phenomenon underlies the general trend in the literature in which the results of PDD field studies have generally outperformed the results of laboratory experiments.[55] Despite potential or observed sampling differences between field and laboratory studies, the NRC (2003) found no significant differences between the results of high quality field and laboratory studies.

Results of this meta-analysis are consistent with the systematic review of Crewson (2003) regarding the accuracy of diagnostic polygraphs. However, these results depart from Crewson's conclusion regarding screening polygraphs. The accuracy rate found for screening polygraphs in this meta-analysis was higher than that reported by Crewson, and the difference is statistically significant (t [1008] = .002). While the exact cause of this difference cannot be known from the present data, we note that the studies and techniques used by Crewson could not be included in this meta-analysis. Four of the screening studies reported by Crewson involved the Relevant-Irrelevant technique (Ansley, 1989; Brownlie, Johnson & Knill, 1997; Honts & Amato, 1999; Jayne, 1989), and the remaining study involved the Reid Technique. Included studies pertaining to criterion independent screening polygraphs were not available at the time of Crewson's review of the published scientific literature.

[54] A possible example of this phenomenon can be seen in Mangan et al., (2008) who reported the results of a survey of the confession-confirmed test results of one experienced examiner. The reported results were 100% accurate, a finding in accord with what would be expected to arise from a confession-based selection bias.

[55] An alternative explanation would hold that the difference is the result of differences in ecological and external validity of the test circumstances. These hypotheses have not been thoroughly evaluated and it would be unwise to attempt to reach any conclusion with the current state of understanding. The NRC (2003) reported that this trend is not inconsistent with experience in other fields of testing and science and should be the focus of future research.

Moderators and Mediators

There was no indication that the study results were a function of, or influenced by, sample sizes. Results were not coded for examinee characteristics, including age, gender, ethnicity, culture, education, or socio-economic status, nor were the studies coded for their quality or methodology. Sample results based on examinees who were subject to some form of experimental manipulation (e.g., medications, fatigue, chronic physical or chronic mental health problems, level of functioning, countermeasure training or instructions, etc.) were not included, and these factors were not evaluated.

Excluding outliers, no significant differences were found in the criterion accuracy of PDD techniques suitable for evidentiary testing, paired testing, and investigative testing. This suggests that these categorical distinctions are arbitrary and therefore meaningless in a scientific sense. However, the value of standardized requirements for test precision becomes clearer when considering policy decisions that emphasize or require the use of evidence-based methods and restrict the use of un-validated or experimental methods. The scientific value of categorical distinctions becomes more obvious when considering the difficulty in answering questions about the scientific accuracy, and the complications that result from the inclusion into accuracy estimates of less accurate and arcane methods that have been supplanted or replaced by more effect modern alternatives. Ethically, it is difficult to imagine some justification, when decisions affect individual lives, community safety and national security, for the use of methods which the scientific evidence has shown to be sub-optimal or sub-standard.

Comparison of accuracy rates for PDD techniques interpreted with the assumption of criterion independence versus non-independence showed no significant differences in decision accuracy. However, a significant interaction effect for inconclusive results suggests there may be subtle differences in inconclusive rates for these types of exams. This would seem to suggest that the selection of different examination strategies, involving independent or non-independent RQs, is a practical matter that should be determined by the needs of the testing circumstances.

Ancillary Analysis

One ancillary analysis was completed. Results were calculated for CQT formats with the exclusion of those studies that did not satisfy a more rigorous set of selection criteria. First, PDD techniques were excluded from the ancillary analysis if both test sensitivity to deception and test specificity to truth-telling were not both statistically significantly greater than chance. This resulted in the exclusion of the AFMGQT, Federal ZCT, and Federal You-Phase techniques when these are scored with the seven-position TDA model, in addition to the Backster You-Phase technique. Statistical outliers, not accounted for by the available evidence, were also excluded. This resulted in the exclusion of the IZCT and MQTZCT and several studies that were seriously confounded. Exclusion of techniques for which there is no published statistics describing test reliability also resulted in the exclusion of the IZCT. Similarly, PDD techniques and studies were excluded if there were significant interaction or main effect differences between the sampling distributions, indicating that the sample distributions are not representative of each other. This also resulted in the exclusion of the IZCT and MQTZCT. Studies were also excluded if statistical descriptions of the sampling distributions were not available or could not be calculated from the available data. This resulted in the removal of two studies on the DLST (Research Division Staff, 1995a; 1995b), one study on the AFMGQT (Senter, Waller & Krapohl, 2008), and two studies on the MQTZCT (Shurani, Stein & Brand, 2009; Shurani, 2011).

CQT formats retained for ancillary analysis produced a combined decision accuracy rate of .898 (.840 - .955) and an inconclusive rate of .092 (.033 - .150)[56] for PDD techniques interpreted with decision rules based on an assumption of

[56] Calculated as the weighted average of unweighted decision accuracy and the unweighted inconclusive rate.

non- independence of the criterion variance of the RQs. PDD techniques interpreted with decision rules based on an assumption of independent criterion variance produced a decision accuracy rate of .857 (.782 - .932) and an inconclusive rate of .117 (.058 - .177). The aggregated decision accuracy rate for all studies and PDD techniques included in the ancillary analysis was .883 (.817 - .950) with an inconclusive rate of .116 (.056 - .175).[57] Two-way ANOVAs showed that neither main effect nor interactions were significant when comparing the decision accuracy for the ancillary analysis with that of the entire meta-analysis. The interaction effect was significant for inconclusive results [$F(1,1992) = 17.335$, $(p < .001)$]. Inconclusive rates were slightly higher for truthful cases with non-independent techniques, and slightly higher for deceptive cases with criterion independent PDD techniques. Post-hoc ANOVAs showed that none of the one-way differences were significant, indicating that these small differences are unlikely to be noticed by field examiners.

One-way ANOVAs showed that the results of the ancillary analysis did not differ significantly from the results of the entire meta-analysis for correct decisions [$F(1,5471) = 0.08$, $(p = 0.777)$] or inconclusive results [$F(1,5471) = 0.08$, $(p = 0.777)$]. This indicates that the use of more rigorous study selection requirements would be unlikely to produce meta-analytic results that differ from the results of this study.

Limitations

Two obvious limitations pertain to this analysis. First, studies were not coded for field and laboratory studies and no attempt was made to investigate any effects from differences in study design. Instead, field and laboratory results were included with equal consideration and the results of all studies were combined regardless of design. Secondly, there was no attempt to investigate decision accuracy at the level of the individual questions for any of the included PDD techniques. Related to this second confound is the fact that the results of studies involving the DLST/TES and AFMGQT PDD were achieved using decision rules that are based on an assumption of criterion independence among the RQs. Generalizability of the results of this meta-analysis may depend, in part, on the correctness of this assumption.

Some of the included studies are impaired by obvious research confounds, the most noticeable of which is that some samples were selected with an emphasis on examinee confession as a central feature of the criterion. Another important confound, observed in some of the included studies, was that the primary author was also the developer of a PDD technique for which there exists some form of proprietary, or financial interest. Indeed, it would appear that one of the markers for these kinds of studies (and typical of advocacy research elsewhere) is that the reported near-perfect accuracy demonstrations are statistical outliers to the distribution of results from less confounded studies.

The absence of critical information and critical commentary in some included study reports gives the impression of a *file-drawer* bias in which less than favorable results are not submitted for publication. Another version of this problem seems to have occurred in the context of this meta-analysis, in which some of the study data initially provided to the committee, and some of the published sampling means, included only those results for which the scorers achieved the correct results, initially withholding the results of inconclusive and error cases. The result of this is that published sampling means for some studies are systematically devoid of error or uncontrolled variance and must therefore be considered not generalizable.

Confounds related to individual studies can complicate the meaning and interpretation of the results of the meta-analysis. These concerns represent an example of the value and need for scientific rigor and independence when evaluating the effectiveness of PDD and lie detection methods. Study selection and inclusion rules for the meta-analysis were intended to be as

[57] Calculated as the unweighted average of all studies included in the ancillary analysis.

inclusive as possible, yet maintain a level of scientific rigor. To reduce the impact of these confounds on the meta-analysis, aggregated results have been provided both with and without outlier results.

Another confounding issue with some studies is that the level of education, training and knowledge regarding psychological, physiological and testing principles may be significantly greater for participating examiners than for most field examiners. Most blind scoring studies of PDD accuracy involve highly experienced experts. Studies of the ESS have been an exception to this trend, making use of inexperienced examiners and scorers, and recent studies by Honts and his colleagues have involved students trained to collect the study data.

Meta-analysis always involves the imposition of study selection rules, and it is always possible that a meta-analysis based on a different set of inclusion criteria would lead to different results. All studies in this analysis were regarded equally if they met the publication requirements and provided sufficient information to evaluate the criterion accuracy and reliability or generalizability of the study results. Qualitative requirements for inclusion in this study pertained only to whether included studies satisfactorily represented field testing instrumentation and components, and satisfactorily represented a PDD technique for which a published description exists for the test question sequence and method for test data analysis. Meta-analytic weighting values were assigned according to the sample size and number of scorers for each study, though there were no obvious effects related to sample size. Although previous statistical analyses have not identified any significant differences in the results of field and laboratory studies, it is possible that meta-analytic results would be slightly different if the included studies were coded and weighted for other dimensions, including study quality, design, sampling methodology, proprietary interests, or inclusion of the primary author as a study participant.

Another limitation of this analysis is that none of the included studies involved juvenile examinees. As a result, a conservative evaluation of the present results would suggest that our present knowledge-base can be considered applicable only to physically and mentally healthy adults of normal functional characteristics. A more generous interpretation would recognize that there is little difference between adults and older juveniles in terms of physiology as measured or utilized by modern polygraph sensors and little difference in the psychological bases for polygraph reactions between adults and older, developmentally mature, juveniles. Generalizing these results to persons who are known outliers compared to the expected distribution of persons from the normative population (i.e., persons whose functional characteristics are outside the normal range) should be done with great caution.

Some of the included studies lacked complete information, though it was possible to calculate the reliability, sampling distributions, and the dimensional profile of criterion accuracy from raw data that was provided to the ad hoc committee. Sample data were not available for some studies, notably those from the U.S. Government. Those studies include the development and validation studies on the TES (DLST) and the AFMGQT techniques. These studies did report reliability and accuracy data that was sufficient to include them in the meta-analysis, and all of these studies have been replicated independently.

An obvious limitation of this meta-analysis is that it did not include the results of computerized scoring algorithms.

One other issue deserves mention. The principal investigator for this meta-analysis was also the primary author of a number of included studies.[58] The committee was aware that his research was significantly, and at times solely, involved in studies that were proffered to validate some of the included techniques. Some of these techniques

[58] Mr. Nelson has no financial, proprietary or personal interest in any of the PDD techniques or methodologies included in this meta-analysis.

would not have been included without these studies. As such, it was the responsibility of the committee to weigh his judgments against factors that may have diminished his independence. This was of vital concern, inasmuch as a bias in study selection or data analysis would seriously compromise the integrity of the final report. Upon closer scrutiny, reservations regarding these two potentially conflicting roles (principal author of the meta-analysis and author of studies included in the meta-analysis) were mitigated by the lack of any apparent personal interests in the outcomes of those studies and his limited participatory role in any study (never having conducted or scored any of the examinations). His published or pending studies did not reveal any discernible pattern of preference for or against particular polygraph techniques. Finally, all decisions for study inclusion were made collectively among the committee members based on the merits of the research: no single committee member had absolute authority to exclude or include a given study. While reasonable individual opinions may differ in some parts of this report, the committee took deliberate care to ensure that personal interests among those in the committee would not be cause for criticism of the report. Full disclosure of the relationship between the principal author of this report and his research is provided here to meet the standard ethical obligation in scientific reports.

Recommendations

Because no significant differences were found among the 14 PDD techniques included in this meta-analysis, no attempt should be made to describe these techniques in terms of a rank order regarding effectiveness. Available evidence does not support any PDD technique as superior to others. Attempts at establishing any hierarchy of efficacy are therefore unwarranted. Instead, less attention should be given to named PDD techniques and meaningless differences in PDD test formats. More emphasis should be given to test construction details for which there is replicated evidence of their contribution to criterion accuracy. More emphasis should be given to the important practical and decision theoretic differences in PDD techniques for which the RQs are interpreted as independent or non-independent.

One practical area of needed research involves the generalizability of normative data and accuracy estimates for PDD techniques interpreted with the assumption of criterion independence, including both multi-facet and multi-issue exams. Another practical area of needed research involves the use of DLCs with additional PDD test formats.

Continued research and improvement is needed for all PDD techniques, and these improvements should be fully integrated into both training and field practices. Additional studies should be completed to increase the knowledge base regarding moderator variables such as examinee characteristics (e.g., juveniles, older persons, persons with mental illness, and persons with medical health complications) in addition to crime details or characteristics that lead to the most effective use of PDD examinations. Additional research is needed in screening examinations, including studies pertaining to the decision theoretic complexities inherent to examinations constructed with multiple independent targets. Researchers should continue to increase their use of Monte Carlo models and other statistical methods that can be used to provide answers to complex research problems that are difficult to investigate through other methods. Results of Monte Carlo studies should be compared to those from live experiments from both field and laboratory settings.

A number of mediator variables have been suggested as having a significant effect on the accuracy of the PDD exam, and some of these involve complex psychological and linguistic assumptions that may or may not be fully testable. Untestable hypotheses should be discarded in favor of testable ones, and additional research should be conducted to understand the merits of procedural and structural hypothesis that have been suggested as related to test accuracy. Evidence from scientific studies should become a standing expectation, and developers and practitioners of PDD exams should resist the temptation to include authority-based and anecdotal theories which have not been tested.

Increased research standards are needed, including requirements for transparency and statements of interest from

all authors and participants. More importantly, because research on PDD test effectiveness is a process of testing the test, primary authors should be required to refrain from also functioning as a study participant. It is especially important, when the primary author is also the PDD technique developer or lacks independence due to a financial or business relationship with the developer, that the research data and methodology be subjected to rigorous objective and external review before a profession or the community is encouraged to rely on the research results.

Researchers should be required to provide statistical descriptions of the sampling distributions. This will facilitate more effective comparison of sampling distributions, and will increase the ability to evaluate and understand the representativeness and generalizability of study results. Just as the results of single un-replicated studies are of little actual value to meta-analytic research, the results of studies that employ a single expert scorer are of little actual value to the profession. Researchers should be encouraged or required to use multiple scorer participants of varied training and experience. Both examiners and examinees should be randomly selected whenever possible. This will increase the ability to study and understand the generalizability of PDD methods.

Some studies included in the meta-analysis are not adequately identified regarding the type of study, and the effect is potentially misleading for the profession. Pilot studies and surveys should be clearly identified as such, and should not be included in future systematic reviews or meta-analysis of criterion accuracy. Criterion studies should also be clearly identified from studies designed to evaluate moderator or mediator variables or questions of construct and causality.

Reliability statistics should be required for all studies, unless precluded by the study design (e.g., computer algorithm or simulation studies). Primary authors should be required to make all raw data and numerical scores available for review and extended analysis.

The results of computerized statistical TDA algorithms should be included in future studies of this type. The use of computers and statistical decision theory is still not common in TDA methods for PDD exams. Instead, TDA methods in field use emphasize manual scoring methods with integer-level and rank-level precision that should be considered blunt and unreliable compared to the precision and reliability that can be obtained via automated measurement and statistical analysis. There is a growing basis of evidence that indicates that computer algorithms can be equally or more effective than manual TDA methods as long as the data are of satisfactory quality. Because PDD examination results may play a decision support role in matters that affect individual lives, community safety and national security, the developers of computer algorithms should be required to provide complete descriptions of the operational procedures, in addition to the evaluation criteria, data transformation and aggregation methodology, normative data, and statistical basis in decision theory, signal detection theory, signal discrimination theory, regression analysis or machine learning.

Continued development and refinement of PDD testing methodologies is needed. PDD testing procedures have changed little over the past decade. Development efforts during this time have focused on improvements to test data analytic methods, including decision rules (Senter, 2003; Senter & Dollins, 2002; 2004; 2008), statistical algorithm development (Nelson, Krapohl & Handler, 2008), numerical transformations (Krapohl, 2010; Nelson, Krapohl & Handler, 2008; Nelson et al., 2011), and an increased use of normative data to calculate error rates and optimal decision cutscores (Krapohl, 2010; Nelson & Handler, 2010; Nelson, Krapohl & Handler, 2008; Nelson et al., 2011). Additional research and more detailed investigation and comparison of numerical transformation models is needed, including seven-position, three-position, ESS, rank-order transformation methods, those of computer algorithms, and the application of these transformations to examinations constructed from independent and non-independent examination targets.

PDD component sensors have changed little for several decades. This may be a mixed blessing. Although critics may point to this as a stagnation of research and development, there is a considerable published knowledge

base supporting and describing the effectiveness of the presently used array of PDD component sensors. It would be premature to abandon that knowledge base in an attempt to satisfy a collective hunger for new methods. Any replacement of the presently used component sensors must be accompanied by published and replicated evidence that the data and information provided by the new sensors is as good as, or better than, the data and information from the presently used sensors. Additionally, the use of new and improved sensors will face a substantial and non-trivial burden of developing and demonstrating the incorporation of new data into new or existing normative data and new or existing structural decision models. Despite these general cautions about the replacement of PDD components and physiological measures, it is also clear that the PDD test remains imperfect and in need of continued advancement. PDD test methods will not be improved without replacing less effective methods and techniques with more effective ones.

To give greater confidence in the effectiveness of the IZCT and the MQTZCT (or any proprietary methods[59]), they should be subject to replication by independent researchers, who did not develop the techniques, have no business relationship with the developers, did not conduct the exams, analyze the data, or report the findings at the time of the exams. All raw data and numerical scores should be made available for extended analysis. If these techniques are inherently superior to others, there should be no great difficulty in confirming this through high-quality independent research.

Finally, this meta-analysis should be repeated at some future time, with the inclusion of new and emerging information. Future meta-analytic studies should code and evaluate for moderators such as examiner characteristics, examinee characteristics, and mediators such as study quality, financial interests, and other possible mediators.

Conclusions

Results of this meta-analysis show that a number of studies are of satisfactory rigorous quality to provide a basis of empirical support describing the generalizability of an array of PDD techniques at criterion accuracy levels that significantly exceed chance expectations. Although somewhat arbitrary, the APA 2012 standards of practice and requirements for test accuracy are helpful to the profession. The goal of professional standards is to promote the use of effective methods, and discourage the use of less effective and unproven methods. Fourteen PDD techniques were found to be supported by multiple published studies and to satisfy the requirements of the APA 2012 Standards of Practice. Normative data are available for each of these 14 PDD techniques. We note that despite the imperfections of the polygraph, the NRC (2003) reported that none of the potential new technologies was ready to replace the polygraph, and this condition appears not to have changed at the present time.

APA standards do not themselves impose qualitative or methodological requirements for scientific evidence,[60] and no quantitative requirements are stated beyond the requirement for two publications that indicate certain levels of precision for examination decisions. Herein rests a potential weakness of the APA standards: a simplistic interpretation of the requirements suggests that anyone could press ink onto

[59] In fairness to the developers of these methods, every "lie detection" method in the past 100 years that was researched by the developer or by an enthusiastic user evaluating his or her own examinations has reported accuracy approaching perfection. It is one of the hallmarks of advocacy research in all fields, not just lie detection. The trend includes Marston's discontinuous blood pressure technique, Summer's Pathometer in the 1930s, MacNitt with the Relevant-Irrelevant technique in the 1940s, Lykken's GKT, Farwell's "Brain Fingerprinting," and the Computer Voice Stress Analyzer. In each case the authors reported stellar accuracy, usually greater than 99%. In all these cases, however, subsequent research was either absent or resulted in accuracies significantly lower than the original reports.

[60] The APA has adopted a standard for research, which can be found online at www.polygraph.org and printed in the journal *Polygraph*.

paper two times in self-published volumes, claiming perfect or near perfect accuracy, and subsequently claim compliance with the APA standards for test validation at the highest categorical level – for evidentiary testing. While this would be viewed by scientific thinkers with some skepticism, the example illustrates the need in meta-analytic research for the definition of more rigorous study inclusion and exclusion criteria. Just as not all evidence is good evidence, not all publications are useful. Publication itself is not an endorsement of fact, and merely indicates that editors and reviewers agreed that the work would be of some interest to the profession.

Nothing in this document should be taken to suggest that we presently know everything we need to know, or everything there is to know, about PDD testing. There is always more to learn and there is always a need for continued research. More information will undoubtedly become available in the future, and it is incumbent on professionals to continue to incorporate practices based on new and improving evidence from high quality scientific studies. To do otherwise is to subject the future of the profession to opinion, which will be vulnerable to personalities, politics, and personal interests. In the strictest sense, PDD techniques for which there is an inadequate basis of published and replicated scientific studies must be considered experimental, regardless of how long they have existed. However, the suggestion of abandoning un-validated, ineffective, or experimental methods that have long been used in field practice is not without controversy.

It can be, and has been, argued that some unstudied or experimental methods may work as well or better than some proven methods, and may provide specialized benefits in certain ways. Conversely, it is also possible that experimental and unproven methods do not work as well as those with evidence of scientific validity. In the worst circumstances, the use of experimental methods could result in otherwise-avoidable adverse consequences. A conservative assessment would suggest that the practice of conducting experimental methods on the public, when effective evidence-based methods are available and easily implemented with no additional costs, may be considered reasonable under some circumstances, but calls for compelling justification and includes ethical requirements for informed consent and notification.

Finally, this meta-analysis should be considered an information resource only, and the results of this study should not be interpreted as APA policy. No attempt should be made to represent or interpret this document or the results of this study as the only or final authority on PDD test validation. Other, equally reasonable, approaches are also possible regarding the evaluation of the scientific literature on PDD testing.

This project was completed with the goal of summarizing the existing published scientific literature regarding PDD techniques and criterion accuracy, and to provide a convenient resource to those who may wish to avoid the burden of reviewing the research literature for themselves. Every effort has been extended to not only provide conclusions, but also in-depth explanations that underlie those conclusions so that readers can better understand their basis.

The information herein is provided to the APA Board to advise its professional membership of the strength of validation of PDD techniques in present use. This information is intended only to help PDD professionals make informed decisions regarding the selection of PDD techniques for use in field settings. It may also assist program administrators, policy makers, and courts to make evidence-based decisions about the informational value of PDD test results in general. Nothing should prevent the use of any PDD technique that is supported by scientific research that demonstrates an accuracy rate significantly greater than chance, so long as that use is compliant with the requirements of local laws, regulations, and enforceable standards.

How to cite this document:

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. [Electronic version] Retrieved <DATE>, from http://www .polygraph.org.

References

* *indicates studies that were included in the meta-analysis.*

√*indicates studies that were cited only in the appendices.*

Abrams, S. (1973). Polygraph validity and reliability: A review. *Journal of Forensic Sciences, 18,* 313-326.

Abrams, S. (1977). *A polygraph handbook for attorneys.* Lexington, MA: Lexington Books.

Abrams, S. (1989). *The complete polygraph handbook.* Lexington, MA: Lexington Books.

Abrams, S. (1984). The question of the intent question. *Polygraph, 13,* 326-332.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions In Psychological Science, 8*, 3-9.

Ansley, N. (1983). A compendium on polygraph validity. *Polygraph, 12,* 53-61.

Ansley, N. (1989). *Accuracy and utility of RI screening by student examiners at DODPI.* Polygraph and Personnel Security Research. Office of Security. National Security Agency. Fort George G. Meade, MD.

Ansley, N. (1990). The validity and reliability of polygraph decisions in real cases. *Polygraph, 19,* 169-181.

Ansley, N. (1992). The history and accuracy of guilty knowledge and peak of tension tests. *Polygraph, 21*, 174-247.

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

Backster School of Lie Detection (2011). *Basic polygraph examiner's course chart interpretation notebook.* Backster School of Lie Detection: San Diego.

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., Kircher, J. C., & Bernhardt, P. C. (2008). New measures improve the accuracy of the directed-lie test when detecting deception using a mock crime. *Physiology and Behavior, 94*, 331-340.

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

Blackstone, K. (2011). *Polygraph, Sex Offenders, and the Court: What Professionals Should Know About Polygraph..., and a Lot More.* Concord, MA: Emerson Books.

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

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

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

Brownlie, C., Johnson, G. J., & Knill, B. (1997) *Validation study of the relevant/irrelevant screening format.* Unpublished report.

Capps, M. H. (1991). Predictive value of the sacrifice relevant. *Polygraph, 20*(1), 1-8.

√Capps, M. H. & Ansley, N. (1992). Comparison of two scoring scales. *Polygraph, 21*, 39-43.

Capps, M. H., Knill, B. L., & Evans, R. K. (1993). Effectiveness of the symptomatic questions. *Polygraph, 22, 2*85-298.

√Correa, E. J. & Adams, H. E. (1981). The validity of the pre-employment polygraph examination and the effects of motivation. *Polygraph*, *10*, 143-155.

Crewson, P. E. (2001). *A comparative analysis of polygraph with other screening and diagnostic tools.* Research Support Service. Report No. DoDPI01-R-0003. Reprinted in Polygraph 32, (57-85).

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

*Driscoll, L. N., Honts, C. R., & Jones, D. (1987). The validity of the positive control physiological detection of deception technique. *Journal of Police Science and Administration, 15,* 46-50. Reprinted in *Polygraph, 16(3),* 218-225.

√Forman, R. F. & McCauley, C. (1986). Validity of the positive control polygraph test using the field practice model. *Journal of Applied Psychology*, *71*, 691-698. Reprinted in *Polygraph*, *16*(2), 145-160.

√Ganguly, A. K., Lahri, S. K., & Bhaseen, V. (1986). Detection of deception by conventional qualitative method and its confirmation by quantitative method - An experimental study in polygraphy. *Polygraph, 15,* 203-210.

√Ginton, A., Daie, N., Elaad, E., & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology*, *67*, 131-137.

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

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

*Gordon, N. J., Mohamed, F. B., Faro, S. H., Platek, S. M., Ahmad, H., & Williams, J. M. (2005). Integrated zone comparison polygraph technique accuracy with scoring algorithms. *Physiology & behavior*, *87(2)*, 251-254. (Same study is described in Mohamed, F. B., Faro, S. H., Gordon, N. J., Platek, S. M., Ahmad, H. & Williams, J.M. (2006).)

√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,* 83-119.

√Handler, M. & Nelson, R. (In press). Criterion validity of the United States Air Force Modified General Question Technique and three position scoring. *Polygraph*.

*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.

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

Hilliard, D. L. (1979). A cross analysis between relevant questions and a generalized intent to answer truthfully question. *Polygraph, 8,* 73-77.

*Honts, C. R. (1996). Criterion development and validity of the CQT in field application. *The Journal of General Psychology*, *123*, 309-324.

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

*Honts, C. R., Amato, S. & Gordon, A. (2004). Effects of outside issues on the comparison question test. *Journal of General Psychology*, *131(1)*, 53-74.

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

√Honts, C. R. & Hodes, R. L. (1983). The detection of physical countermeasures. *Polygraph*, *12*, 7-17.

*Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Psychology*, *70(1)*, 177-187.

Honts, C. R. & Peterson, C. F. (1997). Brief of the Committee of Concerned Social Scientists as Amicus Curiae United States v Scheffer. Available from the author.

*Honts, C. R. & Raskin, D. (1988). A field study of the validity of the directed lie control question. *Journal of Police Science and Administration*, *16(1)*, 56-61.

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

Honts, C. R., & Reavy, R. (2009). *Effects of Comparison Question Type and Between Test Stimulation on the Validity of Comparison Question Test*. US Army Research Office: Grant Number W911NF-07-1-0670.

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

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

√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. S. (1994). The value and effectiveness of the sacrifice relevant question: An empirical assessment. *Polygraph, 23,* 261-279.

√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.

Jayne, B. (1989). A comparison between the predictive value of two common preemployment screening procedures. *The Investigator,* 5(3).

√Jayne, B. C. (1990). Contributions of physiological recordings in the polygraph technique. *Polygraph, 19,* 105-117.

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

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

Kokish, R., Levenson, J. S., & Blasingame, G. D. (2005). Post-conviction sex offender polygraph examination: client-reported perceptions of utility and accuracy. *Sexual Abuse : A Journal of Research and Treatment, 17,* 211-21.

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

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

√Krapohl, D. J. (2005). Polygraph decision rules for evidentiary and paired testing (Marin protocol) applications. *Polygraph, 34,* 184-192.

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

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

*Krapohl, D. J. & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: A replication. *Polygraph*, *35(1)*, 55-63.

√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. J. & 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. & Ryan, A. H. (2001). A belated look at symptomatic questions. *Polygraph, 30,* 206-212.

√Krapohl, D. J., Senter, S. M., & Stern, B. A. (2005). An exploration of methods for the analysis of multiple-issue Relevant/Irrelevant screening data. *Polygraph, 34*(1), 47-62.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43,* 385-388.

*MacLaren, V. V. (2001). A quantitative review of the guilty knowledge test. *The Journal of Applied Psychology, 86*, 674-683.

*Mangan, D. J., Armitage, T. E., & Adams, G. C. (2008). A field study on the validity of the Quadri-Track Zone Comparison Technique. *Physiology and Behavior,* 17-23.

√Matte, J. A. (1990). *Validation study on the polygraph Quadri-Zone Comparison Technique.* Research Abstract LD 01452, Vol. 1502, 1989, University Microfilm International (UMI), Ann Arbor, MI.

√Matte, J. A. (2010). A field study of the Backster Zone Comparison Technique's Either Or Rule and scoring system versus two other scoring systems when relevant question elicits strong response. *European Polygraph*, *4*, 53-69.

*Matte, J. A. & Reuss, R. M. (1989). A field validation study of the Quadri-Zone Comparison Technique. *Polygraph, 18*, 187-202.

√Meiron, E., Krapohl, D. J., & Ashkenazi, T. (2008). An assessment of the Backster "Either-Or" Rule in polygraph scoring. *Polygraph*, *37*, 240-249.

Mohamed, F. B., Faro, S. H., Gordon, N. J., Platek, S. M., Ahmad, H., & Williams, J. M. (2006). Brain mapping of deception and truth telling about an ecologically valid situation: functional MR imaging and polygraph investigation--initial experience. *Radiology, 238,* 679-88.

National Research Council (2003). *The Polygraph and Lie Detection.* Washington, D.C.: National Academy of Sciences.

*Nelson, R. (In press). Monte Carlo study of criterion validity of Backster You-Phase examinations. *Polygraph*.

*Nelson, R. (In press). Monte Carlo study of criterion validity of the Directed Lie Screening Test using the seven-position, three-position and Empirical Scoring Systems. *Polygraph*.

*Nelson, R. (2011). Monte Carlo study of criterion validity for two-question zone comparison tests with the Empirical Scoring System, seven-position, and three-position scoring models. *Polygraph*, 40, 146-156.

*Nelson, R. & Blalock, B. (In press). Extended analysis of Senter, Waller and Krapohl's AFMGQT examination data with the Empirical Scoring System and the Objective Scoring System, version 3. *Polygraph*, (In press).

*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, 172-179.

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

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

*Nelson, R. & Handler, M. (In press). Monte Carlo study of the United States Air Force Modified General Question Technique with two three and four questions. *Polygraph*.

*Nelson, R., Handler, M., Adams, G., & Backster, C. (In press). Survey of reliability and criterion validity of Backster numerical scores of You-Phase exams from confirmed field investigations. *Polygraph*.

*Nelson, R., Handler, M., Blalock, B., & Cushman, B. (In press). Blind scoring of confirmed federal You-Phase examinations by experienced and inexperienced examiners: Criterion validity with the Empirical Scoring System and the seven-position model. *Polygraph*.

*Nelson, R., Handler, M., Blalock, B., & Hernández, N. (In press). Replication and extension study of Directed Lie Screening Tests: Criterion validity with the seven- and three-position models and the Empirical Scoring System. *Polygraph*.

*Nelson, R., Handler, M., & Morgan, C. (In press). Criterion validity of the Directed Lie Screening Test and the Empirical Scoring System with inexperienced examiners and non-naive examinees in a laboratory setting. *Polygraph*.

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

*Nelson, R., Handler, M., & Senter, S. (In press). Monte Carlo study of criterion validity of the Directed Lie Screening Test using the Empirical Scoring System and the Objective Scoring System version 3. *Polygraph*.

*Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph*, 40(2), 67-78.

*Nelson, R. & Krapohl, D. (2011). Criterion validity of the Empirical Scoring System with experienced examiners: Comparison with the seven-position evidentiary model using the Federal Zone Comparison Technique. *Polygraph*, *40*, 79-85.

*Nelson, R., Krapohl, D., & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph*, *37*, 185-215.

Office of Technology Assessment (1983). The validity of polygraph testing: A research review and evaluation. Washington, D.C.: U.S. Congress, Office of Technology Assessment.

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

√Patrick, C. J. & Iacono, W. G. (1991). Validity of the control question polygraph test: The problem of sampling bias. *Journal of Applied Psychology*, *76*, 229-238.

Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, *15*, 344-359.

Podlesny, J., Raskin, D., & Barland, G. (1976). *Effectiveness of Techniques and Physiological Measures in the Detection of Deception.* Report No. 76-5, Contract 75-N1-99-001 LEAA (available through Department of Psychology, University of Utah, Salt Lake City).

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.

Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J. & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of applied psychology,* 89, 1099-105.

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

Raskin, D. C. & Honts, C. R. (2002). Handbook of polygraph testing. In M. Kleiner (Ed.), *Handbook of Polygraph Testing.* San Diego: Academic Press.

Raskin, D. C. & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin, 86,* 54-59.

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). Baltimore, MD: Williams & Wilkins.

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

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

√Research Division Staff (2001). *Test of a mock theft scenario for use in the Psychophysiological Detection of Deception: IV*. Report No. DoDPI00-R-0002. Department of Defense Polygraph Institute. Reprinted in *Polygraph 30*(4) 244-253.

√Rovner, L. I. (1986). Accuracy of physiological detection of deception for subjects with prior knowledge. *Polygraph*, *15(1)*, 1-39.

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

Senter, S. M. & Dollins, A. B. (2002). *New Decision Rule Development: Exploration of a two-stage approach.* Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in *Polygraph 37*, 149-164.

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

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

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

Senter, S., Weatherman, D., Krapohl, D., & Horvath, F. (2010). Psychological set or differential salience: A proposal for reconciling theory and terminology in polygraph testing. Polygraph, 39 (2), 109-117.

*Shurani, T. (2011). Polygraph verification test. *European Polygraph*, 16.

*Shurani, T. & Chaves, F. (2010). Integrated Zone Comparison Technique and ASIT PolySuite algorithm: A field validity study. *European Polygraph*, *4(2)*, 71-80.

*Shurani, T., Stein, E. & Brand, E. (2009). A Field Study on the Validity of the Quadri-Track Zone Comparison Technique. *European Polygraph*, *1*, 5-24.

√Slowik, S. M. & Buckley, J. 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.

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

Verschuere, B., Meijer, E., & Merckelbach, H. (2008). The Quadri-Track Zone Comparison Technique: It's just not science. A critique to Mangan, Armitage, and Adams (2008). *Physiology and Behavior,* 1-2, 27-28.

√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

Sample Sizes of Included Studies

PDD Technique	Study	Total N	N Deceptive	N Truthful	Total Scores	Deceptive Scores	Truthful Scores	Scorers
AFMGQT (7-position)	Senter, Waller & Krapohl (2008) [1]	69	33	36	69	33	36	1
AFMGQT (7-position)	Nelson, Handler, Morgan & O'Burke (In press) [2]	22	11	11	66	33	33	3
AFMGQT (7-position)	Nelson, Handler, & Senter (In press) [3 A]	-	-	-	100	50	50	1
AFMGQT (ESS)	Nelson, Blalock & Handler (2011) [2]	-	-	-	66	33	33	3
AFMGQT (ESS)	Nelson & Blalock (In press) [1]	-	-	-	69	33	36	1
AFMGQT (ESS)	Nelson, Handler, & Senter (In press) [3 A]	100	50	50	100	50	50	1
Backster You-Phase (Backster)	Nelson, Handler, Adams & Backster (In press) [4]	22	11	11	154	77	77	7
Backster You-Phase (Backster)	Nelson (In press)	100	50	50	100	50	50	1
CIT	MacLaren 2001	1,070	666	404	1,070	666	404	39
DLST/TES (7-position)	Research Division Staff 1995a	94	26	68	94	26	68	3
DLST/TES (7-position)	Research Division Staff 1995b	85	30	55	85	30	55	10
DLST/TES (7-position)	Nelson (In press) B	100	50	50	100	50	50	1
DLST/TES (7-position)	Nelson Handler Blalock & Hernández (In press) [5 C]	49	25	24	98	50	48	2
DLST/TES (ESS)	Nelson & Handler (In press)	100	50	50	100	50	50	1
DLST/TES (ESS)	Nelson, Handler & Morgan (In press)	49	24	25	49	24	25	1
DLST/TES (ESS)	Nelson (In press) [B]	-	-	-	100	50	50	1
DLST/TES (ESS)	Nelson, Handler, Blalock & Hernández (In press) [5 C]	-	-	-	98	50	48	2
Federal You-Phase (7-position)	Nelson (2011) [D]	100	50	50	100	50	50	1
Federal You-Phase (7-position)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-	-	-	220	110	110	10
Federal You-Phase (ESS)	Nelson (2011) [D]	100	50	50	100	50	50	1
Federal You-Phase (ESS)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-	-	-	220	110	110	10
Federal ZCT (7-position)	Blackwell (1998)	100	65	35	300	195	105	3
Federal ZCT (7-position)	Krapohl & Cushman (2006) [7 F]	100	50	50	1,000	500	500	10
Federal ZCT (7-position)	Honts, Amato & Gordon (2004) as reported in Honts in Grahnag (2004)	48	24	24	144	72	72	3
Federal ZCT (7-position evidentiary)	Krapohl & Cushman (2006) [7 F]	-	-	-	1,000	500	500	10
Federal ZCT (7-position evidentiary)	Nelson & Krapohl (2011) [8 G]	60	30	30	60	30	30	6
IZCT (Horizontal)	Shurani & Chavez (2010)	84	44	40	84	44	40	4
IZCT (Horizontal)	Gordon, Mohamed, Faro, Platek, Ahmad & Williams (2005)	11	6	5	11	6	5	1
IZCT (Horizontal)	Shurani (2011)	84	36	48	84	36	48	3
MQTZCT (Matte)	Matte & Reuss (1989) dissertation	122	64	58	122	64	58	2
MQTZCT (Matte)	Shurani, Stein & Brand (2009)	57	28	29	57	28	29	4
MQTZCT (Matte)	Mangan, Armitage & Adams (2008)	140	91	49	140	91	49	1
Utah-RCMP/CPC (Utah)	Honts, Hodes & Raskin (1985)	38	19	19	38	19	19	1
Utah-RCMP/CPC (Utah)	Driscoll, Honts & Jones, 1987)	40	20	20	40	20	20	1
Utah-RCMP/CPC (Utah)	Honts (1996)	32	21	11	32	21	11	1
Utah-DLC (Utah)	Honts & Raskin (1988)	25	12	13	25	12	13	1
Utah-DLC (Utah)	Horowitz, Kircher, Honts & Raskin (1997)	30	15	15	30	15	15	1
Utah-DLC (Utah)	Kircher & Raskin (1988)	100	50	50	200	100	100	2
Utah-DLC (Utah)	Honts, Raskin & Kircher (1987)	20	10	10	20	10	10	1
ZCT (ESS)	Nelson, Krapohl & Handler (2008) [7]	-	-	-	700	350	350	7
ZCT (ESS)	Nelson, Blalock, Oelrich & Cushman (2011) [7]	-	-	-	250	150	100	25
ZCT (ESS)	Nelson & Krapohl (2011) [8 G]	-	-	-	60	30	30	6
ZCT (ESS)	Nelson et al (2011)	572	304	268	1,382	741	641	74
ZCT (ESS)	Blalock, Cushman & Nelson (2009) [7]	-	-	-	900	450	450	9
ZCT (ESS)	Handler, Nelson, Goodson & Hicks (2010) [7]	-	-	-	1,900	950	950	19

[1-8] Sample scores based on the same sample cases.
[A-G] Sample scores published in the same study.

Appendix B

Criterion Accuracy of Included Studies

PDD Technique	Study	Sens.	Spec.	FN	FP	D-INC	T-INC	Unweighted Accuracy	Unweighted INC
AFMGQT (7-position)	Senter, Waller & Krapohl (2008) [1]	.758	.917	.212	.083	.030	.001	.849	.015
AFMGQT (7-position)	Nelson, Handler, Morgan & O'Burke (In press) [2]	.818	.364	.001	.333	.182	.303	.761	.242
AFMGQT (7-position)	Nelson, Handler, & Senter (In press) [3 A]	.780	.420	.040	.200	.140	.420	.814	.280
AFMGQT (ESS)	Nelson, Blalock & Handler (2011) [2]	.831	.616	.010	.175	.158	.208	.883	.183
AFMGQT (ESS)	Nelson & Blalock (In press) [1]	.511	.862	.211	.027	.277	.028	.839	.152
AFMGQT (ESS)	Nelson, Handler, & Senter (In press) [3 A]	.806	.639	.067	.131	.127	.229	.876	.178
Backster You-Phase (Backster)	Nelson, Handler, Adams & Backster (In press) [4]	.943	.543	.009	.274	.048	.183	.828	.116
Backster You-Phase (Backster)	Nelson (In press)	.668	.592	.019	.079	.313	.329	.927	.321
CIT	MacLaren 2001	.815	.832	.185	.168	.001	.001	.823	.000
DLST/TES (7-position)	Research Division Staff 1995a	.654	.676	.154	.206	.192	.118	.788	.155
DLST/TES (7-position)	Research Division Staff 1995b	.833	.909	.167	.073	.000	.018	.880	.009
DLST/TES (7-position)	Nelson (In press) B	.910	.677	.037	.184	.053	.139	.874	.096
DLST/TES (7-position)	Nelson Handler Blalock & Hernández (In press) [5 C]	.583	.940	.271	.020	.145	.039	.831	.092
DLST/TES (ESS)	Nelson & Handler (In press)	.917	.587	.036	.253	.047	.160	.831	.104
DLST/TES (ESS)	Nelson, Handler & Morgan (In press)	.625	.950	.210	.040	.165	.010	.854	.088
DLST/TES (ESS)	Nelson (In press) [B]	.935	.730	.046	.195	.020	.075	.871	.048
DLST/TES (ESS)	Nelson, Handler, Blalock & Hernández (In press) [5 C]	.665	.839	.207	.040	.126	.119	.859	.123
Federal You-Phase (7-position)	Nelson (2011) [D]	.833	.417	.010	.138	.157	.444	.870	.301
Federal You-Phase (7-position)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	.844	.730	.036	.171	.119	.097	.885	.108
Federal You-Phase (ESS)	Nelson (2011) [D]	.813	.729	.050	.126	.090	.102	.897	.096
Federal You-Phase (ESS)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	.859	.770	.027	.143	.145	.325	.906	.235
Federal ZCT (7-position)	Blackwell (1998)	.923	.448	.015	.295	.062	.257	.793	.159
Federal ZCT (7-position)	Krapohl & Cushman (2006) [7 F]	.824	.560	.044	.180	.132	.260	.853	.196
Federal ZCT (7-position)	Honts, Amato & Gordon (2004) as reported in Honts in Grahnag (2004)	.917	.917	.001	.083	.083	.000	.958	.042
Federal ZCT (7-pos. evidentiary)	Krapohl & Cushman (2006) [7 F]	.792	.824	.122	.116	.086	.060	.872	.073
Federal ZCT (7-pos. evidentiary)	Nelson & Krapohl (2011) [8 G]	.933	.667	.000	.233	.067	.133	.870	.100
IZCT (Horizontal)	Shurani & Chavez (2010)	.955	.900	.023	.001	.023	.100	.988	.061
IZCT (Horizontal)	Gordon, Mohamed, Faro, Platek, Ahmad & Williams (2005)	.999	.800	.001	.001	.001	.200	.999	.100
IZCT (Horizontal)	Shurani (2011)	.999	.999	.001	.001	.001	.001	.999	.000
MQTZCT (Matte)	Matte & Reuss (1989) dissertation	.969	.914	.000	.001	.031	.086	.999	.059
MQTZCT (Matte)	Shurani, Stein & Brand (2009)	.929	1.000	.071	.001	.000	.001	.964	.000
MQTZCT (Matte)	Mangan, Armitage & Adams (2008)	.978	1.000	.001	.001	.022	.001	.999	.011
Utah-RCMP/CPC (Utah)	Honts, Hodes & Raskin (1985)	.895	.421	.001	.211	.105	.368	.833	.237
Utah-RCMP/CPC (Utah)	Driscoll, Honts & Jones, 1987)	.900	.900	.001	.001	.100	.100	.999	.100
Utah-RCMP/CPC (Utah)	Honts (1996)	.714	.818	.048	.001	.238	.182	.969	.210
Utah-DLC (Utah)	Honts & Raskin (1988)	.917	.846	.083	.001	.001	.154	.958	.077
Utah-DLC (Utah)	Horowitz, Kircher, Honts & Raskin (1997)	.733	.867	.133	.133	.133	.001	.856	.067
Utah-DLC (Utah)	Kircher & Raskin (1988)	.880	.860	.060	.060	.060	.080	.935	.070
Utah-DLC (Utah)	Honts, Raskin & Kircher (1987)	.800	.700	.001	.200	.200	.100	.889	.150
ZCT (ESS)	Nelson, Krapohl & Hanlder (2008) [7]	.749	.814	.154	.077	.097	.109	.872	.103
ZCT (ESS)	Nelson, Blalock, Oelrich & Cushman (2011) [7]	.793	.930	.073	.001	.133	.070	.958	.102
ZCT (ESS)	Nelson & Krapohl (2011) [8 G]	.833	.633	.001	.133	.167	.233	.913	.200
ZCT (ESS)	Nelson et al (2011)	.863	.789	.047	.093	.103	.107	.921	.105
ZCT (ESS)	Blalock, Cushman & Nelson (2009) [7]	.773	.727	.122	.102	.104	.171	.870	.138
ZCT (ESS)	Handler, Nelson, Goodson & Hicks (2010) [7]	.865	.881	.103	.089	.040	.039	.901	.040

[1-8] Sample scores based on the same sample cases.
[A-G] Sample scores published in the same study.

Appendix C

Reliability Statistics for Included Studies

PDD Technique	Study	Fleiss' Kappa	Decision Agreement	Correlation
AFMGQT (7-position)	Senter, Waller & Krapohl (2008) [1]	.750	.930	.940
AFMGQT (7-position)	Nelson, Handler, Morgan & O'Burke (In press) [2]	-	1.000	-
AFMGQT (7-position)	Nelson, Handler, & Senter (In press) [3 A]	-	-	-
AFMGQT (ESS)	Nelson, Blalock & Handler (2011) [2]	-	1.000	.931
AFMGQT (ESS)	Nelson & Blalock (In press) [1]	-	-	-
AFMGQT (ESS)	Nelson, Handler, & Senter (In press) [3 A]	-	-	-
Backster You-Phase (Backster)	Nelson, Handler, Adams & Backster (In press) [4]	-	-	.567
Backster You-Phase (Backster)	Nelson (In press)	-	-	-
CIT	MacLaren 2001	-	-	-
DLST/TES (7-position)	Research Division Staff 1995a	.760	.890	-
DLST/TES (7-position)	Research Division Staff 1995b	-	-	-
DLST/TES (7-position)	Nelson (In press) B	-	-	-
DLST/TES (7-position)	Nelson Handler Blalock & Hernández (In press) [5 C]	-	.722	-
DLST/TES (ESS)	Nelson & Handler (In press)	-	-	-
DLST/TES (ESS)	Nelson, Handler & Morgan (In press)	-	.911	-
DLST/TES (ESS)	Nelson (In press) [B]	-	-	-
DLST/TES (ESS)	Nelson, Handler, Blalock & Hernández (In press) [5 C]	-	.769	-
Federal You-Phase (7-position)	Nelson (2011) [D]	-	-	-
Federal You-Phase (7-position)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-	.852	-
Federal You-Phase (ESS)	Nelson (2011) [D]	-	-	-
Federal You-Phase (ESS)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-	.897	-
Federal ZCT (7-position)	Blackwell (1998)	.570	.800	-
Federal ZCT (7-position)	Krapohl & Cushman (2006) [7 F]	-	-	-
Federal ZCT (7-position)	Honts, Amato & Gordon (2004) as reported in Honts in Grahnag (2004)	-	-	-
Federal ZCT (7-position evidentiary)	Krapohl & Cushman (2006) [7 F]	-	.870	-
Federal ZCT (7-position evidentiary)	Nelson & Krapohl (2011) [8 G]	-	-	-
IZCT (Horizontal)	Shurani & Chavez (2010)	-	-	-
IZCT (Horizontal)	Gordon, Mohamed, Faro, Platek, Ahmad & Williams (2005)	-	-	-
IZCT (Horizontal)	Shurani (2011)	-	-	-
MQTZCT (Matte)	Matte & Reuss (1989) dissertation	-	-	.990
MQTZCT (Matte)	Shurani, Stein & Brand (2009)	-	-	-
MQTZCT (Matte)	Mangan, Armitage & Adams (2008)	-	-	-
Utah-RCMP/CPC (Utah)	Honts, Hodes & Raskin (1985)	.480	.950	.880
Utah-RCMP/CPC (Utah)	Driscoll, Honts & Jones, 1987)	-	-	.860
Utah-RCMP/CPC (Utah)	Honts (1996)	-	.930	.910
Utah-DLC (Utah)	Honts & Raskin (1988)	-	-	.940
Utah-DLC (Utah)	Horowitz, Kircher, Honts & Raskin (1997)	-	-	.920
Utah-DLC (Utah)	Kircher & Raskin (1988)	.730	.990	.970
Utah-DLC (Utah)	Honts, Raskin & Kircher (1987)	.730	.960	-
ZCT (ESS)	Nelson, Krapohl & Hanlder (2008) [7]	.610	-	-
ZCT (ESS)	Nelson, Blalock, Oelrich & Cushman (2011) [7]	-	.950	-
ZCT (ESS)	Nelson & Krapohl (2011) [8 G]	-	-	-
ZCT (ESS)	Nelson et al (2011)	-	-	-
ZCT (ESS)	Blalock, Cushman & Nelson (2009) [7]	.560	-	-
ZCT (ESS)	Handler, Nelson, Goodson & Hicks (2010) [7]	.590	-	.840

[1-8] Sample scores based on the same sample cases.
[A-G] Sample scores published in the same study.

Appendix D

Means and Standard Deviations of Criterion Deceptive and Criterion Truthful Scores

PDD Technique	Study	Mean D	StDev D	Mean T	StDev T
AFMGQT (7-position)	Senter, Waller & Krapohl (2008) [1]	-	-	-	-
AFMGQT (7-position)	Nelson, Handler, Morgan & O'Burke (In press) [2]	-2.995*	4.727*	2.365*	3.879*
AFMGQT (7-position)	Nelson, Handler, & Senter (In press) [3 A]	-2.827*	4.504*	3.556*	3.766*
AFMGQT (ESS)	Nelson, Blalock & Handler (2011) [2]	-3.850*	4.730*	4.530*	5.180*
AFMGQT (ESS)	Nelson & Blalock (In press) [1]	-2.000*	5.030*	3.420*	3.470*
AFMGQT (ESS)	Nelson, Handler, & Senter (In press) [3 A]	-3.031*	4.535*	3.265*	3.661*
Backster You-Phase (Backster)	Nelson, Handler, Adams & Backster (In press) [4]	-19.649	6.482	3.612	10.010
Backster You-Phase (Backster)	Nelson (In press)	-12.460	8.353	6.820	10.572
CIT	MacLaren 2001	-	-	-	-
DLST/TES (7-position)	Research Division Staff 1995a	-	-	-	-
DLST/TES (7-position)	Research Division Staff 1995b	-	-	-	-
DLST/TES (7-position)	Nelson (In press) B	-2.418*	3.818*	2.653*	3.618*
DLST/TES (7-position)	Nelson Handler Blalock & Hernández (In press) [5 C]	-1.833*	4.099*	3.670*	3.443*
DLST/TES (ESS)	Nelson & Handler (In press)	-2.442*	3.531*	2.086*	3.460*
DLST/TES (ESS)	Nelson, Handler & Morgan (In press)	-1.271*	3.131*	4.660*	2.299*
DLST/TES (ESS)	Nelson (In press) [B]	-3.031*	5.104*	3.265*	3.935*
DLST/TES (ESS)	Nelson, Handler, Blalock & Hernández (In press) [5 C]	-1.781*	3.437*	3.636*	2.917*
Federal You-Phase (7-position)	Nelson (2011) [D]	-6.398	4.914	5.485	5.106
Federal You-Phase (7-position)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-7.991	6.733	6.514	6.680
Federal You-Phase (ESS)	Nelson (2011) [D]	-6.685	6.881	6.735	6.045
Federal You-Phase (ESS)	Nelson, Handler, Blalock & Cushman (In press) [6 E]	-8.606	5.842	6.018	7.107
Federal ZCT (7-position)	Blackwell (1998)	-10.385	9.510	6.981	7.495
Federal ZCT (7-position)	Krapohl & Cushman (2006) [7 F]	-6.264	10.863	9.776	8.212
Federal ZCT (7-position)	Honts, Amato & Gordon (2004) as reported in Honts in Grahnag (2004)	-8.420	6.837	6.640	9.187
Federal ZCT (7-position evidentiary)	Krapohl & Cushman (2006) [7 F]	-6.264	10.863	9.776	8.212
Federal ZCT (7-position evidentiary)	Nelson & Krapohl (2011) [8 G]	-9.600	7.356	6.926	10.709
IZCT (Horizontal)	Shurani & Chavez (2010)	-8.847	15.264	21.181	5.097
IZCT (Horizontal)	Gordon, Mohamed, Faro, Platek, Ahmad & Williams (2005)	-36.000	12.946	8.750	1.635
IZCT (Horizontal)	Shurani (2011)	-19.667	9.607	28.948	5.963
MQTZCT (Matte)	Matte & Reuss (1989) dissertation	-9.148+	2.843+	3.099+	6.002+
MQTZCT (Matte)	Shurani, Stein & Brand (2009)	-6.949+	1.630+	5.388+	1.246+
MQTZCT (Matte)	Mangan, Armitage & Adams (2008)	-10.037+	2.995+	7.190+	3.189+
Utah-RCMP/CPC (Utah)	Honts, Hodes & Raskin (1985)	-11.950	6.520	9.000	10.660
Utah-RCMP/CPC (Utah)	Driscoll, Honts & Jones, 1987)	-10.700	6.000	10.350	6.470
Utah-RCMP/CPC (Utah)	Honts (1996)	-15.000	5.564	8.170	5.270
Utah-DLC (Utah)	Honts & Raskin (1988)	-11.500	5.803	9.000	5.803
Utah-DLC (Utah)	Horowitz, Kircher, Honts & Raskin (1997)	-7.000	13.500	8.500	11.500
Utah-DLC (Utah)	Kircher & Raskin (1988)	-7.710	8.420	10.785	8.671
Utah-DLC (Utah)	Honts, Raskin & Kircher (1987)	-14.000	7.490	9.600	7.490
ZCT (ESS)	Nelson, Krapohl & Hanlder (2008) [7]	-9.606	9.743	9.162	8.564
ZCT (ESS)	Nelson, Blalock, Oelrich & Cushman (2011) [7]	-10.740	8.263	8.690	4.585
ZCT (ESS)	Nelson & Krapohl (2011) [8 G]	-11.833	7.764	6.000	9.592
ZCT (ESS)	Nelson et al (2011)	-11.354	9.392	7.373	9.270
ZCT (ESS)	Blalock, Cushman & Nelson (2009) [7]	-11.253	9.786	7.191	8.785
ZCT (ESS)	Handler, Nelson, Goodson & Hicks (2010) [7]	-7.953	10.017	11.212	8.619

[1-8] Sample scores based on the same sample cases.
[A-G] Sample scores published in the same study.
* Means and standard deviations are reported as subtotal scores for individual questions.
+ Means and standard deviations are reported for subtotal scores for individual test charts.

Appendix E-1

AFMGQT / Seven-position TDA

Study	Senter, Waller & Krapohl (2008)	Nelson, Handler, Morgan & O'Burke (In press)	Nelson & Handler (In press)
Sample N	69	22	100
N Deceptive	33	11	50
N Truthful	36	11	50
Scorers	12	3	1
D Scores	33	33	50
T Scores	36	33	50
Total Scores	69	66	100
Mean D	-2.000	-2.995	-2.827
StDev D	5.030	4.727	4.504
Mean T	3.420	2.365	3.556
StDev T	3.470	3.879	3.766
Reliability Kappa	.750	-	-
Reliability Agreement	.930	.999	-
Reliability Correlation	.940	-	-
Unweighted Average Accuracy	.849	.761	.814
Unweighted Inconclusives	.015	.242	.280
Sensitivity	.758	.818	.780
Specificity	.917	.364	.420
FN Errors	.212	.000	.040
FP Errors	.083	.333	.200
D-INC	.030	.182	.140
T-INC	.000	.303	.420
PPV	.901	.711	.796
NPV	.812	.999	.913
D Correct	.781	.999	.951
T Correct	.917	.522	.677

Appendix E-2

AFMGQT / ESS

Study	Nelson, Blalock Handler (In press)	Nelson & Blalock (In press)	Nelson, Handler & Senter (In press)
Sample N	22	69	100
N Deceptive	11	33	50
N Truthful	11	36	50
Scorers	3	1	1
D Scores	33	33	50
T Scores	33	36	50
Total Scores	66	69	100
Mean D	-3.850	-2.000	-3.031
StDev D	4.730	5.030	4.535
Mean T	4.530	3.420	3.265
StDev T	5.180	3.470	3.661
Reliability Kappa	-	-	-
Reliability Agreement	.999	-	-
Reliability Correlation	.931	-	-
Unweighted Average Accuracy	.883	.839	.876
Unweighted Inconclusives	.183	.152	.178
Sensitivity	.831	.511	.806
Specificity	.616	.862	.639
FN Errors	.010	.211	.067
FP Errors	.175	.027	.131
D-INC	.158	.277	.127
T-INC	.208	.028	.229
PPV	.826	.951	.860
NPV	.984	.803	.905
D Correct	.988	.708	.923
T Correct	.779	.970	.830

Appendix E-3

Backster You-Phase

Study	Nelson, Handler, Adams & Backster (In press)	Nelson (In press)
Sample N	22	100
N Deceptive	11	50
N Truthful	11	50
Scorers	7	1
D Scores	77	50
T Scores	77	50
Total Scores	154	100
Mean D	-19.649	-12.460
StDev D	6.482	8.353
Mean T	3.612	6.820
StDev T	10.010	10.572
Reliability Kappa	-	-
Reliability Agreement	-	-
Reliability Correlation	.567	-
Unweighted Average Accuracy	.825	.927
Unweighted Inconclusives	.117	.321
Sensitivity	.948	.668
Specificity	.532	.592
FN Errors	.001	.019
FP Errors	.286	.079
D-INC	.052	.313
T-INC	.182	.329
PPV	.768	.894
NPV	.999	.969
D Correct	.999	.972
T Correct	.650	.882

Appendix E-4

Concealed Information Test / Guilty Knowledge Test

as reported by MacLaren (2001)

MacLaren, V. V. (2001). A quantitative review of the guilty knowledge test. *Journal of Applied Psychology*, 86, 674-683.

Results are reported with all informed participants, and with only those informed participants who also engaged in the behavioral acts.

Mean (St. Er.) {95% CI}		
	Informed/guilty and uniformed participants	All informed and uninformed participants
Number of studies	39	50
N Deceptive	666	843
N Truthful	404	404
Total N	1070	1243
Unweighted Accuracy	.823 (.011) {.801 to .846}	.795 (.043) {.711 to .880}
Unweighted Inconclusives	-	-
Sensitivity	.815 (.014) {.789 to .842}	.759 (.053) {.655 to .864}
Specificity	.832 (.019) {.795 to .868}	.832 (.068) {.698 to .965}
FN Errors	.185 (.014) {.158 to .211}	.241 (.053) {.136 to .345}
FP Errors	.168 (.019) {.132 to .205}	.168 (.068) {.035 to .302}
D-INC	-	-
T-INC	-	-
PPV	.889 (.011) {.868 to .909}	.904 (.041) {.824 to .984}
NPV	.732 (.021) {.69 to .774}	.623 (.075) {.477 to .770}
D Correct	.815 (.014) {.789 to .842}	.759 (.053) {.655 to .864}
T Correct	.832 (.019) {.795 to .868}	.832 (.068) {.698 to .965}

Appendix E-5

Directed Lie Screening Test (TES) / Seven-position TDA

Study	Research Division Staff (1995a)	Research Division Staff (1995b)	Nelson (In press)	Nelson, Handler, Blalock & Hernández (In press)
Sample N	94	85	100	49
N Deceptive	26	30	50	25
N Truthful	68	55	50	24
Scorers	3	10	1	2
D Scores	26	30	50	50
T Scores	68	55	50	48
Total Scores	94	85	100	98
Mean D	-	-	-2.418	-1.833
StDev D	-	-	3.818	4.099
Mean T	-	-	2.653	3.670
StDev T	-	-	3.618	3.443
Reliability Kappa	.760	-	-	-
Reliability Kappa	.890	-	-	.722
Reliability Agreement	-	-	-	-
Unweighted Average Accuracy	.788	.880	.874	.831
Unweighted Inconclusives	.155	.009	.096	.092
Sensitivity	.654	.833	.910	.583
Specificity	.676	.909	.677	.940
FN Errors	.154	.167	.037	.271
FP Errors	.206	.073	.184	.020
D-INC	.192	.001	.053	.145
T-INC	.118	.018	.139	.039
PPV	.761	.920	.832	.967
NPV	.815	.845	.948	.776
D Correct	.810	.833	.961	.683
T Correct	.767	.926	.786	.979

Appendix E-6

Directed Lie Screening Test (TES) / ESS

Study	Nelson & Handler (In press)	Nelson, Handler & Morgan (In press)	Nelson (In press)	Nelson, Handler, Blalock & Hernández (In press)
Sample N	100	49	100	49
N Deceptive	50	24	50	25
N Truthful	50	25	50	24
Scorers	1	1	1	2
D Scores	50	24	50	50
T Scores	50	25	50	48
Total Scores	100	49	100	98
Mean D	-2.442	-1.271	-3.031	-1.781
StDev D	3.531	3.131	5.104	3.437
Mean T	2.086	4.660	3.265	3.636
StDev T	3.460	2.299	3.935	2.917
Reliability Kappa	-	-	-	-
Reliability Kappa	-	.911	-	.769
Reliability Agreement	-	-	-	-
Unweighted Average Accuracy	.831	.854	.871	.859
Unweighted Inconclusives	.104	.088	.048	.123
Sensitivity	.917	.625	.935	.665
Specificity	.587	.950	.730	.839
FN Errors	.036	.210	.046	.207
FP Errors	.253	.040	.195	.040
D-INC	.047	.165	.020	.126
T-INC	.160	.010	.075	.119
PPV	.784	.940	.827	.943
NPV	.942	.819	.941	.802
D Correct	.962	.749	.953	.763
T Correct	.699	.960	.789	.954

Appendix E-7

Federal You-Phase / Seven-position TDA

Study	Nelson (In press)	Nelson, Handler, Blalock & Cushman (In press)
Sample N	100	22
N Deceptive	50	11
N Truthful	50	11
Scorers	1	10
D Scores	50	110
T Scores	50	110
Total Scores	100	220
Mean D	-6.398	-7.991
StDev D	4.914	6.733
Mean T	5.485	6.514
StDev T	5.106	6.680
Reliability Kappa	-	-
Reliability Kappa	-	.852
Reliability Agreement	-	-
Unweighted Average Accuracy	.870	.885
Unweighted Inconclusives	.301	.108
Sensitivity	.833	.844
Specificity	.417	.730
FN Errors	.010	.036
FP Errors	.138	.171
D-INC	.157	.119
T-INC	.444	.097
PPV	.858	.832
NPV	.977	.953
D Correct	.988	.959
T Correct	.751	.810

Appendix E-8

Federal You-Phase / ESS

Study	Nelson (In press)	Nelson, Handler, Blalock & Cushman (In press)
Sample N	100	22
N Deceptive	50	11
N Truthful	50	11
Scorers	1	10
D Scores	50	110
T Scores	50	110
Total Scores	100	220
Mean D	-6.685	-8.606
StDev D	6.881	5.842
Mean T	6.735	6.018
StDev T	6.045	7.107
Reliability Kappa	-	-
Reliability Kappa	-	0.9
Reliability Agreement	-	-
Unweighted Average Accuracy	.897	.906
Unweighted Inconclusives	.096	.235
Sensitivity	.813	.859
Specificity	.729	.770
FN Errors	.050	.027
FP Errors	.126	.143
D-INC	.090	.145
T-INC	.102	.325
PPV	.866	.857
NPV	.936	.966
D Correct	.942	.970
T Correct	.853	.843

Appendix E-9

Federal ZCT / Seven-position TDA

Study	Blackwell (1998)	Krapohl & Cushman (2006)	Honts, Amato & Gordon (2004) as reported in Grahnag (2004)
Sample N	100	100	48
N Deceptive	65	50	24
N Truthful	35	50	24
Scorers	3	10	3
D Scores	195	500	72
T Scores	105	500	72
Total Scores	300	1,000	144
Mean D	-10.385	-6.264	-8.420
StDev D	9.510	10.863	6.837
Mean T	6.981	9.776	6.640
StDev T	7.495	8.212	9.187
Reliability Kappa	.570	-	-
Reliability Kappa	.800	-	-
Reliability Agreement	-	-	-
Unweighted Average Accuracy	.793	.852	.958
Unweighted Inconclusives	.159	.198	.042
Sensitivity	.923	.824	.917
Specificity	.448	.556	.917
FN Errors	.015	.044	.000
FP Errors	.295	.180	.083
D-INC	.062	.132	.083
T-INC	.257	.264	.001
PPV	.758	.821	.917
NPV	.967	.927	.999
D Correct	.984	.949	.999
T Correct	.603	.755	.917

Appendix E-10

Federal ZCT / Seven-position TDA with Evidentiary Rules

Study	Krapohl & Cushman (2006)	Nelson & Krapohl (2011)
Sample N	100	60
N Deceptive	50	30
N Truthful	50	30
Scorers	10	6
D Scores	500	30
T Scores	500	30
Total Scores	1,000	60
Mean D	-6.264	-9.600
StDev D	10.863	7.356
Mean T	9.776	6.926
StDev T	8.212	10.709
Reliability Kappa	-	-
Reliability Agreement	0.870	-
Reliability Correlation	-	-
Unweighted Average Accuracy	.872	.870
Unweighted Average Accuracy	.073	.100
Sensitivity	.792	.933
Specificity	.824	.667
FN Errors	.122	.001
FP Errors	.116	.233
D-INC	.086	.067
T-INC	.060	.133
PPV	.872	.800
NPV	.871	.999
D Correct	.999	.999
T Correct	0.88	.741

Appendix E-11

Integrated Zone Comparison Technique / Horizontal Scoring System

Study	Gordon, Mohamed, Faro, Platek, Ahmad & Williams (2005)	Shurani & Chaves (2010)	Shurani (2011)
Sample N	11	84	84
N Deceptive	6	44	36
N Truthful	5	40	48
Scorers	1	4	1
D Scores	6	44	36
T Scores	5	40	48
Total Scores	11	84	84
Mean D	-36.000	-8.847	-19.667
StDev D	12.946	15.264	9.607
Mean T	8.750	21.181	28.948
StDev T	1.635	5.097	5.963
Reliability Kappa	-	-	-
Reliability Agreement	-	-	-
Reliability Correlation	-	-	-
Unweighted Average Accuracy	.999	.988	.999
Unweighted Inconclusives	.100	.061	.001
Sensitivity	.999	.955	.999
Specificity	.800	.900	.999
FN Errors	.001	.023	.001
FP Errors	.001	.001	.001
D-INC	.001	.023	.001
T-INC	.200	.100	.001
PPV	.999	.999	.999
NPV	.999	.975	.999
D Correct	.999	.977	.999
T Correct	.999	.999	.999

Appendix E-12

Matte Quadri-Track Zone Comparison Technique

Study	Matte & Reuss (1989)	Shurani, Stein & Brand (2009)	Mangan, Armitage & Adams (2008)
Sample N	122	57	140
N Deceptive	64	28	91
N Truthful	58	29	49
Scorers	2	4	1
D Scores	64	28	91
T Scores	58	29	49
Total Scores	122	57	140
Mean D	-9.148	-6.949*	-10.037*
StDev D	2.843	1.630*	2.995*
Mean T	3.099	5.388*	7.190*
StDev T	6.002	1.246*	3.189*
Reliability Kappa	-	-	-
Reliability Agreement	-	-	-
Reliability Correlation	.990	-	-
Unweighted Average Accuracy	.999	.964	.999
Unweighted Inconclusives	.059	.001	.011
Sensitivity	.969	.929	.978
Specificity	.914	.999	.999
FN Errors	.001	.071	.001
FP Errors	.001	.001	.001
D-INC	.031	.001	.022
T-INC	.086	.001	.001
PPV	.999	.999	.999
NPV	.999	.933	.999
D Correct	.999	.929	.999
T Correct	.999	.999	.999

Appendix E-13

Utah PLC / Utah Numerical Scoring

Study	Kircher & Raskin (1988)	Honts, Raskin & Kircher (1987)
Sample N	100	20
N Deceptive	50	10
N Truthful	50	10
Scorers	1	1
D Scores	50	10
T Scores	50	10
Total Scores	100	20
Mean D	-7.710	-14.000
StDev D	8.420	7.490
Mean T	10.785	9.600
StDev T	8.671	7.490
Reliability Kappa	.730	.730
Reliability Agreement	.990	.960
Reliability Correlation	.970	-
Unweighted Average Accuracy	.935	.889
Unweighted Inconclusives	.070	.150
Sensitivity	.880	.800
Specificity	.860	.700
FN Errors	.060	.000
FP Errors	.060	.200
D-INC	.060	.200
T-INC	.080	.100
PPV	.936	.800
NPV	.935	.999
D Correct	.936	.999
T Correct	.935	.778

Appendix E-14

Utah DLC / Utah Numerical Scoring

Study	Honts & Raskin (1988)	Horowitz, Kircher, Honts & Raskin (1997)
Sample N	25	30
N Deceptive	12	15
N Truthful	13	15
Scorers	1	1
D Scores	12	15
T Scores	13	15
Total Scores	25	30
Mean D	-11.500	-7.000
StDev D	5.803	13.500
Mean T	9.000	8.500
StDev T	5.803	11.500
Reliability Kappa	-	-
Reliability Agreement	-	-
Reliability Correlation	.940	.920
Unweighted Average Accuracy	.958	.856
Unweighted Inconclusives	.077	.067
Sensitivity	.917	.733
Specificity	.846	.867
FN Errors	.083	.133
FP Errors	.001	.133
D-INC	.001	.133
T-INC	.154	.001
PPV	.999	.846
NPV	.910	.867
D Correct	.917	.846
T Correct	.999	.867

Appendix E-15

Utah RCMP Zone / Utah Numerical Scoring

Study	Honts Hodes & Raskin (1985)	Honts (1996)	Driscoll, Honts & Jones, (1987)
Sample N	38	32	40
N Deceptive	19	21	20
N Truthful	19	11	20
Scorers	1	1	1
D Scores	19	21	20
T Scores	19	11	20
Total Scores	38	32	40
Mean D	-11.950	-15.000	-10.700
StDev D	6.520	5.564	6.000
Mean T	9.000	8.170	10.350
StDev T	10.660	5.270	6.470
Reliability Kappa	.480	-	-
Reliability Agreement	.950	.930	-
Reliability Correlation	.880	.910	.860
Unweighted Average Accuracy	.833	.969	.999
Unweighted Inconclusives	.237	.210	.100
Sensitivity	.895	.714	.900
Specificity	.421	.818	.900
FN Errors	.001	.048	.001
FP Errors	.211	.001	.001
D-INC	.105	.238	.100
T-INC	.368	.182	.100
PPV	.810	.999	.999
NPV	.999	.945	.999
D Correct	.999	.938	.999
T Correct	.667	.999	.999

Appendix E-16

Utah PLT DLC Combined / Utah Numerical Scoring

Technique	Utah PLC	Utah DLC	RCMP
Sample N	120	55	110
N Deceptive	60	27	60
N Truthful	60	28	50
Scorers	2	2	3
D Scores	60	27	60
T Scores	60	28	50
Total Scores	120	55	110
Mean D	-10.855	-9.250	-12.550
StDev D	7.955	9.652	6.028
Mean T	10.193	8.750	9.173
StDev T	8.081	8.652	7.467
Reliability Kappa	.730	-	.480
Reliability Agreement	.975	-	.940
Reliability Correlation	.970	.930	.883
Unweighted Average Accuracy	.927	.902	.939
Unweighted Inconclusives	.083	.073	.185
Sensitivity	.867	.815	.833
Specificity	.833	.857	.700
FN Errors	.050	.111	.017
FP Errors	.083	.071	.080
D-INC	.083	.074	.150
T-INC	.083	.071	.220
PPV	.912	.919	.912
NPV	.943	.885	.977
D Correct	.945	.880	.980
T Correct	.909	.923	.897

Appendix E-17

Zone Comparison Techniques / ESS

Study	Nelson, Krapohl & Handler (2008)	Nelson, Blalock, Cushman & Oelrich (2011)	Nelson & Krapohl (2011)	Nelson *et al.* (2011)	Blalock, Cushman & Nelson (2009)	Handler, Nelson, Goodson & Hicks (2010)
Sample N	100	10	60	562	100	100
N Deceptive	50	6	30	298	50	50
N Truthful	50	4	30	264	50	50
Scorers	7	25	6	74	9	19
D Scores	350	150	30	741	450	950
T Scores	350	100	30	641	450	950
Total Scores	700	250	60	1,382	900	1,900
Mean D	-9.634	-10.740	-11.833	-11.354	-11.253	-7.953
StDev D	8.475	8.263	7.764	9.392	9.786	10.017
Mean T	8.849	8.690	6.000	7.373	7.191	11.212
StDev T	7.457	4.585	9.592	9.270	8.785	8.619
Reliability Kappa	.610	-	-	-	.560	.590
Reliability Agreement	-	.950	-	-	-	-
Reliability Correlation	-	-	-	-	-	.840
Unweighted Average Accuracy	.872	.958	.913	.883	.870	.867
Unweighted Inconclusives	.103	.102	.200	.114	.138	.115
Sensitivity	.749	.793	.833	.804	.773	.666
Specificity	.814	.930	.633	.761	.727	.873
FN Errors	.154	.073	.001	.090	.122	.196
FP Errors	.077	.001	.133	.117	.102	.035
D-INC	.097	.133	.167	.105	.104	.138
T-INC	.109	.070	.233	.122	.171	.092
PPV	.907	.999	.862	.873	.883	.950
NPV	.841	.927	.999	.894	.856	.817
D Correct	.829	.916	.999	.899	.864	.773
T Correct	.914	.999	.826	.867	.877	.961

Appendix F

PDD Techniques for which no published studies could be located or for which no published studies could be included in the meta-analysis.

- Backster Exploratory
- Backster SKY
- Backster ZCT
- IZCT Screening
- Law Enforcement Pre-Employment Test (LEPET)*
- Marcy Technique
- Matte Quinque Track
- Matte SGK
- RCMP B Series
- Searching Peak of Tension
- Utah MGQT*

* Although there is no published research to describe these techniques, the LEPET and Utah MGQT formats are structurally nearly identical to the AFMGQT. Validation data for the AFMGQT can be generalized to these techniques if they are evaluated with the TDA models described in the published studies.

Appendix G

Studies that could not be included in the meta-analysis

These studies were excluded for a variety of reasons, including the use of instrumentation that does not match that used in the field, the use of test question sequences or test data analysis models that do not match field testing protocols, non-representative study samples, or insufficient statistical data to calculate the criterion accuracy profile or sampling distributions.

Ansley (1989) produced a study on the effectiveness of Relevant-Irrelevant screening exams with student examiners. Crewson (2001) included the results of this study in a survey of diagnostic and screening tests in the polygraph, medicine and psychology professions. However, the study report is not available, and the study could not be included in the meta-analysis.

Barland, Honts and Barger (1989) studied the accuracy of multi-issue screening exams. No reliability data and no sampling means or standard deviations were available. Data are not available for further analysis. Numerical scoring was completed according to older training and field practice protocols from the U.S. Department of Defense and may not reflect current practices. Without the ability to compare these results to the results from other studies, this series of studies could not be included in the meta-analysis.

Bell, Kircher and Bernhardt (2008) compared the Utah PLC and DLC methods, and concluded no significant differences exist. TDA was limited to automated methods, and therefore was not included in the meta-analysis.

Brownlie, Johnson, and Knill (1997) completed a study on the Relevant-Irrelevant technique. The study is described in the report from the NRC (2003) and Crewson (2001). Crewson (2001) included the results of this study in a survey of diagnostic and screening tests in the polygraph, medicine and psychology professions. However, the study report is not available, and the study could not be included in the meta-analysis.

Correa and Adams (1981) reported the results of a study of the RI technique. The testing procedure does not match field practice, and includes the use of a thermistor respiration sensor instead of standard thoracic and abdominal pneumograph sensors. In addition, an EKG was used in place of the cardiograph sensors. Results, as reported, were 100% accuracy. No data could be obtained for review.

Forman and McCauley (1986) reported the results of a study on the positive-control technique. Reported results did not provide mean or standard deviation statistics for the distributions of scores. Additionally, the study employed a quasi-numerical scoring system that does not reflect PCT field practice as described in other studies.

Ganguly, Lahri and Bhaseen (1986) reported the results of a study based on the Reid Technique. However, testing procedures do not reflect field practices in that only the pneumograph and cardiograph data were evaluated during the study.

Ginton, Daie, Elaad and Ben-Shakhar (1982) reported the results of an interesting field study involving a unique modification of the MGQT technique, using an overall truth question at position 3. This study was not included because this question sequence does not reflect field practices.

Gordon, Fleisher, Morsie, Habib, and Salah (2000) reported the results of a field study of the IZCT, including 309 reported confirmed cases and 1 error. No reliability data was included in the publication, nor any statistical description of the sampling distributions of

deceptive and truthful scores. The authors advised that the data belong to the intelligence service of a foreign government, and the primary author informed the ad hoc committee (personal communication June 10, 2011) that he completed the study report without ever seeing the data.

Honts and Amato (1999) reported the results of an automated presentation of the Relevant-Irrelevant polygraph technique. This procedure does not reflect field practices, and the information could therefore not be included in the meta-analysis.

Honts and Hodes (1983) reported the same results as experiment 1 in the Honts, Hodes and Raskin (1985) study, which did not include a standard cardiovascular arm cuff.

Honts, Hodes and Raskin (1985) experiment 1 involved the Backster You-Phase technique, but did not include a standard cardio cuff and therefore could not be included in the meta-analysis.

Honts and Reavy (2009) used the Federal ZCT in a large-scale laboratory experiment. However, this study was designed to evaluate and compare the effectiveness of PLCs and DLCs. Data as reported could not be used to calculate a dimensional profile of generalizable criterion accuracy estimates.

Horowitz, Kircher, Honts and Raskin (1997) included the results of a study of the RI techniques. RI examination results could not be included because the scoring protocol did not reflect field practices (global analysis and subjective evaluation of consistent and significant responses), and involved the use of seven-position numerical scoring procedures in which the RQ responses were compared with the responses to neutral questions.

Horvath (1988) reported the results of a study on the Reid technique, involving two blind evaluators who used a 7-position scoring model described as similar to the one used by Barland and Raskin (1975) with +/- 5 decision thresholds. This is not the Reid scoring method, which has been described as a three position TDA model that does not employ fixed cut scores. They also included a fifth RQ which is not consistent with current field practices.

Horvath and Palmatier (2008) reported the results of an MGQT format structured like the Reid technique. The study involved two blind evaluators who used a 7-position scoring model described as similar to the one used by Barland and Raskin (1975) with +/- 6 decision thresholds. This is not the Reid scoring method, which has been described as a three position TDA model that did not have cut scores. They also included a fifth RQ which is not consistent with field practices. Additionally, there was only one evaluator so reliability statistics could not be calculated.

Horvath and Reid (1971) reported the results of a study on the Reid technique. The study could not be included in the meta-analysis for several reasons, including a highly selective sample and the lack of a clearly structured decision model. Of the original 75 polygraph exams, 40 were chosen by the author for inclusion in the study sample, and 35 exams were removed from the sample because the author felt they were too easy to score. The act of selecting out exams from individual case files which they felt were not appropriate for the study brings into question whether the resulting sample would be representative of real world cases. The examiners were precluded from making an inconclusive call, and were required to make a DI or NDI call for every case. Some cases had five RQs, which is inconsistent with field testing techniques currently used. No mean and standard deviation scores were provided and no data could be provided to calculate them. Additionally, no interrater reliability statistics were reported.

Hunter and Ash (1973) reported the results of a study of the Reid technique. The Reid technique does not employ a structured decision model or fixed cutscores, but relies on impressionistic decisions from the examiner. The report does not include any inter-rater reliability information and does not include any sampling distribution data that can be used to compare the sampling distributions. Additionally, the study sample was constructed using a highly selective process involving verified cases conducted by the primary author. Without data on reliability and without sampling distribution parameters, or access to the examination data, the sample is of is of unverifiable representativeness and generalizability.

Jayne (1989) reported the results of a field study on the predictive value of polygraph screening tests. The study design was intended to compare screening polygraph accuracy with that of other preemployment screening methods. In addition to highly selective and non-random case selection requirements, the criterion status of the sample screening cases was determined as a function of the results of a subsequent diagnostic polygraph regarding an employee theft investigation that was independent of the screening polygraph. Although an innovative attempt to study future outcomes related to polygraph screening, the results of this study are not suitable for use as a criterion accuracy study.

Jayne (1990) reported the result of a study of the Reid technique. Although the results were previously described by Krapohl (2006), this study could not be included in the present meta-analysis because two different scoring approaches were used, neither of which reflect field practices. One scoring method involved making precise linear measurements of the test data. The other, more common, numerical scoring method was completed while excluding the fourth RQ, which was described as a secondary RQ that could also be scored as a CQ. A test for which the nature and purpose of the stimulus is decided post-hoc lacks scientific rigor. Current field practices do not endorse the exclusion of individual RQs or the use of a secondary RQ as a CQ. In addition, no decision rules or numerical cutscores were used in this study, and decisions were made according to the subjective opinion of the scorer.

Krapohl (2005) reported the results of seven-position evidentiary scoring of Federal ZCT exams. Study data included cases from four different archival samples for which the sampling distributions could not be effectively compared to the sampling distributions from other studies.

Krapohl (2010) showed that seven-position scores of You-Phase exams could be transformed to ESS scores. Sample data was a highly selective and non-representative sample of Backster You-Phase exams, reported by Meiron, Krapohl & Ashknazi (2008).

Matte (1990) is an abstract of a dissertation study completed at an institution that was approved by the State of California but not accredited by an institution recognized by the U.S. Government or foreign equivalent. UMI later became part of ProQuest who subsequently adopted a policy that only those dissertations from regionally accredited universities would be listed and available to the public. The dissertation is maintained by ProQuest with Matte as the author and ProQuest as the publisher. Data from this dissertation study was included in this meta-analysis because the data was previously published in the journal *Polygraph* (Matte & Ruess, 1989).

Matte (2010) reported the results of a study of the Backster Either-Or rule. However, no truthful cases were included in the sample data. This study cannot be construed as a criterion accuracy study, and could not be included in the meta-analysis.

Meiron, Krapohl and Ashkenazi (2008) reported the results of a study of the Backster Either-Or rule. Data, as reported at the 2008 APA annual conference in Indianapolis, included a highly selective and non-random sample from which the results of problematic

examinations were not included. The result of this form of highly selective sampling is that the sample data are systematically devoid of error variance.

Patrick and Iacono (1991) reported the results of a study of the Federal ZCT while reviewing the CQs between charts. Instrumentation recorded both skin conductance and heart rate, which were scored against pre-stimulus levels in a manner that does not reflect field practices.

Patrick and Iacono (1989) in a replication of an earlier study by Raskin and Hare (1978) reported the results of a study involving a technique that resembles the Federal ZCT, using the Utah scoring system and a grand total decision rule. This study is interesting but does not resemble field practices closely enough to be included in the present meta-analysis of criterion accuracy.

Podlesny and Raskin (1978) recorded eight different physiological channels using a test question sequence that resembles the Federal ZCT with a guilt-complex question instead of the symptomatic question at position 8. Numerical scoring was based on the method described by Barland and Raskin (1975) and Raskin and Hare (1978) which is an early version of the Utah numerical scoring system. Guilt complex questions and differences between exclusive and non-exclusive CQs were studied. In addition, the CQT was compared to the GKT. The absence of standard deviations prevents the comparison of the sampling distribution of numerical scores with those from other studies. This study was designed to evaluate a number of important questions, but could not fit into a clear category of test question sequence and TDA model for which comparable replication studies could be located. As a result, this study could not be included in the meta-analysis of criterion accuracy of PDD techniques presently used in field settings.

Raskin and Hare (1978) reported the results of an important study using examinees who were considered to be criminal psychopaths. Results from this study could not be included in the meta-analysis criterion accuracy of field PDD methods because of the use of non-standard testing instrumentation that did not include a standard blood-pressure cardio-activity cuff sensor.

Research Division Staff (2001) described a study used on the development of a laboratory scenario used to manipulate research subjects. This study was not designed to be a criterion accuracy study.

Rovner (1986) in an interesting study of polygraph countermeasures did not report decisions, errors and inconclusive results separately for truthful and deceptive cases. Data are unavailable for further analysis. As a result, the complete dimensional profile of criterion accuracy could not be calculated and the study could not be included as a criterion accuracy study.

Senter and Dollins (2002) published an interesting and informative study of decision rules that is not suitable for use as a criterion study.

Senter (2003) published an interesting and informative study of decision rules that is not suitable for use as a criterion study.

Senter and Dollins (2004) published an interesting and informative study of decision rules that is not suitable for use as a criterion study.

Senter and Dollins (2008) published an interesting and informative study of decision rules that is not suitable for use as a criterion study.

Slowik and Buckley (1975) reported the results of a study of the Reid technique. The study sample was selected from verified cases but does not describe the verification process. The Reid technique does not employ a structured decision model or fixed cutscores, but relies on impressionistic decisions from the examiner. The report does not include any inter-rater reliability information and does not include any sampling distribution parameters that can be used to compare the sampling distributions. Without sampling distribution statistics, or access to the examination data, the sample is of unverifiable representativeness and unknown generalizability.

Van Herk (1990) reported the results of a pilot study and what may be the first publication on three-position TDA. One-third of the cases were unconfirmed, and consequently, this study could not be included in the meta-analysis.

Wicklander and Hunter (1975) reported the results of a study of the Reid technique. The study sample was selected from verified cases but does not describe the verification process. The Reid technique does not employ a structured decision model or fixed cutscores, but relies on impressionistic decisions from the examiner. The report does not include any inter-rater reliability information and does not include any sampling distribution data that can be used to compare the sampling distributions. Without sampling distribution statistics, or access to the examination data, the sample is of unverifiable representativeness and unknown generalizability.

Appendix H

Techniques for which there exists only one study that met the qualitative and quantitative criteria for inclusion in the meta-analysis.

The APA Standards of Practice require a minimum of two published studies of a given technique.

Arther Technique

Horvath, F.S. (1977) reported the results of a field study using the Arther modification of the Reid technique.

Reid Technique

Horvath (1988) reported the results of a laboratory study using the Reid technique.

These studies were previously combined by Krapohl (2006). However, subsequent review suggests that the Reid and Arther techniques are sufficiently dissimilar that the results of these studies cannot be regarded as replicating each other.

The Reid Technique differs from other CQT formats in some important ways. It sometimes employs a fifth RQ while all other PDD techniques are limited to four RQs. Also, the Reid technique does not employ fixed numerical cutscores or a structured decision model. Instead, decisions are made impressionistically by the examiner, using information from the test data, pretest interview, behavioral observations, and case file information. Inclusion of clinical impressions in the decision process severely restricts the ability to validate it in the way that has been done with the other techniques. Although the Reid technique can be credited as the source of many important innovations, and is indeed the wellspring from which all other CQTs have evolved, it is currently not being taught at any polygraph schools accredited by the APA. The number of practitioners using the technique has decreased substantially since the closure of the Reid Polygraph School. The majority of the other techniques in the meta-analysis are currently taught and/or practiced on a considerable scale. A review of the research in support of the technique resulted in an inability to satisfy the requirements of the study selection criteria, including interrater reliability statistics and normative parameters with which to calculate the generalizability of sample data. Results described in the published literature, and data available to the committee, do not permit the statistical treatments applied to all of the other methods. Despite these limitations, the average accuracy level of studies on the Reid technique was not significantly different from the results of this meta-analysis.

Appendix I-1

AFMGQT / Three-position TDA

Two usable studies on the AFMGQT with three-position TDA produced a combined unweighted average accuracy level of .816 (.059), with an inconclusive rate of .443 (.044). The weighted average of correct decisions for the AFMGQT three-position model, was .989 (.016) for criterion deceptive cases and .643 (.116) for criterion truthful cases. The weighted average inconclusive rates were .237 (.060) for criterion deceptive cases and .648 (.067) for criterion truthful cases.

Study	Nelson & Handler (In press)	Handler & Nelson (In press)
Sample N	100	22
N Deceptive	50	11
N Truthful	50	11
Scorers		3
D Scores	50	33
T Scores	50	33
Total Scores	100	66
Mean D	-1.886	-1.903
StDev D	3.161	2.986
Mean T	2.427	1.424
StDev T	2.557	2.722
Reliability Kappa	-	-
Reliability Agreement	-	.945
Reliability Correlation	-	-
Unweighted Accuracy	.869	.740
Unweighted Inconclusives	.457	.421
Sensitivity	.737	.780
Specificity	.260	.180
FN Errors	.007	.010
FP Errors	.088	.186
D-INC	.256	.209
T-INC	.658	.633
PPV	.894	.807
NPV	.974	.947
D Correct	.991	.987
T Correct	.748	.492

Appendix I-2

Army MGQT / Seven-position TDA

Two usable studies on the Army MGQT resulted in an unweighted accuracy level of .694 (.043), with an inconclusive rate of .133 (.038). The weighted average of correct decisions for the Army MGQT, using the seven-position TDA model, was .999 (.050) for criterion deceptive cases and .039 (.085) for criterion truthful cases. The weighted average inconclusive rates were .043 (.034) for criterion deceptive cases and .224 (.065) for criterion truthful cases.

Study	Krapohl & Norris (2000)	Blackwell (1999)
Sample N	32	100
N Deceptive	16	80
N Truthful	16	20
Scorers	3	3
D Scores	16	240
T Scores	16	60
Total Scores	32	300
Mean D	-	-
StDev D	-	-
Mean T	-	-
StDev T	-	-
Reliability Kappa	-	-
Reliability Agreement	-	-
Reliability Correlation	.750	.907
Unweighted Accuracy	.833	.660
Unweighted Inconclusives	.219	.125
Sensitivity	.813	.967
Specificity	.500	.250
FN Errors	.001	.001
FP Errors	.250	.533
D-INC	.188	.033
T-INC	.250	.217
PPV	.765	.644
NPV	.999	.999
D Correct	.999	.999
T Correct	.667	.319

Appendix I-3

Directed-Lie Screening Test / Three-position TDA

Two usable studies on the DLST/TES with three-position TDA produced a combined unweighted average accuracy level of .869 (.037) with an inconclusive rate of .228 (.043). The weighted average of correct decisions for the DLST three-position model, was .827 (.060) for criterion deceptive cases and .912 (.043) for criterion truthful cases. The weighted average inconclusive rates were .427 (.063) for criterion deceptive cases and .210 (.060) for criterion truthful cases.

Study	Nelson (In press)	Nelson, Handler, Blalock & Hernández (In press)
Sample N	100	49
N Deceptive	50	25
N Truthful	50	24
Scorers	1	2
D Scores	50	50
T Scores	50	48
Total Scores	100	98
Mean D	-1.585	-1.458
StDev D	2.382	2.784
Mean T	1.719	2.470
StDev T	2.253	1.853
Reliability Kappa	-	-
Reliability Agreement	-	0.762
Reliability Correlation	-	-
Unweighted Accuracy	.893	.817
Unweighted Inconclusives	.208	.248
Sensitivity	.829	.415
Specificity	.595	.848
FN Errors	.033	.228
FP Errors	.127	.009
D-INC	.138	.355
T-INC	.277	.141
PPV	.867	.979
NPV	.947	.788
D Correct	.962	.645
T Correct	.824	.989

Appendix I-4

Federal You-Phase / Three-position TDA

Two usable studies on the Federal You-Phase technique with three-position TDA produced combined unweighted average accurate level of .881 (.041), with an inconclusive rate of .282 (.046). The weighted average of correct decisions for the Federal You-Phase three-position model, was .977 (.023) for criterion deceptive cases and .786 (.078) for criterion truthful cases. The weighted average inconclusive rates were .181 (.055) for criterion deceptive cases and .348 (.072) for criterion truthful cases.

Study	Nelson (In press)	Nelson, Handler, Blalock & Cushman (In press)
Sample N	100	22
N Deceptive	50	11
N Truthful	50	11
Scorers	1	10
D Scores	50	110
T Scores	50	110
Total Scores	100	220
Mean D	-4.720	-5.394
StDev D	3.345	4.219
Mean T	3.901	3.982
StDev T	3.804	4.432
Reliability Kappa	-	-
Reliability Agreement	-	0.877
Reliability Correlation	-	-
Unweighted Accuracy	.889	.878
Unweighted Inconclusives	.387	.235
Sensitivity	.740	.826
Specificity	.380	.530
FN Errors	.001	.027
FP Errors	.107	.143
D-INC	.260	.145
T-INC	.513	.325
PPV	.874	.852
NPV	.997	.952
D Correct	.999	.968
T Correct	.780	.788

Appendix I-5

Federal ZCT / Three-position TDA (+/-6)

Two usable studies on the Federal ZCT with three-position TDA, using tradition cutscores (+/-6 and -3) produced a combined unweighted average accuracy level of .883 (.048), with an inconclusive rate of .318 (.043). The weighted average of correct decisions for the Federal ZCT three-position model with traditional cutscores, was .990 (.016) for criterion deceptive cases and .675 (.095) for criterion truthful cases. The weighted average inconclusive rates were .158 (.050) for criterion deceptive cases and .477 (.069) for criterion truthful cases.

Study	Capps & Ansley (1992)	Blackwell (1998)
Sample N	100	100
N Deceptive	52	65
N Truthful	48	35
Scorers	1	3
D Scores	52	195
T Scores	48	105
Total Scores	100	300
Mean D	-9.640	-5.938
StDev D	5.146	5.503
Mean T	4.780	4.867
StDev T	5.461	5.580
Reliability Kappa	-	.360
Reliability Agreement	-	.660
Reliability Correlation	-	-
Unweighted Average Accuracy	.977	.779
Unweighted Inconclusives	.386	.293
Sensitivity	.769	.851
Specificity	.438	.314
FN Errors	.001	.010
FP Errors	.021	.238
D-INC	.231	.138
T-INC	.542	.448
PPV	.974	.781
NPV	.999	.968
D Correct	.999	.988
T Correct	.955	.569

Appendix I-6

Federal ZCT / Three-position TDA (+/-4)

Two usable studies on the Federal ZCT with three-position TDA, using improved cutscores (+/-4 and -3) produced a combined unweighted average accuracy level of .939 (.028), with an inconclusive rate of .269 (.044). The weighted average of correct decisions for the Federal ZCT three-position model with improved cutscores, was .932 (.042) for criterion deceptive cases and .946 (.035) for criterion truthful cases. The weighted average inconclusive rates were .314 (.066) for criterion deceptive cases and .225 (.059) for criterion truthful cases.

Study	Krapohl (1998)	Harwell (2000)
Sample N	100	88
N Deceptive	50	60
N Truthful	50	28
Scorers	5	3
D Scores	250	180
T Scores	250	84
Total Scores	500	264
Mean D	-7.700	-6.194
StDev D	5.876	5.576
Mean T	6.300	5.113
StDev T	5.317	5.521
Reliability Kappa	-	-
Reliability Agreement	-	.990
Reliability Correlation	.900	-
Unweighted Accuracy	.929	.937
Unweighted Inconclusives	.264	.296
Sensitivity	.604	.689
Specificity	.768	.631
FN Errors	.068	.017
FP Errors	.032	.071
D-INC	.328	.294
T-INC	.200	.298
PPV	.950	.906
NPV	.919	.974
D Correct	.899	.976
T Correct	.960	.898

Appendix I-7

Positive Control Technique / Seven-position TDA*

Two studies on the Positive Control Technique produced a combined unweighted average accuracy level of .820 (.043), with an inconclusive rate of .292 (.045). The weighted average of correct decisions for the Positive Control Technique, was .679 (.080) for criterion deceptive cases and .962 (.032) for criterion truthful cases. The weighted average inconclusive rates were .333 (.066) for criterion deceptive cases and .250 (.063) for criterion truthful cases.

Study	Driscoll, Honts & Jones (1987)	Forman & McCauley (1986)
Sample N	40	38
N Deceptive	20	22
N Truthful	20	16
Scorers	1	1
D Scores	20	22
T Scores	20	16
Total Scores	40	38
Mean D	-2.000	-
StDev D	3.800	-
Mean T	6.600	-
StDev T	5.700	-
Reliability Kappa	-	-
Reliability Agreement	-	.800
Reliability Correlation	.840	.800
Unweighted Accuracy	.889	.777
Unweighted Inconclusives	.450	.131
Sensitivity	.350	.545
Specificity	.650	.750
FN Errors	.100	.318
FP Errors	.001	.063
D-INC	.550	.136
T-INC	.350	.125
PPV	.999	.897
NPV	.867	.702
D Correct	.778	.632
T Correct	.999	.923

* Driscoll, Honts & Jones (1987) used seven-position numerical scoring. Forman & McCauley used a non-numerical TDA approach, and is included here only for comparison purposes.

Appendix I-8

Relevant-Irrelevant Technique

One usable published study exists for the Relevant-Irrelevant techniques (Krapohl, Senter & Stern, 2005), which resulted in an unweighted accuracy rate of .732 (.044). Krapohl (2006) previously reported the accuracy of the RI technique at .83 with zero inconclusives, while including the results of Correa and Adams (1981) who reported an accuracy level of 100% in a laboratory study that employed methodology that does not reflect field practices.

Study	Krapohl, Senter & Stern (2005)
Sample N	100
N Deceptive	59
N Truthful	41
Scorers	1
D Scores	59
T Scores	41
Total Scores	100
Mean D	-
StDev D	-
Mean T	-
StDev T	-
Reliability Kappa	-
Reliability Agreement	.700
Reliability Correlation	-
Unweighted Accuracy	.732
Unweighted Inconclusives	.001
Sensitivity	.831
Specificity	.634
FN Errors	.169
FP Errors	.366
D-INC	.001
T-INC	.001
PPV	.694
NPV	.789
D Correct	.831
T Correct	.634

Appendix I-9

Zone Comparison Techniques / Rank Order Scoring System

Two usable studies on the ZCT with Rank Order TDA produced a combined unweighted average accuracy level of .886 (.036), with an inconclusive rate of .213 (.042). The weighted average of correct decisions for the Zone Comparison Technique with Rank Order TDA, was .885 (.050) for criterion deceptive cases and .887 (.053) for criterion truthful cases. The weighted average inconclusive rates were .020 (.061) for criterion deceptive cases and .225 (.057) for criterion truthful cases.

Study	Krapohl, Dutton & Ryan (2001)	Honts & Driscoll (1987)
Sample N	100	60
N Deceptive	50	30
N Truthful	50	30
Scorers	3	2
D Scores	50	30
T Scores	50	30
Total Scores	100	60
Mean D	-24.650	-20.900
StDev D	18.970	8.660
Mean T	7.400	13.400
StDev T	17.060	8.660
Reliability Kappa	-	-
Reliability Agreement	-	-
Reliability Correlation	-	.930
Unweighted Accuracy	.879	.906
Unweighted Inconclusives	.150	.317
Sensitivity	.720	.600
Specificity	.720	.633
FN Errors	.120	.033
FP Errors	.080	.100
D-INC	.100	.367
T-INC	.200	.267
PPV	.900	.857
NPV	.857	.950
D Correct	.857	.947
T Correct	.900	.864

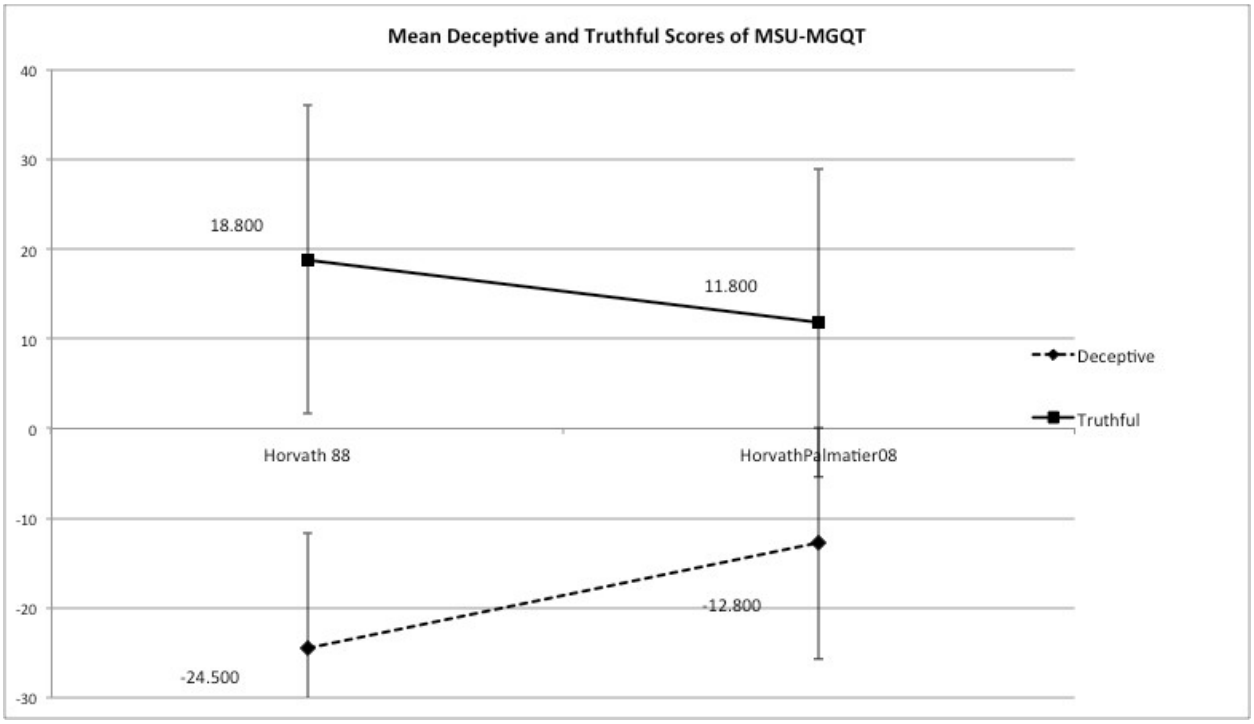
Addendum to the 2011 Meta-analytic Survey – MSU-MGQT

Raymond Nelson

This variant of the Modified General Question Technique was first described by Horvath (1988) and was replicated by Horvath and Palmatier (2008). Both studies were designed to investigate effects related to types of comparison questions. Both studies were conducted at the University of Michigan, and the name "Michigan State University Modified General Question Technique" (MSU-MGQT) is used herein. Operation and execution of the MSU-MGQT is, in some ways, not dissimilar to that for other MGQT formats, and the details can be found in the previously cited studies.

Horvath (1988) described a laboratory study involving a test question sequence similar to that described by Reid (1947), though including a fifth relevant question that was used to describe the guilty status of the participant. Data were evaluated by two blind evaluators who used a 7-position scoring model described as similar to the one used by Barland and Raskin (1975) with grand total decision thresholds set at +/-5. Unweighted decision accuracy of blind numerical scores was .871, with an inconclusive rate of .025.

Figure 1. Mean and standard deviations for the scores from truthful and deceptive samples in the two MSU-MGQT studies.



Horvath and Palmatier (2008) reported the results of another study of comparison question types, also involving the MSU-MGQT format, but also including a fifth relevant question that was used to describe the guilty status of the participant. Scoring tasks were completed used a 7-position

scoring model described as similar to the one used by Barland and Raskin (1975), with grand total decision thresholds set at +/- 6. Unweighted decision accuracy of blind numerical scores was .882, with an inconclusive rate of .167.

Figure 1 shows a mean and standard deviation plot of the scores of the sampling distributions of the included MSU-MGQT studies. A two-way ANOVA showed that the interaction of sampling distribution and criterion status was significant [F (1,92) = 8.7, (p = .004)]. The source of the interaction is at least partially attributable to the fact that scores for both deceptive and truthful participants were substantially stronger for the Horvath (1988) study, and both were closer to zero in the Horvath and Palmatier (2008) study. One-way ANOVAs showed no significant differences in the scores of the two studies for either the deceptive samples [F (1,46) = 0.231, (p = .633)] or truthful samples [F (1,46) = 0.147, (p = .703)].

Table 1 shows the summary of the two studies combined. Table 2 shows the profile and statistical confidence intervals for the criterion accuracy metrics. Table 3 shows a summary of the individual studies.

Table 1. Summary MSU-MGQT studies.	
Number of studies	2
Total N	50
N Deceptive	25
N Truthful	25
Number of Examiners/Scorers	3
Total Scores	100
D Scores	50
T Scores	50
Mean D	-18.650
StDev D	-
Mean T	15.300
StDev T	-
Reliability Kappa	-
Reliability Agreement	0.95
Reliability Correlation	0.92

Table 2. Criterion accuracy and confidence intervals for MSU-MGQT studies.	
Unweighted Average Accuracy	.877 (.049) {.781 to .972}
Unweighted Inconclusives	.110 (.043) {.025 to .195}
Sensitivity	.820 (.068) {.686 to .954}
Specificity	.740 (.072) {.599 to .881}
FN Errors	.120 (.064) {.001 to .246}
FP Errors	.100 (.060) {.001 to .218}
D-INC	.060 (.047) {.001 to .153}
T-INC	.160 (.073) {.017 to .303}
PPV	.891 (.065) {.765 to .999}
NPV	.860 (.075) {.713 to .999}
D Correct	.872 (.068) {.739 to .999}
T Correct	.881 (.072) {.740 to .999}

Table 3. Summary of individual studies using the MSU-MGQT.		
Study	Horvath (1988)	Horvath & Palmatier (2008)
Sample N	20	30
N Deceptive	10	15
N Truthful	10	15
Scorers	2	1
D Scores	20	30
T Scores	20	30
Total Scores	40	60
Mean D	-24.500	-12.800
StDev D	-	17.200
Mean T	18.800	11.800
StDev T	-	12.900
Reliability Kappa	-	
Reliability Agreement	0.95	
Reliability Correlation	0.920	
Unweighted Average Accuracy	0.871	0.882
Unweighted Inconclusives	0.025	0.167
Sensitivity	0.900	0.767
Specificity	0.800	0.700
FN Errors	0.100	0.133
FP Errors	0.150	0.067
D-INC	0.000	0.100
T-INC	0.050	0.233
PPV	0.857	0.920
NPV	0.889	0.840
D Correct	0.900	0.852
T Correct	0.842	0.913

The combined decision accuracy level of the MSU-MGQT studies, weighted for sample size and number of scorers, was .877 with a combined inconclusive rate of .11. Reliability for MSU-MGQT exams, expressed as the concordance rate for categorical decisions, was reported by Horvath (1988) at .95, with a correlation coefficient of .92 for numerical scores of the two evaluators.

References

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

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*, 198-209.

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.

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.