Received: 10 June 2020 | Revised: 4 December 2020 | Accepted: 8 December 2020

DOI: 10.1002/acp.3779

SPECIAL ISSUE ARTICLE

WILEY

A comprehensive meta-analysis of the comparison question polygraph test

Charles R. Honts[1] | Steven Thurber[2] | Mark Handler[3]

[1]Department of Psychological Science, Boise State University, Boise, Idaho, USA

[2]Child and Adolescent Behavioral Health Services, Minnesota Department of Human Services, Saint Paul, Minnesota, USA

[3]Converus, Inc., Lehi, Utah, USA

Correspondence
Charles R. Honts, Department of Psychological Science, Boise State University, 1910 University Drive MS-1715, Boise ID 83725-1715, USA.
Email: chonts@boisestate.edu

Abstract

We conducted a meta-analysis on the most commonly used forensic polygraph test, the Comparison Question Test. We captured as many studies as possible by using broad inclusion criteria. Data and potential moderators were coded from 138 datasets. The meta-analytic effect size including inconclusive outcomes was 0.69 [0.66, 0.79]. We found significant moderator effects. Notably, level of motivation had a positive linear relationship with our outcome measures. Information Gain analysis of CQT outcomes representing the median accuracy showed a significant information increase over interpersonal deception detection across almost the complete range of base rates. Our results suggest that the CQT can be accurate, that experimental studies are generalizable, and no publication bias was detected. We discussed the limitations of the field research literature and problems within polygraph profession that lower field accuracy. We suggest some possible solutions.

KEYWORDS
comparison question test, deception detection, polygraph, psychophysiological deception detection

1 | INTRODUCTION

Lying is a ubiquitous human behavior. In a now classic study DePaulo et al. (1996) reported that college students lied twice a day in their interactions (conversations that lasted more than 10 min) with others, that was in approximately one-third of their daily interactions. DePaulo et al. (1996) also reported that college students lied to 38% of the people with whom they interacted. Subsequent research has consistently demonstrated the high frequency and ubiquity of lying (Hartwig & Bond, 2014). Although many of these lies are trivial, clearly many are not and, if successful, lies can have devastating impacts on relationships, societies, employment, criminal justice, politics, public health, and national security (Granhag & Strömwall, 2004).

The commonality of lying might not be so serious if people could detect lies inter-personally. Unfortunately, a substantial body of research indicates average people have a truth bias (i.e., they tend to judge people as truthful) and are only about 54% accurate. Moreover, professionals (e.g., police officers) charged with making credibility judgments are no more accurate, showing approximately the same accuracy but with a lie bias (they tend to judge people as liars). The research findings that indicate poor accuracy for interpersonal deception detection seem to be well established science and interested readers are referred to Vrij, Mann, et al. (2008) for an overview and to Hartwig and Bond (2011, 2014) for meta-analyses.

One response to poor interpersonal deception detection accuracy is to look to technology for a solution. One of the oldest technological

Portions of these findings were presented as a paper (Honts & Thurber, 2019a) at the annual meeting of the American Psychology Law Society, Portland, Oregon, USA. The authors would like to thank Adela Anderson for her help in editing the completed manuscript.

approaches to credibility assessment is the use of physiological measures to make inferences about the credibility of people's statements (Munsterberg, 1908). In the United States such testing came to be known as polygraph testing. For a history of development in polygraph testing see Trovillo (1939a, 1939b). Raskin (1986) and Raskin and Honts (2002) provide descriptions of the development of modern scientific research on the most commonly used forensic polygraph test, the Comparison Question Test (CQT).

Polygraph tests are psychological tests that are used worldwide as a screening tool in law enforcement, national security, and private employment. Polygraph tests are also widely employed as forensic tests in investigations and in legal proceedings. The largest professional association of polygraph examiners, the American Polygraph Association (APA), shows more than 2800 members from 58 countries (APA, 2019a). Estimates indicate that there are more than 8000 polygraph examiners operating in China alone (Zhang, 2011). While the critics of the polygraph (e.g., Iacono & Ben-Shakhar, 2019) acknowledge that the polygraph is used in some countries outside the United States, they often fail to acknowledge the broad international use of the polygraph. A brief series of online searches revealed resident polygraph examiners in 65 countries, with 24 professionally recognized training schools and 12 professional organizations all with international memberships. We have provided documentation of the international use of the polygraph our supplementary information Archive A (Data S1).

With regard to the use of polygraph tests results in courts of law there is a great deal of variability. In the United States, polygraph tests are admissible in courts of law in about half the states under stipulation (Iacono & Ben-Shakhar, 2019). Since 1975, the State of New Mexico has allowed the general admission, without stipulation, of the results of polygraph tests under the New Mexico Rule of Evidence 11–707 (Raskin, 1986, also see, Lee et al., v. Martinez et al., 2004 for a reaffirmation of admissibility under the *Daubert* standard). The U.S. Federal courts may also admit the results of polygraph tests at the trial judge's discretion (U.S. v. Scheffer, 1998) under the rules from *Daubert v. Merrell Dow Pharmaceuticuals* (1993). Beyond the issue of admissibility of polygraph tests at trial as a practical matter they are used throughout the criminal justice systems of many countries to influence decisions about the continued investigation of potential suspects, the decision to interrogate suspects, the decision to charge crimes, and in sentencing hearings. Additionally polygraph tests are used in a few countries to make decisions about the continuation or modification of conditions of treatment, parole or probation for persons convicted of sex-related offenses (Grubin et al., 2019).

As in the United States the status of the polygraph in international courts is mixed. It appears that in most countries polygraph testing is used primarily as an investigative forensic and security tool. However, there are a number of countries that allow for the admission of polygraph test results as evidence in their courts of law. Most recently, Belgium (Philippe, 2020) has determined that the results of CQT polygraph examinations may be used as evidence in criminal cases. Polygraph tests results have been admissible in Poland since 1976 (Widacki, 2007). In Europe, it is also noted that in a few cases polygraph results were presented in courts in Finland, Norway, Sweden (Meijer & von Koppen, 2008) and Lithuania (Kraujalis et al., 2007). In Asia, polygraph test results are admissible in China in civil but not criminal cases (Guodong, 2020). In South America, polygraph test results are admissible in Colombia (Bermudez & Arias, 2011).

Despite the widespread application of polygraph testing, and the important role it plays in national security, forensics, and criminal justice around the world, polygraph tests have received relatively little attention in academic psychology and often, that attention has been in the form of negative commentary. Most of the published polygraph research has focused on forensic uses of the various polygraph techniques. There are two qualitatively different families of polygraph tests used in forensic application. The first family of tests are designed to detect hidden information. Those tests are known variously as Guilty Knowledge Tests or Concealed Information Tests. Although such tests have good psychometric qualities and have been shown to be accurate in experimental settings, their accuracy has never been established in field settings where their necessary preconditions rarely exist (Podlesny, 1993), and where there is an abject lack of theory about what details of a crime scene are likely to be remembered (Honts, 2004). Moreover, the existing field data indicate high false negative rates (Elaad et al., 1992; and reviews by Honts, Raskin, et al., 2008 and Vrij, 2008). Japan is the only country were the CIT is widely applied in criminal investigations (Matsuda et al., 2019). In Japan 80–100 examiners conduct about 5000 test a year (Hira & Furumitsu, 2002; Matsuda et al., 2019). Although 5000 tests might seem to be a relatively large number of tests that must be contrasted with the number of criminal acts investigated. In 2018 there were 817,338 criminal acts investigated (Osumi, 2019). Thus, assuming that 5000 CIT examinations were administered, then the CIT was used on only 0.6% of criminal cases in Japan in 2018. This indicates the use of the CIT is extremely rare even in the one country that is focused on the forensic use on the information tests in criminal investigations.

Worldwide the most commonly used polygraph test, the Comparison Question Test (CQT), takes a direct approach to forensic credibility assessment by asking simple accusatory questions. Honts and Thurber (2019b) recently noted that the CQT comes in several variants with generally common characteristics. During testing, the subject's autonomic physiology (usually, respiration, electrodermal activity, relative blood pressure, and often peripheral vasomotor activity) is monitored while the subject answers a series of questions. There are two categories of critical questions (usually three of each) in the series. Relevant questions are semantically simple questions that directly address the matters under investigation. Comparison questions are designed and presented in such a way that every subject lies, or is at least uncertain about their truthfulness in their response to them during the test. Subjects' physiological responses are expected to show a full cross-over interaction between their guilt status and the critical question type. That is, subjects who are deceptive to the relevant questions are expected to show larger physiological responses to relevant questions as compared to comparison questions. Innocent subjects who are being truthful to the relevant questions are expected

to show the opposite pattern, with physiological responses to comparison questions being larger than those to relevant questions.

The CQT research literature was the subject of a number of reviews over the years. Typical of those reviews are: Kircher et al. (1988), Raskin et al. (1997), Iacono and Lykken (1997), National Research Council of the National Academy of Sciences (NRC) (2003), Honts (2004), Vrij, Mann, et al. (2008); APA (2011); and Raskin et al. (2014). There is variation across the reviews, but nevertheless they generally produced overall accuracy estimates of over 85%.

However, only one of those reviews used meta-analytic techniques to examine moderator variables (Kircher et al., 1988). Kircher et al. (1988) sampled only experiments and analyzed only 14 studies. The small number of studies considered by Kircher et al. (1988) reflected the size of the experimental literature at the time and their criteria for inclusion. They found significant moderator effects of Subjects (Student vs. Other), Incentives (Minimal vs. Stronger) and Decision Policy (Standard Field and Other). All three variables were found to be predictive of accuracy, but all three showed high covariations within the studies and analyses that examined their relative association with accuracy were not reported. The moderator effects in Kircher et al. (1988) are thus confounded and difficult to interpret.

Unfortunately, all of the prior reviews can be criticized for selective study choices and, with the single exception of Kircher et al. (1988) a lack of meta-analytic scrutiny. Nevertheless, the reviewers sometimes reached conclusions that hypothesized or even assumed powerful moderator effects. The recent publication by Iacono and Ben-Shakhar (2019) is particularly egregious in that regard. Iacono and Ben-Shakhar (2019) focus their review on the National Research Council of the National Academy of Sciences (NRC) (2003) review of polygraph testing and ultimately conclude, "In 2003, the National Academy of Sciences concluded that polygraph testing had a weak scientific basis and unknown error rate. Analysis of research conducted over the last 15 years indicates that these conclusions remain valid" (p. 86). Iacono and Ben-Shakhar (2019) base their conclusion on the following lines of argument: (1) Many authors have misrepresented the NRC analysis as indicating high accuracy for the CQT. (2) A thought experiment, that Iacono and Ben-Shakhar (2019) treat as evidence, that shows a possible set of factors that could result in a test with chance producing high accuracy in a study. (3) An attack upon the venues where polygraph research has been published rather than on the quality of the research. (4) A broad dismissal of experimental studies as providing a useful index of the CQT in application. (5) An implicit assumption that the contingency associated with the outcome of a CQT examination is a powerful moderator of the test's accuracy. Finally, (6) an assertion that there is a lack of theory underlying the test. Iacono and Ben-Shakhar (2019) state these arguments as fact, but most are unsupported speculation, and they deserve critical and empirical examination. Those arguments have been analyzed elsewhere and they were found to lack merit (Honts & Thurber, 2019a, 2019b).

Iacono and Ben-Shakhar (2019) arguments 3, 4, and 5 assert that accuracy and venue of publication are correlated and they generally dismiss experimental research as not generalizable to the CQT in field applications. These arguments are made as if they were statements of factual conditions, but they were presented without empirical evidence. However, such issues of eternal validity represent potential moderator variables for meta-analysis. Interestingly, Iacono and Ben-Shakhar's (2019) arguments 3, 4, and 5 are some of the same arguments as the criticisms raised about experimental research on interpersonal deception detection that were a motivating factor for one of the meta-analysis published by Hartwig and Bond (2014).

Hartwig and Bond (2014) reviewed the stated concerns for the external validity of the interpersonal deception detection research. That review contained many striking similarities to the criticisms of psychophysiological deception detection with concerns about strong moderator effects of experimental venue, subject population, and the strength of outcome contingency and the inadequacy of theory. Hartwig and Bond conducted a meta-analysis of interpersonal deception detection to address the concerns about limited external validity of deception detection research. Specifically, they addressed the following potential moderator variables, Liar's Demographic Background (student, other), Motivation to Lie (None, Moderate, High), Social Setting (Monolog, Interview, Interaction), Deception Medium (Face-to-face, Other) Affective State (Strong Emotion, No Emotion), and Content of Lie (Feelings, Facts). Hartwig and Bond (2014) reported non-significant findings for all of the potential moderator variables. Hartwig and Bond (2014) conclude, "The primary finding of our analysis is that lie detectability remains stable across contexts. Notably, the finding on external validity mirrors those of meta-analyses that have compared laboratory research to field research in other domains" (p. 667).

2 | AIMS OF THE PRESENT STUDY

As with Hartwig and Bond (2014), our primary aim was to address concerns about the external validity of the psychophysiological deception detection research. A secondary interest was to provide a meta-analytic assessment of the ability of the CQT to detect deception. We were also motivated by the fact that there was a dramatic increase in CQT research since the publication of the NRC (2003) report. Our approach was to make our inclusion criteria as broad as possible so that we would be able to test the full range of potential effects of a number of moderator variables that the critics have said are important and also avoid any suggestion of bias in our sampling of cases. Our broad selection criteria were adopted with the knowledge that we would be including studies that previous reviewers found to have substandard methods. We realized that this decision would likely have an impact on our effect size estimate. However, we put our focus on inclusion of as many studies as possible so that we could examine widest possible range or our prospective moderator variables in our assessment of external validity.

As with Hartwig and Bond (2014), there were two potential outcomes of this meta-analysis. First, it may be that the critics are correct and there are moderators that are strongly associated with the accuracy of CQT polygraph examinations. It may be that under real world motivational and testing settings, CQT tests are more, or less, accurate than in the laboratory. Alternatively, it may be that

psychophysiological deception detection is stable across a range of potential moderator variables in a manner similar to the findings of Hartwig and Bond (2014). As Hartwig and Bond (2014) noted, the implications of these two outcomes for research and application are quite different. If significant moderators were to be identified, end users in applied settings would have better information upon which to base judgments about the weight to be applied to CQT test outcomes in their various applications. The existence of significant moderators would also provide a guide for people conducting research about how to increase the external validity of their experimental paradigms. However, if the accuracy of the CQT is found to be stable across a range of moderator variables, or only weakly impacted by them, then the criticisms of experimental research on the CQT for weak external validity would appear to be unfounded and brought into question. The latter finding would suggest that the results of experimental research on the CQT should not be dismissed as a laboratory artifact and those results should be given serious weight in estimating the validity of the CQT.

3 | ACCESSING THE ACCURACY OF THE CQT

Standard practice with the CQT poses an unusual problem for traditional effect size analysis where the focus is usually on a binary outcome variable. With a CQT the standard outcome is not binary but instead has three levels that are based upon an underlying continuum of scores. The standard outcomes in a CQT are Truthful, Inconclusive, or Deceptive. That three-level decision continuum generally follows an underlying interval scale of numerical scores in the same way the terms, cold, medium, and hot follow underlying interval or ratio scales of temperature. Across the reviews a number of approaches have been taken to quantifying the accuracy of the CQT. The NRC (2003) used Area Under the Curve (*AUC*) as an index of accuracy and ignored inconclusive outcomes. Honts and Schweinle (2009) used Information Gain (Wells & Olson, 2002) and provided three information gain curves for truthful, inconclusive and deceptive outcomes. Other studies have simply calculated weighted means from a 2 (Innocent or Guilty) by 3 (Truthful, Inconclusive, Deceptive) contingency table (Raskin et al., 1997) or some variation thereof (Iacono & Lykken, 1997). The use of three outcomes thus increases the complexity of interpretation of the aggregated data. In response to this problem Kircher et al. (1988) developed and used a single measure of accuracy they called a Detection Efficiency Coefficient (r_{dec}). The r_{dec} is simply a correlation between the binary reality state, Guilty or Innocent, coded −1 and 1 respectively, and test outcomes, Deceptive, Inconclusive, or Truthful, coded −1, 0 and 1 respectively. The r_{dec} thus is sensitive to the impact of inconclusive outcomes where their occurrence reduces the value of the r_{dec}, but not by as much as an error. We adopted r_{dec} as our primary measure of accuracy for the CQT in our analyses. However, we also planned to look at the more traditional analyses of sensitivity, specificity and *AUC*.

Once the effects of moderators were known and an estimate of CQT accuracy was obtained from the meta-analysis, we planned to assess the added value of having a CQT test outcome as compared to the information that is readily available to individuals attempting to assess credibility in an interpersonal setting. To be useful in application, a diagnostic test must provide information beyond what is available without the test. In the credibility assessment situation, there are two sources of information that are available before conducting the test. One of those is the interpersonal decision of credibility based upon the individual's overt behavior during an interaction or formal questioning. Unfortunately, credibility assessments made interpersonally are consistently estimated to be about 54% accurate (Vrij, 2008).

An important, and often overlooked, source of information in forensic decision making is the underlying base rate of the target condition (Honts & Schweinle, 2009). In the settings where polygraph testing is used the base rate of guilt may vary dramatically. For example, in the national security employment screening situation, the base rate of guilt (i.e., the probability of a given subject being an agent of a foreign government or terrorist organization) is likely to be very low. In some forensic settings, the base rate may be relatively low, for example when there are a number of suspects and the polygraph is used to reduce the size of the suspect pool. In other forensic polygraph settings the base rate of guilt may be high, for example after a long investigative process has narrowed the pool of suspects to one or two individuals, or when an individual has been formally charged with a crime. What is needed is a method to evaluate the usefulness of a test across the range of base rates so that end users of the information can estimate how much weight to give a test outcome and make judgments about when the test may be useful. Fortunately, there is such a method. First described by Wells and Lindsay (1980) and expanded by Wells and Olson (2002), Information Gain (IG) analysis uses a Bayesian-based approach to describe the impact of base rates on the information provided by eyewitness identification procedures. Honts and Schweinle (2009) adapted the Wells and Olson IG procedures for use with the CQT and its three levels of outcome. We used IG analysis to evaluate the applied value of the CQT based upon the meta-analytic estimates of accuracy of the CQT in comparison with interpersonal deception detection.

4 | METHOD

4.1 | Literature search procedures

For our database we attempted to find all of the available English-language studies of CQT accuracy conducted in forensic settings or paradigms. We began our search with the first author's personal library. The first author has been involved in conducting research on the CQT since 1980. Computer-based searches were then conducted of Criminal Justice Abstracts, Defense Technical Information Center (DTIC), Google Scholar, JSTOR, ProQuest Theses and Dissertations Global, PsychINFO, and PsychARTICLES. Searches were made with the following terms: Comparison Question Test, CQT, Polygraph, Psychophysiological Deception Detection, Psychophysiological Detection of Deception, and PDD. We also reviewed the complete volume of the journal *Polygraph*, now known as *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field*

Practice. The reference sections of articles were searched as they were obtained, and cross indexed against studies already in the database. References not in the database were obtained and added to the database. The search for additional studies was closed on July 1, 2018.

4.2 | Criteria for the inclusion of studies

Our goal for this study was to include every English-language report with sufficient information for analysis. Studies were included if they met the following criteria: (1) The study addressed the validity of the CQT in a setting or paradigm that addressed a focused specific issue or issues (broad pre-employment screening tests were not included). (2) Sufficient information was available to determine frequencies for the various test outcomes. (3) In field studies, there was a description of the criterion used to classify cases as Innocent or Guilty. (4) There was sufficient information to determine the method used for evaluation of the data and the generation of an outcome. (5) At least two of the standard physiological measures (respiration, electrodermal activity, relative blood pressure, or vasomotor activity) were used in the collection of data. (6) The study did not duplicate data and analyses already in the database. (e.g., the same data in a grant report and a publication would be represented in the database by only the publication.) (7) Study data were collected from actual subjects and were not based upon bootstrapping, Monte Carlo, or other statistical estimation methods.

4.3 | Samples of interest

Our unit of analysis was a sample of data from liar (Guilty) and truth-teller (Innocent) subjects analyzed with the same scoring technique. In some reports, the same sample of subjects was evaluated by multiple evaluators. In some of those reports, only averages were reported. In that case the averaged data were used in our analysis. When averages were used the number of tests averaged was retained as the N and not the number of scorings. In some reports, data were provided for multiple scorings of the same data. For those studies we selected the data from one evaluator by random selection and used only the data from that evaluator in our analysis. In some reports the data were scored with different scoring methods. One exemplar of each scoring method from a study was included in the data for this study.

4.4 | Justification and retention of moderator variables

Potential variables for coding were selected by several methods. Initially we began with the relevant variables coded by Hartwig and Bond (2014) for their meta-analysis of interpersonal deception detection and with the variables coded in the one existing meta-analysis of the CQT (Kircher et al., 1988). We also looked for variables that were, and are, the topic of continued scientific debate about their theoretical importance for the understanding of CQT research (e.g., Honts, 2014; Iacono & Ben-Shakhar, 2019). That initial set of moderator variables included, the sampling frame for subject selection, contingent motivation associated with test outcome, the study's status as an experiment, peer-reviewed status of the report,[1] and examiner orientation (defense/law enforcement). For non-experimental studies we also collected data on how the determination of truth status was made and the setting where the data were collected (e.g., the workplace, criminal justice, national security) and the nature of the topics addressed. Basic data were collected concerning, the number of persons tested, subject age, subject sex, examiner characteristics and the test outcome frequencies. We also coded the following variables that are of interest in developing evidence based best practice standards for the profession: CQT type, number of issues addressed, and scoring method. Unfortunately, a number of potential moderator variables we initially examined were not included in the meta-analysis because there were an insufficient number of studies (>65%) reporting the data (e.g., age, years of education, specific type of crime in field studies, and years of experience as an examiner) or there was insufficient variability for meaningful analysis as a moderator (e.g., type of mock crime in experiment, method of confirmation of guilt status in field study, nature of the topics addressed in field studies [defense vs. law enforcement orientation], and number of issues in the CQT).

Although we did have a sufficient number of studies that addressed scoring method we ultimately did not include it in this analysis because a number of the scoring methods did not have sufficient studies for meaningful analysis and we felt that grouping the low frequency methods into an Other category would be meaningless. Moreover, the two methods with sufficient representation for meta-analysis had been tested on the same data set (Utah and US Federal 7-position; Honts, Amato, et al., 2000) and total scores were not found to be significantly different and were therefore unlikely to have any value as a moderator.

4.5 | Variables retained for the meta-analysis

The following variables had sufficient data for meaningful meta-analysis and were retained for analysis: Setting, Subject Source, Motivation, Issues, Type of Comparison Question, Peer Review, Subject Sex, and Age. Those variables were coded as follows: Setting contrasts experiments with field studies. Subject Source indexed where the sample of subjects was obtained and had four levels: students, community, work, and criminal justice. Motivation indexed contingencies that were associated with test outcome and had three levels: nothing, something awarded, and real-world consequences. Issues coded two levels: Single versus Multiple and indexed if the polygraph examination addressed only a single incident or multiple independent incidents. Type of Comparison Question indexed the two types of comparison questions in common use in field practice, probable lie and directed lie. Peer review indexed if the report was peer reviewed. Subject Sex indexed if only males, only females, or if a mix of sexes was included. When available we also recorded frequency data for sex. Average subject age in years was recorded.

4.6 | Meta-analytic procedures

The meta-analytic statistics were calculated using *Comprehensive Meta-Analysis* (*Version 3*; Borenstein et al., 2014). Other statistical analyses were calculated with *SPSS* (IBM, 2017). Information Gain analyses were calculated with the Excel spreadsheet developed by Honts and Schweinle (2009).

The Detection Efficiency Coefficient (r_{dec}; Kircher et al., 1988) a point biserial statistic was our primary effect size estimate. A logit transformation of event rates was used for computation of point estimates and confidence intervals. There are two models that can be selected for data analyses. The fixed effects model assumes a single, true effect size among aggregate, independent studies. The random effects model posits variability among the investigations. The statistic I^2 indicates the percentage of heterogeneity among the studies; elevated heterogeneity supports the appropriateness of the use of the random model. We used the random effects approach for all meta-analysis computations.

A potential source of publication bias is that smaller studies tend to produce inordinately large effect sizes, and correspondingly, this disproportionate impact is not balanced by the inclusion of smaller investigations with extreme, non-significant effect sizes. This possible bias was evaluated via a funnel plot (Duval & Tweedie, 2000). In the absence of bias, the funnel plot would show a symmetrical distribution of effect sizes around a summary value.

5 | RESULTS

We obtained and examined 173 documents, of which 112 met our selection criteria and were coded for analysis. Sixty-one documents did not meet our criteria for the following reasons: one was a meta-analysis, eight were duplicate studies, 34 were not CQT studies, 15 contained inadequate information for outcome calculation, and 3 were reports concerning individual cases. From the 112 selected documents, we coded 221 datasets that contained 16,278 polygraph decisions. However, many of those datasets contained reliability data (i.e., different people scoring the same data with the same scoring system). When the redundant data were removed there were 138 data sets that represented independent decisions. Those 138 data sets contained 11,053 decisions. However, three of the data sets contained only guilty subjects. Those three data sets were not available for analyses that assessed both innocent and guilty subjects, but they were retained for sensitivity analysis. Notably, 59 (43.0%) of the data sets were published or reported after NRC's (2003) close of data collection and were thus not included in the NRC review.

5.1 | Characteristics of the research literature

A summary of the research literature based on our coding is presented in Table 1. The number of subjects in these studies varied widely, from a low of seven to a high of 500. There was insufficient age data to provide a meaningful estimate of subject age. Only one study reported a focus on juveniles (Craig et al., 2011) and it seems safe to assume that the other studies tested persons over the age of 18 years. Similarly, it was difficult to develop information on participation rates by sex. Fifty-nine (43.7%) of the samples had no information about the sex of their subjects. Seventeen (12.6%) of the samples were male only, while 59 (43.7%) of the samples indicated the participation of both males and females. Within the mixed samples that reported frequency data for sex, (16 samples did not), the Mean number of male subjects was 53.7 and the mean number of female

TABLE 1 Characteristics of the research literature

Quantitative Variables				
Variable	**Minimum**	**Maximum**	**Mean**	**Median**
N Subjects	7	500	73.32	60.5
Male	6	257	54.28	42.0
Female	0	167	20.84	10.0
DEC	0.10	0.99	0.65	0.66

Categorical Variables		
Variable	**# (%) of subject samples**	**Percent coding agreement**
Motivation		98%
None	27 (19.6%)	
Some	57 (41.3%)	
Real	53 (38.4%)	
Setting		98%
Experiment	88 (63.8%)	
Field	50 (36.2%)	
Subject Source		94%
Student	31 (22.5%)	
Community	36 (26.1%)	
Work	16 (11.6%)	
Criminal Justice	53 (38.4%)	
Issues		93%
Multiple	26 (19.3%)	
Single	100 (74.1%)	
CQT Type		
Probable-Lie	113 (81.9%)	95%
Directed-Lie	20 (14.5%)	
Both	5(3.6%)	
Peer-Reviewed		95%
Yes	104 (75.4%)	
No	33 (23.9%)	
Sex		
All Male	17 (12.6%)	
All Female	0 (0%)	
Mixed	59 (43.7%)	
Unknown	59 (43.7%)	

Note: Percentages are based on 138 datasets. Due to missing data the results may not sum to 100%.

subjects was 30.05. The few studies that have explicitly tested for sex differences have failed to reveal any significant effects (e.g., Honts, Raskin, et al., 1994). A majority of the data sets (86, 63.7%) were from experiments. Within the experiments 31 (36%) were student samples, 37 (43.0%) were community samples, 15 (17.4%) were work samples, and 5 (5.8%) were samples from a forensic setting (e.g., a prison population, Raskin & Hare, 1978).

5.1.1 | Reliability

Table 1 also contains reliability data for the coding of the moderator variables. The first and third author independently coded the first 97 data sets obtained in our analysis representing 70% of our retained data sets. Those data were analyzed for agreement in coding. As is shown in Table 1, agreement was high for all the moderators and ranged from a low of 93% with Issues to a high of 98% with Motivation and Setting. A calibration of disagreements was made between the two evaluators and the consensus coding was retained for analysis. A significant delay in analyzing the data resulted in us reopening the search for studies in early 2018. An additional 41 data sets were obtained and were coded by the first author.

5.2 | Results of the meta-analysis of r_{dec}

We were able to calculate r_{dec} (Kircher et al., 1988) for 135 of the 138 datasets and those 135 values were subjected to meta-analysis using a random effects model. All confidence intervals here were calculated at 95%.

The obtained meta-analytic effect size for r_{dec} was 0.694 [.66, .79], $p < 0.0001$. That effect size converts (Salgado, 2018) to a Cohen's $d = 1.92$, and an $AUC = 0.91$. Although these values appear to be close replications of the NRC (2003) results, our estimate of *AUC* was reduced by the inclusion of inconclusive outcomes while the NRC estimate of *AUC* did not consider inconclusive outcomes. We also calculated the Cohen (1988) U_3 index to be 0.973. That value of U_3 indicates that the upper half of the Innocent population exceeds 97.3% of the members of the Guilty population. This results in an Improvement Index value of 47.3%, a value that represents the difference in percentile rank of an average Guilty subject and an average Innocent subject in their respective distributions (What Works Clearinghouse, 2008).

With regard to the effect size obtained for r_{dec}, Cohen (1988, 1992) indicates that an effect size *r* of .50 and above is considered "large." In the binomial effect size approach of Rosenthal (Rosenthal, 1983; Rosnow & Rosenthal, 2003), an r_{pb} of .00 yields an equal percentage, 50/50 for success (e.g., true positives) events over failure (e.g., false positive) events. When the r_{pb} is at the level of .50, the indication is that there is a "separation" between success and failure of 75% and 25% respectively. With an r_{pb} of .60, that separation increases to 80% versus 20%. As indicated, the obtained point estimate or summary r_{dec} of .694 in the current data set was in Cohen's large effect classification. In the current study, this suggests a percentage of classification accuracy of well over 80% (Rosenthal, 1983). It is noteworthy that the obtained funnel plot was symmetrical about the effect size mean, indicating that smaller studies with larger sampling error still displayed a broad range of values toward the bottom of the funnel graph. This symmetry was further corroborated by the trim and fill procedure in that no studies had to be inserted to improve that symmetry. A figure illustrating the Funnel Plot is provided in our Figure S1.

5.2.1 | r_{dec} and heterogeneity

The extent to which disparities exist among the obtained effect sizes is an object of concern. The meta-analysis returned an I^2 of 92.63. That value represents real differences in the effect sizes, unrelated to sampling error. We followed the recommendations of Borenstein et al. (2009) for dealing with effect size variability. First, because heterogeneity (I^2) was over 50%, the random model was correctly employed. Second, each effect size was weighted by Tau (T), the "true" standard deviation of the effect sizes in the DEC metric units ($r_{dec} = .398$). The pooled result yielded a point effect size estimate and confidence interval that does not include zero. The practical significance is that despite the varying effect sizes, the true effect size for r_{dec} is almost certainly positive and substantial.

5.2.2 | Moderator variable effects with r_{dec}

A moderating variable is one that affects the strength or direction of an outcome or relationship (Shadish & Sweeney, 1991). In meta-analyses, a moderator will influence the magnitude of an effect size. Table 2 lists the moderators we tested and the statistical results with r_{dec}. For each moderator subset in the table, we list an *R* for deception detection. The *R* corresponds to a weighted mean Fisher's Z_R for the effect in question. Also listed is a Q-statistic that tests the significance of each moderator variable. Most notable was the significant effect for Motivation, $Q = 333.15$, $p < .001$, indicating that as motivation increased detection accuracy as indexed by the r_{dec} also increased. A separate analysis for linearity between Motivation and r_{dec} was significant, $F(1, 132) = 15.27$, $p = .001$, while a test for deviations from linearity was not, $F(1, 132) = .279$, *ns*. A similar pattern of results was seen with Setting (Experiment vs. Field), and Source (Students, Community, and Forensic), but with much smaller values for *Q*, of 12.12 and 17.77, respectively. This is not surprising as those three moderators were highly correlated with each other, Motivation versus Setting, $r = .83$, $p < .01$, Motivation versus Source $r = .70$, $p < .01$, and Setting versus Source, $r = .83$, $p < .01$.

Other moderators were also found to have significant effects. There was a significant moderator effect associated with CQT Type, $Q = 9.16$, $p = .01$. The *R* value for probable lie tests was .71 while the *R* for directed lie tests was .61. An examination of the confidence intervals indicates that there was much larger variability in the relatively small sample of directed lie tests. The relatively small value of *Q* for this moderator and the similarity of the *R* values suggests that these differences are likely of little applied importance. Peer reviewed studies were significantly more

TABLE 2 Results of the meta-analysis of moderator variables on the r_{dec}

Moderator	Level	N	R	95% CI	Z	*p* <	I2	Q
Motivation	None	26	0.61	0.53, 0.68	11.63	.001	86.67	333.15, *p* < .001
	Some	57	0.65	0.61, 0.69	21.42	.001	71.66	
	Field	51	0.76	0.71, 0.81	15.59	.001	95.17	
Setting	Exp	85	0.64	0.60, 0.67	23.85	.001	78.75	12.15, *p* < .001
	Field	50	0.77	0.71, 0.81	15.57	.001	95.27	
Source	Student	30	0.59	0.53, 0.65	14.36	.001	69.44	17.77, *p* < .001
	Com	36	0.68	0.63, 0.73	16.89	.001	80.89	
	Work	16	0.61	0.50, 0.69	9.07	.001	83.24	
	CJ	51	0.76	0.70, 0.81	15.40	.001	95.24	
Issues	Single	28	0.64	0.58, 0.70	13.95	.001	83.23	*3.58, ns*
	Multiple	102	0.71	0.65, 0.74	19.67	.001	93.03	
CQT Type	PL	110	0.71	0.67, 0.75	21.03	.001	93.17	*9.16, p = .01*
	DL	20	0.61	0.53, 0.68	12.00	.001	74.83	
	Both	5	0.60	0.51, 0.68	9.98	.001	48.14	
Peer Review	Yes	101	0.71	0.66, 0.75	20.17	.001	92.71	*72.09, p = .001*
	No	33	0.64	0.57, 0.69	14.72	.001	85.17	

Abbreviations: CI, confidence interval; CJ, criminal justice; Com, community; DL, directed lie; I^2, % heterogeneity; PL, probable lie; Q, total between group variance.

TABLE 3 Results of the meta-analysis of experimental study moderator variables on the r_{dec}

Moderator	Level	*n* of studies	*R*	95% CI	*Z*	*p* <	I2	Q
Motivation	None	26	.61	[.53, .68]	11.63	.001	86.07	2.53. *ns*
	Some	57	.66	[.61, .97]	20.63	.001	75.72	
Source	Student	30	.60	[.53, .65]	14.36	.001	69.41	6.91, *ns*
	Community	36	.68	[.63, .73]	16.34	.001	83.62	
	Work	15	.56	[.48, .66]	10.03	.001	72.97	
Issues	Single	66	.65	[.60, .69]	20.32	.001	80.99	1.08, *ns*
	Multiple	18	.60	[.52, .68]	11.21	.001	78.17	
CQT	PL	60	.65	[.60, .70]	18.45	.001	82.67	1.92, *ns*
	DL	20	.60	[.53, .67]	12.02	.001	76.19	
	Both	5	.60	[.51, .68]	9.98	.001	48.14	
Peer Review	Yes	59	.65	[.60, .70]	13.92	.001	80.47	0.90, *ns*
	No	26	.61	[.53, .6.8]	13.22	.001	79.95	

accurate than the studies that were not peer reviewed, $Q = 72.09$, $p = .001$, although the *R* values were relatively similar. The relatively large value of Q suggests that Peer Review may be a moderator of more applied importance in interpreting research results where more weight should be given to results in peer-reviewed journals.

5.3 | Separate meta-analyses of experimental and field studies

The covariation of Motivation, Setting, and Source resulted in a suggestion that data from field studies of the CQT might produce qualitatively different results from the experimental data. To explore that possibility we conducted two additional meta-analyses of the potential moderators one on the experimental studies and a second on the field studies. Full summary results tables for those two analyses are provided in our online archive as Tables 3 and 4. The meta-analytic effect size estimates for the Field and Experimental Studies were .76 [.71, .81] and .64 [.60, .67] respectively. Across the two meta-analyses only one moderator, Peer Review with the Field Studies, produced a significant effect, $Q = 30.42$, $p < .01$.

To summarize, in our initial moderation variable analyses all 135 r_{dec} effect sizes were used (50 field; 85 experimental). These analyses indicated significant moderation for all categorical variables: Motivation, Source, Issues, CQT Types, and Peer Review (see Table 2). However, separate analyses for the field and experimental

TABLE 4 Results of the meta-analysis of field study moderator variables on the r_{dec}

Moderator	Level	*n* of studies	*R*	95% CI	*Z*	*p* <	*I2*	Q
Motivation	RW	50	.76	[.70, .81]	15.56	.001	95.22	
Source	CJ	49	.76	[.70, .81]	15.58	.001	95.36	
Issues	Single	36	.78	[.71, .83]	13.08	.001	96.08	2.03, *ns*
	Multiple	10	.70	[.60, .78]	0.93	.001		
CQT	PL	50	.76	[.71, .81]	15.96	.001	95.27	
Peer Review	Yes	42	.77	[.71, .82]	14.28	.001	95.31	30.42, $p < .001$
	No	7	.70	[.56, .80]	7.29	.001	91.78	

studies yielded different results. First, only the experimental investigations had data in each sub-category of the moderating variables. Exclusion of the elevated summary effect size for field studies rendered all moderating meta-analyses non-significant for the experimental investigations (Table 3). Second, the field studies provided insufficient data for the sub-categories of the moderating variables "Motivation," "Source," and "CQT Type" and moreover the "field" sub-category of the "Motivation" moderator and the field "Setting field" sub-category were the same. Therefore, only the "Issues" and "Peer Review" moderators could be fully evaluated with the latter obtaining a significant value (see Table 4).

5.4 | Other effect-size measures

We computed meta-analyses of the sensitivity and specificity of our data with summary estimates of effect sizes of .879 and .843 respectively. However, the legitimacy of these pooled findings is suspect because the individual studies varied in the criteria used to ascertain a positive result (e.g., many different types of scoring combined with a variety of decision rules) and there were marked differences in the number of participants among the investigations. Under such conditions, (especially the different scoring thresholds), there will likely be a negative relationship between sensitivity and specificity across the studies and the pooled results will not accurately reflect the overall accuracy of the data (as sensitivity increases, specificity decreases). Indeed, the *r* between sensitivity and specificity was −.197, $p = .02$. The further implication of that finding is that meta-analytic summary ROC and AUC analyses cannot be justified (Jones & Athanasiou, 2005). Moreover, applied statisticians over the last 10 years have concluded that AUC analysis is so flawed and potentially misleading that is simply should be abandoned in favor of other analyses (Hand, 2009a, 2009b; Honts & Schweinle, 2009) such as Gain analysis (Elder, 2020).

5.5 | Information gain analysis

5.5.1 | A median sample of studies: Information gain analysis and accuracy

To conduct an Information Gain (IG) analysis we initially examined the sample of r_{dec} results. There was a study at the median r_{dec} value of 0.66. We selected that study, the five studies with the closest r_{dec} values below 0.66 and the five studies with the closest r_{dec} values above 0.66. Those 11 studies contained 998 decisions and made up our Median Sample. The r_{dec} values of the Median Sample ranged from 0.645 to 0.673. IG within the Median Sample was calculated using the software developed by Honts and Schweinle (2009). The IG for the Median Sample and for interpersonal deception detection (Honts & Schweinle, 2009) are illustrated in Figure 1. The curve for deceptive outcomes can be viewed as an indication of the gain in the CQT's sensitivity of detecting deception as compared to predicting the baserate. Similarly, the IG curve for truthful outcomes can be viewed as an indication of the gain in the CQT's specificity by accurately identifying the truthful as compared to predicting the baserate.

Following the methods described by Honts and Schweinle (2009), CQT deceptive outcome IG was found to peak at 0.37 at a base rate of guilt of 32%. Deceptive outcomes provided significantly more IG ($p < .05$, 1 tailed) than interpersonal deception detection decisions made by lay people in the base rate range of Guilt from 1% through 93% inclusive. IG for Truthful CQT outcomes peaked at 0.48 at a base rate of guilt of 78%. Truthful CQT outcomes provided significantly more IG ($p < .05$, 1 tailed) than truthful interpersonal deception detection decisions made by lay people in the base rate range of guilt from 5% through 99% inclusive. IG for lay persons never exceeded IG for the CQT for either type of decision at any base rate of Guilt. A classification table for the Median Sample is provided in our Table S3. There were more correct outcomes with Guilty subjects than there were with Innocent subjects. There were roughly twice as many Inconclusive outcomes with Innocent than with Guilty subjects, 18.3% versus 10%. The differential in Inconclusive outcomes results in an Information Gain with regard to the innocence of the subject. Excluding inconclusive outcomes, truthful decisions in the Median Sample were 78.9% correct and Deceptive decisions were 91.6% correct. Overall CQT decisions in the Median Sample were 86% correct.

In light of the significant moderator effects, we elected to also illustrate the impact of the strongest moderator effect, Motivation. Figure 2 illustrates the IG of the various outcomes for the three levels of Motivation. The IG curves were based on three median samples of the combined frequencies of 11 studies from and around the median *rdec* for each level of Motivation. For the three levels of Motivation (No Explicit, Some, and Real World) IG for Truthful outcomes peaked at .35, .45, and .46, respectively, at base rates of Guilt of 67%, 73%, and 76%, respectively. IG for Deceptive outcomes peaked at .42, .39,

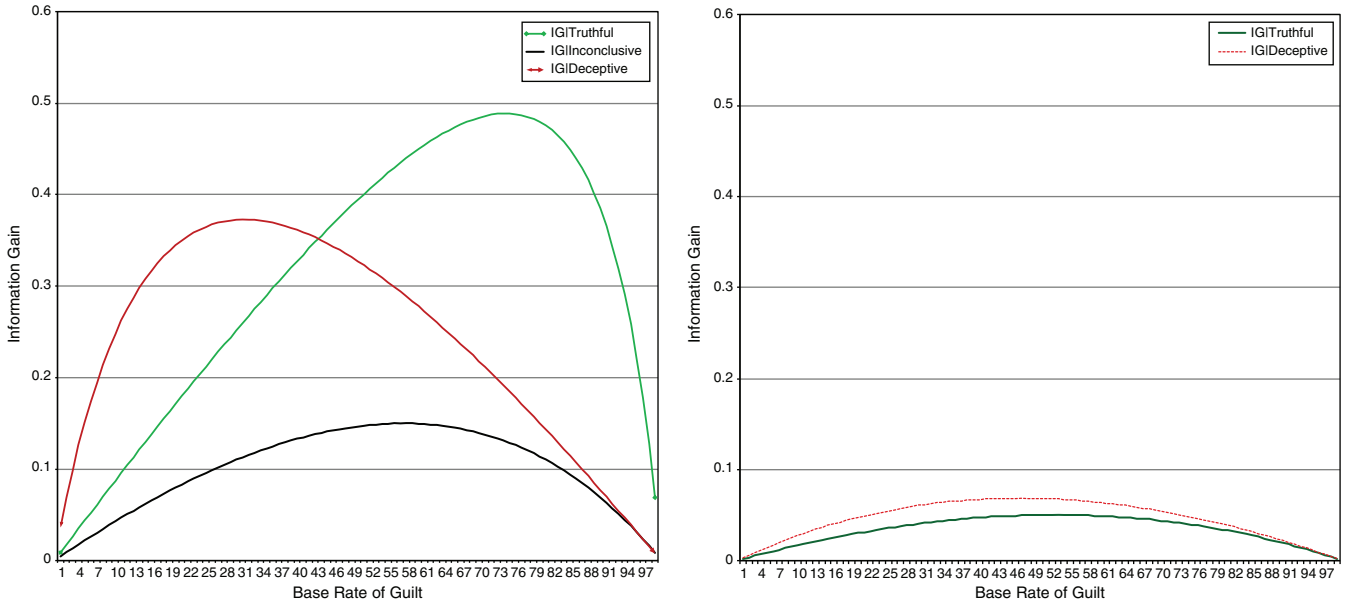

FIGURE 1 Information gain curves for the median sample from this study (left panel) and for interpersonal deception detection after Honts and Schweinle's (2009) Figure 1 (right panel)

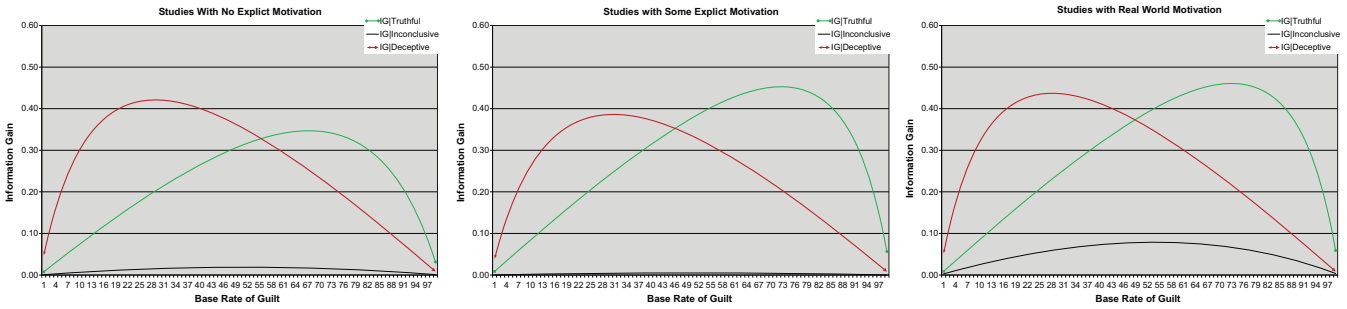

FIGURE 2 Information gain curves for the 11 studies around the median rdec values for each level of the moderator motivation

and .44, respectively, at base rates of Guilt of 29%, 31%, and 28%, respectively. These results show every little impact of the Motivation moderator on the IG provided by Truthful and Deceptive outcomes across the range of motivation. However, in absolute terms the greatest information gain for both Truthful and Deceptive conditions were found under real world motivational conditions. It is also interesting that in the Real-World category the peak IG for Truthful and Deceptive outcomes are approximately equal.

However, Inconclusive outcomes present a different pattern. In the No Explict and Some motivational categories, Inconclusive outcomes provide almost no IG. However, under in the Real-World category, Inconclusive outcomes provide IG indicative of Innocence because the frequency of Inconclusive outcomes was higher with Innocent than with Guilty subjects. A classification table for the 11 studies around the Real-World motivation Median *rdec* is provided as Table S4. However, the IG for Inconclusive outcomes in the Real-World Motivation condition were not significantly better ($p < .05$, 1-tailed) than interpersonal deception detection outcomes of truthful from Honts and Schweinle (2009).

6 | DISCUSSION

Following the approach reported by Hartwig and Bond (2014) we explored the potential validity of a number of moderator variables that critics of the CQT have either hypothesized or simply asserted were powerful determinants of the validity of the CQT. The nature of the criticisms about interpersonal deception detection research has some similarity to the criticisms of the research on the CQT. We sampled the CQT research broadly to maximize the scope of our examination of external validity through the moderator variables and to avoid any criticism that we were biased in our selection of studies.

As described above, the critics of the CQT research have generally dismissed experimental research as lacking external validity. Generally, that criticism has stated that the motivational setting in experiments is qualitatively different from the motivational setting in real world settings where jobs, money, freedom and even life are at stake. Iacono and Ben-Shakhar (2019) noted the longevity of the skepticism in the external validity of CQT experiments, "Lykken (1978)

argued 40 years ago that there is no reason for laboratory subjects to find the experience frightening or guilt-provoking, with the circumstances more akin to a challenging game in which relevant questions are more likely to elicit orienting responses than reactions associated with genuine fear or guilt" (p. 93). Implicit in the critics' argument about the importance of motivation is an assumption that fear and guilt are necessary components of real-world polygraphs but are not present in experimental studies. Iacono and Ben-Shakhar (2019) also note that this assertion and its necessarily implicit assumption, have never been tested.

Our analyses provided a partial test of the potential impact of motivation on CQT outcomes and found a significant moderator effect of Motivation. However, in stark contrast to the assertions of the critics we found that the effect of motivation was linear and it was not dramatic. In this regard our results replicate and extend the meta-analytic results reported by Kircher et al. (1988). Our results, based upon a large sample of highly varied CQT studies, strongly suggest that the long-standing assertion of dramatic qualitative differences between CQT experiments and field studies based upon the moderation of Motivation is without support. Our results indicated that with the CQT liars and truth tellers produce similar results in experimental and in field settings that differ quantitatively, but not qualitatively. Significant and powerful discrimination is seen even in experimental settings that provided no explicit contingency associated with test outcome. However, our results do show that experiments without any explicit motivation underestimate the discriminative power of the CQT. The implication of our results with motivation suggest that researchers who conduct CQT experiments should build in an explicit reward/punishment contingency, as experiments with such a contingency produce estimates that are closer to the effects sizes found in field settings although they appear to somewhat underestimate effect sizes in the field.

Our findings concerning the general potential moderator variables, Motivation, Subject Source, and Experiment versus Field, did not replicate the results of Hartwig and Bond (2014). They found weak detection effects and no significant moderator effects while we significant effects for peer review status and for motivation/experiment versus field. Given those differences it seems possible that the expressive phenomena in interpersonal deception detection and those in psychophysiological deception detection may represent different processes. Research is needed to explore those differences. Nevertheless, the moderator effects we found although significant were not substantial. The results of our analyses provide little or no support to the long-standing claims, sometimes stated as facts, by the critics that experiments are not useful for estimating field accuracy of the CQT. The available data suggest that psychophysiological deception detection works the much the same way in an experiment as it does in applied settings in the field.

Finally, we examined a potential moderator that was of specific interest to CQT practitioners, Comparison Question Type. We did find a statistically significant difference for CQT Type as a moderator in the initial analysis that difference was not found in the separate analyses of the Experimental data. The range of the Comparison Question Type variable was not represented in the Field data and thus could not be tested. While this remains a bit of an open question for field application, it is notable that direct comparisons between the two techniques have generally failed to find significant differences (e.g., Honts & Reavy, 2015).

To provide useful practical information to end users of CQT outcomes we conducted Information Gain analyses. Within a median sample of 11 studies both Truthful and Deceptive CQT outcomes provided significantly more IG than interpersonal deception detection between base rates of guilt between 4% and 94%. Similar but stronger results were found for a median sample from the Real-World motivation studies.

The information gain for both the entire sample and for the Real-World motivation studies show separate peaks for information gain for truthful and deceptive outcomes at different ends of the base rate continuum. This is not surprising, but it has different implications for end users. In cases where the base rate of guilt is relatively high, for example in the population of persons formally charged with a crime, truthful outcomes should be given more weight than deceptive outcomes by the end user of the information. The United States Bureau of Justice Statistics (2019) reports that the base rate of guilt (convictions + guilty pleas) in the state courts was about 66% in 2006. Within our Median Sample at a base rate of guilt at 66% the IG for a deceptive CQT outcomes was .24 and the IG for truthful CQT outcomes was .47. This indicates that at the base rate of guilt for charged subjects in state courts a truthful outcome is about twice as informative as a deceptive CQT outcome. In the median sample from the Real-World motivation studies with a base rate of guilt of 66% the IG for deceptive outcomes was .27 and for Truthful outcomes as .45, essentially the same IG values as those that are representative of the entire sample of studies. At the same 66% base rate of guilt, layperson's decisions that a person is truthful have an IG of .05 and deceptive decision have and IG of .06. Thus, at the critical base rate for persons charged with a crime, a truthful CQT outcome is approximately 9 times more informative than a layperson conclusion that a person is truthful. At that critical base rate, a CQT deceptive outcome is approximately 4 times more informative than a layperson conclusion that the person is a liar.

At the other end of the base rate continuum the information gain situation is different. Consider a case where there are three suspects but only one person could have committed the crime, so the base rate of guilt is .33. Within our Median Sample and Median Sample for Real World Motivation a CQT truthful outcome had an IG of .28 and .27, respectively at a base rate of guilt of 33%. Deceptive outcomes had an IG of .37 and .43, respectively. Again, little difference is seen in IG between the entire sample and those data from Real World motivation studies. At the same base rate of guilt, a layperson's decision that a person is a truth teller has an IG of 0.04 and a layperson's decision that a person is a liar has an IG of 0.06. Thus, a CQT truthful outcome is approximately 7 times more informative than a layperson's decision of truth teller and a CQT deceptive outcome is approximately 6 times more informative that a layperson's decision that a person is a liar.

6.1 | Caveats, cautions, and areas of concern

Our results show that, when estimated across the research literature, the CQT discriminates truth tellers from liars with a large magnitude of effect, r_{dec} = .694. Given our inclusion of a number of studies previous reviews have found to be substandard, our effect size estimate should be viewed as conservative. Although our effect size estimate was moderated by several variables, the moderator effects were small and had little impact on the IG provided by CQT test results. Moreover, all but two of those moderator effects went to non-significance when we conducted separate meta-analyses of the Experimental and Field data. However, our results should not be interpreted as indicating that all CQT polygraph tests have high accuracy. There was a large range of results that does not appear to be due to the tested moderators and thus there are likely other factors at work.

6.1.1 | Unanswered questions about the CQT field studies

Although the 50 field studies examined in this meta-analysis varied in many ways, the methodology was surprisingly and disturbingly invariant. In an effort to explore the reasons for the high variability in the field studies we attempted to examine a number of possible moderators. However, for most of those potential moderators there was either no information or there was minimal variation. In short, the field studies of the CQT can be generally characterized as quasi-experiments with non-equivalent groups where subject condition is determined retrospectively from a confession given either by a participant or by someone else involved in the investigation. Almost all of the data were generated in forensic settings by law enforcement polygraph examiners conducting investigations. The crimes being investigated with the polygraphs generally were not specified and presumably ran the full range of criminal activity. Not surprisingly many of the researchers involved in conducting field research on the CQT are involved professionally in polygraph testing, either in conducting CQT tests, conducting funded research, or appearing as experts in courts of law supporting or opposed to CQT testing. However, such connections are often not made explicit and for the older literature they are impossible to code.

The homogeneity of methods in the field research of the CQT is clearly a weakness. The one clear exception to the general homogeneity of the credibility criterion are the two field studies that used a paired testing protocol to determine a criterion of guilt and innocence (Ginton, 2013; Mao et al., 2014). Since the Ginton (2013) approach determines the guilt criterion by algorithm it seems reasonable to expect that if the field literature were highly biased in favor of CQT accuracy then the paired test algorithm approach should show reduced accuracy. However, that was not the case in these data. The mean r_{dec} value for field studies was .71 while the r_{dec} value for Mao et al. (2014) was .72 and for Ginton (2013) was .80. This is not to say the Ginton (2013) approach is the solution to the criterion problem as there is some disagreement about it as well (e.g., Ginton, 2020; Iacono & Ben-Shakhar, 2019). We are simply noting that this one clear contrast in methods fails to provide any support to the assertions of the CQT critics. Clearly there are weaknesses in the field data for the CQT and additional research that takes different and innovate approaches are needed to supplement the current literature.

6.1.2 | Assessing credibility versus interrogation ploy

In application in police and national security we see the polygraph being used in two ways. Some agencies use the polygraph as a credibility assessment test with the intent to use the outcome for its own value in focusing investigations, providing evidence, or in using the information for other decisions. Many of the studies included in our meta-analysis appear to fit that increased information model.

However, there is a second use in the field where examiners and/or their agencies use the polygraph as an evidence ploy to further an interrogation with the goal of obtaining a confession. Honts (2017) described the policies of the FBI polygraph program as they were revealed in a criminal case (U.S. vs. Jamico Tennison, 2016). In that case testimony was given by an FBI Supervisory Special Agent (SSA) who was also an instructor at the U.S. Government's only polygraph training facility the National Center for Credibility Assessment (NCCA). The SSA's testimony was clear that as policy the FBI put high emphasis on minimizing false negative errors with almost no concern for false positive errors. Moreover, FBI had chosen a specific combination CQT variant, scoring system and decision rule to reach that goal that included treating inconclusive outcomes as indications of deception. Honts (2017) provided an analysis based upon U.S. Government generated polygraph data that indicated that under the FBI's policy only 17% of the actually innocent people given FBI polygraph examinations will avoid interrogation and, thus 83% of the actually innocent are needlessly subjected to a risk of making a false confession.

Concerns about the impact of misused or misinterpreted polygraph examinations are well documented in the false confession literature (e.g., see Kassin, Drizin, et al., 2010). That concern is amplified by the fact that people, including police (Honts, Kassin, et al., 2014; Kassin, Meissner, et al., 2005) and polygraph examiners (Honts, Forrest, et al., 2019) are unable to discriminate true from false confessions. The fact that four out of five actually innocent subjects tested by the FBI will be interrogated strongly suggests that under those conditions polygraph tests may be an important factor leading to false confessions. However, also see Bonpasse (2013) who documented the sometimes role (14.4%) of polygraph tests in wrongful convictions but also documented that for the majority of the wrongly convicted where there were polygraph tests conducted before conviction, 62.9% of those tests supported the defendant's innocence but did not prevent a wrongful conviction.

6.1.3 | Weak standards for training and practice

Although polygraph tests are unquestionably psychological tests, Psychology as a profession has never claimed them as falling under the

domain of psychological test regulation. In the United States this fact has left the setting of standards and the regulation of training and practice for polygraph examiners to the various states and to professional organizations. Currently only 26 of the 50 United States license polygraph examiners (APA, 2019a, 2019b). Licensing requirements vary dramatically from state to state. Ethical standards are provided by the various professional groups (e. g., APA, 2015). However, those standards are rarely and inconsistently enforced and have no force over non-members. Similarly, the professional organizations have provided standards of practice (e. g., APA, 2018). However, those standards appear to be advisory and not binding. The APA accredits polygraph schools (APA, 2019a, 2019b), but there are a number of active polygraph examiner schools that do not have accreditation and a substantial number of practicing examiners have not graduated from an APA accredited school (for an example and additional information see, Honts & Handler, 2013). Moreover, even under the best of situations the requirements to conduct polygraph tests are far below those required to administer and interpret even the simplest of psychological tests. In the United States this lack of unified regulation, standards, practices and ethics has created a situation where end users are left to make decisions about the quality of the polygraph practice that generated the test results presented to them. Unfortunately, that seems to be an assessment that they are often poorly prepared to make.

6.1.4 | Countermeasures

Countermeasures are anything that the subject of a test might do in order to distort or change the outcome of that test. Polygraph tests in general and specifically the CQT, were shown to be vulnerable to countermeasures (see the review by Honts, 2014) in experiments. However, the frequency and effectiveness of countermeasures in field practice remains anecdotal. This is an area where additional research is needed. However, it is critical to note that this vulnerability to, and concern about, countermeasures is common to all tests where the subject of the test has something to gain or lose from the outcome of the test. The CQT is not at all unique in this regard and the existence of countermeasures should no more eliminate the CQT from applied use than it would any IQ test, personality test or other psychological assessment.

6.2 | Possible remedies

Unified, universal, and binding regulation defining standards for training, practices, and ethics along with universal licensing of polygraph examiners would be highly desirable. However, in lieu of Psychology as a profession owning the fact that polygraph tests are psychological tests and that polygraph tests should be regulated as such, this seems highly unlikely. In the meantime, transparency would seem to be the most readily achievable remedy. The 44-year long experience of the State of New Mexico with admitting the results of polygraph tests in courts of law may provide some guidance. The central part of the apparent success of New Mexico admitting polygraph test results as evidence seems to be their Rule of Evidence 11–707 (N.M. R. Evid. 11–707, 2015). Table S5 in our Supplementary Archive B describes the requirements for a polygraph test result to be admissible in the New Mexico courts. Rule 11–707 requires documentation through the provision of all of the polygraph test data and a recording of the examination to any opposing party. Those materials must be provided at least 30 days in advance of any legal proceeding and all polygraph tests taken by the examinee must be revealed. This transparency appears to have worked well in New Mexico for 44 years, and the requirements of Rule 11–707 would seem to be a good place for the polygraph profession to start in order to provide transparency for all polygraph testing.

6.3 | Theory and the CQT

Finally, we would like to address the long-standing criticism, most recently restated by Iacono and Ben-Shakhar (2019), that the CQT should not be used because there is a lack of a comprehensive theory to explain how the CQT works. We find the Iacono and Ben-Shakhar (2019) critique lacking on two grounds. First, the Iacono and Ben-Shakhar (2019) argument is an illogical straw man argument. There is no requirement that a full explanatory theory be in place before using a technology (Honts & Reavy, 2015). Honts and Reavy (2015) specifically detail the fact that aspirin, in clinical use since the late 1800s, still lacks a complete theoretical explanation of its medical action. Despite this lack of a complete theory the worldwide medical consumption of aspirin in 1998 exceeded 40,000 metric tons a year (Warner & Mitchell, 2002). The results of the present analyses clearly show that the CQT does work, albeit not perfectly. Moreover, the CQT works much better at assessing credibility than unassisted humans doing interpersonal deception detection. Despite this finding Iacono and Ben-Shakhar (2019) would have law enforcement around the world abandon the CQT in favor of near chance interpersonal deception detection. We find that position indefensible.

The second striking weakness of the Iacono and Ben-Shakhar (2019) lack of CQT theory argument is simply that their argument is disingenuous on its face. There are relatively recent theoretical offerings that are consistent with the existing research literature. Ginton (2009) proposed a cognitive theory that focuses on attention. Senter et al. (2010) offered another cognitive theory based upon question salience. Honts (2014) proposed a theory of the CQT that adapted the Cognitive Load (Demand) theory proposed by Vrij and his colleagues (Vrij, 2008; Vrij, Fisher, et al., 2006) as a theoretical framework for understanding interpersonal deception.

Many field practitioners state a belief that fear of detection is the underlying mechanism of the CQT. Iacono and Ben-Shakhar (2019) echo similar beliefs in their instance that the field and laboratory are qualitatively different due the emotional content of the field settings. However, the results of our meta-analysis have clearly falsified both of those positions by showing that the CQT provides a high-level of

discrimination in both the experimental and in field settings including experiments with no explicit contingency associated with test outcome. If, as the results of our meta-analysis show, the results are similar in no incentive laboratory studies and in studies where people are facing the loss of wealth, freedom and/or even life, then clearly neither fear, nor for that matter any emotion, are the sine qua non for the CQT to work, or to be scientifically studied.

However, the significant linear effect of motivation on the degree of discrimination is easily accounted for within any of the cognitive theories cited above. The increase in motivation simply helps to define the focus of the subject on the test questions critical to them, that is, the comparison questions for the actually innocent and the relevant questions for the actually guilty. However, what is lacking in the current CQT research literature are studies deliberately designed to test predictions that follow from these cognitive theories and research oriented toward the construct validation of the proposed cognitive mechanisms. Studies similar to Vrij, Mann, et al. (2008) where the effects of manipulating cognitive load on the ability to do interpersonal detection deception should be relatively easy to do with the CQT if scientists and funding agencies are willing to take on the work to directly advance our theoretical understanding in this domain.

6.4 | Concluding comments

The modern academic disagreement over the CQT has now lasted over five decades and several generations of scientists. We have no illusions that this meta-analysis will resolve this conflict. We ask only that undecided readers view the data with an open mind and consider the following two points. First, we would note that the arguments against the validity of the CQT are now almost completely lacking in data. Although there are some experiments and field studies with very low accuracy, those studies are shown by our analyses to be outliers and not representative of the central tendency of the research literature. Moreover, against the assumptions and predictions of the CQT critics we found that strong outcome contingent motivation was not required for the CQT work and that the relationship between CQT accuracy and motivation was positive, continuous, and linear.

Second, in the absence of data, Iacono and Ben-Shakhar (2019) tout a thought experiment that they say shows that it is possible for a chance technique to produce high accuracy in a field study. However, there are no data that support that thought experiment and it, like all thought experiments, is a pure invention of the mind. Moreover, their invention was based upon so many untenable assumptions that it can easily be seen as having been most likely derived as a backtrack. That is, it seems likely that the thought experiment started with their desired conclusion and they worked backward for the unique preconditions that would produce that desired conclusion (Honts & Thurber, 2019a, 2019b). Moreover, here we reported on 42 field studies published in peer reviewed journals. In contrast to the thought experiment there is not a single study where the number of false positive outcomes equals or exceeds the number of true positive outcomes. The critics of the CQT would have you believe that all 42 of those peer-reviewed studies are invalid artifacts and that the peer reviews for those journals and/or the editors of those journals are either incompetent or dishonest. We ask our readers to consider which of the following propositions is the more logical, parsimonious, and likely? First, the accuracy of the CQT is no better than chance in the real world and the CQT has a discontinuous non-linear relationship with motivation that is invisible in the peer-reviewed literature because all of the real-world studies are inaccurate and were published only by dishonesty and incompetence on the part of the scientific journals involved. Alternatively, the CQT is an imperfect tool that makes some errors, the CQT has a positive linear continuous relationship with motivation, and the CQT is accurate enough to provide substantial information gain to decision makers who are only able to detect deception with 54% accuracy.

CONFLICT OF INTEREST

The first author is licensed as a polygraph examiner and conducts forensic polygraph examinations. He also serves as a consulting and testifying expert witness concerning the quality of polygraph examinations and about the use of polygraph examinations as contributors to false confessions. The second author has no known conflicts of interest to disclose. The third author is licensed as a polygraph examiner and conducts forensic polygraph examinations. He also serves as a consulting and testifying expert witness concerning the quality of polygraph examinations. The third author is the editor of the journal, *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice*.

ACKNOWLEDGEMENT

This work was supported by Department of Psychological Science, Boise State University.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are available in the Supplementary Materials for this report.

ORCID

Charles R. Honts https://orcid.org/0000-0002-6925-731X

ENDNOTE

[1] The critics of the CQT have also often raised criticism of the venues where the research was published. In particular, the journal *Polygraph*, now known as *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice (PFCA)* was dismissed as not a valid scientific venue by Iacono and Ben-Shakhar (2019), "*Polygraph* is not currently edited by a scientist, nor has it been in the past; it is not a peer-reviewed scientific journal." (p. 89). Most of Iacono and Ben-Shakhar's (2019) assertions about *PFCA* are simply false. Scientific articles submitted to *Polygraph*/*PFCA* have been peer reviewed at least since 1983, as the current first author has personal knowledge that Honts and Hodes (1983) was peer reviewed and revisions were requested prior to publication. All articles published in *Polygraph*|*PFCA* have been peer reviewed since the early 2000's. Since 2002, *Polygraph*|*PFCA* has been indexed by *Criminal Justice Abstracts* and *Criminal Justice Abstracts With Full Text* (EBSCO, 2019). While it is true that the current Editor of *PFCA* does not

have academic credentials, he has coauthored a number of published peer-reviewed papers and is a coauthor of this manuscript. Moreover, Iacono and Ben-Shakhar fail to mention that persons with academic credentials and academic appointments have consistently been associate editors of *Polygraph|PFCA*. Currently nine of the associate editors have academic credentials. The direct involvement of academics on the editorial board of *Polygraph|PFCA* has been true since at least 1988. Notwithstanding Iacono and Ben-Shakhar's misrepresentation of the status of *Polygraph|PFCA*, we correctly coded it as a peer-reviewed journal.

REFERENCES

American Polygraph Association. (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, *40*(4), 194–305. https://doi.org/10.1016/b978-0-12-802924-4.09986-2.

American Polygraph Association. (2015). *Code of ethics: American polygraph association*. American Polygraph Association.

American Polygraph Association. (2018). *APA standards of practice*. American Polygraph Association.

American Polygraph Association. (2019a). Find a member. Retrieved from https://apoa.memberclicks.net/find-a-member#/

American Polygraph Association. (2019b). State licensing boards and associations. Retrieved from https://apoa.memberclicks.net/state-licensing-boards-associations

Bermudez, M. N., & Arias, S. W. (2010). Polygraph testing in Colombia. *Polygraph*, *40*(2), 124–130.

Bermudez M. N., & Arias S. W. (2011). Polygraph testing in Colombia. *Polygraph*. *40*(2), 124–130.

Bonpasse, M. (2013). Polygraph and 215 wrongful conviction exonerations. *Polygraph*, *42*(2), 112–127.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis, version 3*. Biostat.

Borenstein, M., Hedges, L., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.15.

Craig, R. A., Raskin, D. C., & Kircher, J. C. (2011). The use of physiological measures to detect deception in juveniles. *Polygraph*, *40*(2), 86–99.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979–995. https://doi.org/10.1037/0022-3514.70.5.979.

Daubert, V. Merrell Dow Pharmaceuticals, 509 U.S. 579, (1993). https://supreme.justia.com/cases/federal/us/509/579/.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x.

EBSCO. (2019). Criminal Justice Abstracts | EBSCO cjacoverage.xls and i3hcoverage.xls. Retrieved from https://www.ebsco.com/products/research-databases/criminal-justice-abstracts

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, *77*(5), 757–767. https://doi.org/10.1037/0021-9010.77.5.757.

Elder, J. (2020). AUC: A fatally flawed model metric. Statistic.com: Blog Retrieved from https://www.statistics.com/auc-a-fatally-flawed-model-metric/?inf_contact_key=ab87743d7421f0c03c18f4eaa836f1aa4dfbc39d7283b2cb89d5189540b69330

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

Ginton, A. (2013). A non-standard method for estimating accuracy of lie detection techniques demonstrated on a self-validating set of field polygraph examinations. *Psychology, Crime & Law*, *19*(7), 577–594. https://doi.org/10.1080/1068316X.2012.656118.

Ginton, A. (2020). A critical examination of Iacono and Ben-Shakhar's critique of Ginton's innovative technique for estimating polygraph CQT accuracy in real-life cases. *Journal of Investigative Psychology and Offender Profiling*, *17*(3), 296–309. https://doi.org/10.1002/jip.1558.

Granhag, P. A., & Strömwall, L. A. (2004). *The detection of deception in forensic contexts*. Cambridge University Press.

Grubin, D., Kamenskov, M., Dwyer, R. G., & Stephenson, T. (2019). Post-conviction testing of sex offenders. *International Review of Psychiatry*, *31*(2), 141–118. https://doi.org/10.1080/09540261.2018.1561428.

Guodong, D. (2020) Is a polygraph test admissible as evidence in China? China Justice Observer. Retrieved from https://www.chinajusticeobserver.com/a/is-a-polygraph-test-admissible-as-evidence-in-china

Hand, D. J. (2009a). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5.

Hand, D. J. (2009b). Mismatched models, wrong results, and dreadful decisions: On choosing appropriate data mining tools. Paper presented at: KDD'09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1–2.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589.

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/acp.3052.

Hira, S., & Furumitsu, I. (2002). Polygrapic examinations in Japan: Application of the guilty knowledge test in forensic investigations. *International Journal of Police Science & Management*, *4*(1), 16–27. https://doi.org/10.1177/146135570200400103.

Honts, C. R. (2004). The psychophysiological detection of deception. In P. Granhag & L. Strömwall (Eds.), *Detection of deception in forensic contexts* (pp. 103–123). Cambridge University Press.

Honts, C. R. (2014). Countermeasures and credibility assessment. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 131–158). Academic Press.

Honts, C. R. (2017). Current FBI polygraph practices put the innocent at high risk of wrongful accusation, interrogation, and false confession [Paper presentation]. American Psychology - Law Society meeting, Seattle, WA.

Honts, C. R., Amato, S., & Gordon, A. (2000). Validity of outside-issue questions in the control question test: Final report on grant no. N00014-98-1-0725. Submitted to the Office of Naval Research and the Department of Defense Polygraph Institute. Applied Cognition Research Institute, Boise State University. DTIC Accession# ADA376666. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA376666.pdf

Honts, C. R., Forrest, K., & Stephanescu, A. (2019). Polygraph examiners unable to discriminate true and false juvenile confessions. *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice*, *48*(1), 1–9.

Honts, C. R., & Handler, M. (2013). A case study of the validity of the Arther examination procedures in a criminal case with DNA confirmation. *Polygraph*, *42*(2), 61–71.

Honts, C. R., & Hodes, R. L. (1983). The detection of physical countermeasures. *Polygraph*, *12*, 7–17. https://doi.org/10.1037/0021-9010.70.1.177.

Honts, C. R., Kassin, S. M., & Craig, R. (2014). "I'd know a false confession if I saw one": A constructive replication with juveniles. *Psychology, Crime & Law*, *20*(7), 695–704. https://doi.org/10.1080/1068316X.2013.854792.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, *79*(2), 252–259. https://doi.org/10.1037/0021-9010.79.2.252.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2008). Scientific status: The case for polygraph tests. In D. L. Faigman, M. J. Saks, J. Sanders, & E. Cheng (Eds.), *Modern scientific evidence: The law and science of expert testimony (volume 5): 2008–2009 edition*. Thompson West.

Honts, C. R., & Reavy, R. (2015). The comparison question polygraph test: A contrast of methods and scoring. *Physiology and Behavior*, *143*, 15–26. https://doi.org/10.1016/j.physbeh.2015.02.028.

Honts, C. R., & Schweinle, W. (2009). Information gain of psychophysiological detection of deception in forensic and screening settings. *Applied Psychophysiology and Biofeedback*, *34*(3), 161–172. https://doi.org/10.1007/s10484-009-9096-z.

Honts, C. R., & Thurber, S. (2019a). A comprehensive meta-analysis of the comparison question polygraph test [Paper presentation]. Annual Meeting of the American Psychology Law Society, Portland, Oregon.

Honts, C. R., & Thurber, S. (2019b). Analyzing Iacono's thought experiment about polygraph field studies: Reason or fantasy? *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice*, *48*, 76–86. https://doi.org/10.13140/RG.2.2.21263.33448.

Iacono, W. G., & Ben-Shakhar, G. (2019). Current status of forensic lie detection with the comparison question test: An update of the 2003 National Academy of Sciences report on polygraph testing. *Law and Human Behavior*, *43*(1), 86–98. https://doi.org/10.1037/lhb0000307.

Iacono, W. G., & Lykken, D. T. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *Science in the law: Social and behavioral sciences issue, American casebook series* (pp. 582–618). West Group.

IBM. (2017). IBM SPSS Statistics, V. 25.

Jones, C. M., & Athanasiou, T. (2005). Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. *The Annals of Thoracic Surgery*, *79*(1), 16–20. https://doi.org/10.1016/j.athoracsur.2004.09.040.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior*, *34*(1), 3–38. https://doi.org/10.1007/s10979-009-9188-6.

Kassin, S. M., Meissner, C. A., & Norwick, R. J. (2005). "I'd know a false confession if I saw one": A comparative study of college students and police investigators. *Law and Human Behavior*, *29*(2), 211–227. https://doi.org/10.1007/s10979-005-2416-9.

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, *12*(1), 79–90. https://doi.org/10.1007/bf01064275.

Kraujalis, L., Kovalenko, A., & Saldziunas, V. (2007). Legal and practical aspects of using the polygraph in the Republic of Lithuania. *European Polygraph*, *1*(1), 17–23.

Lee, V. Martinez, 2004-NMSC-027, 136 N.M. 166, 96 P.3d 291 (2004). https://law.justia.com/cases/new-mexico/supreme-court/2004/8c0a.html.

Lykken, D. T. (1978). The psychopath and the lie detector. *Psychophysiology*, *15*(2), 137–142. https://doi.org/10.1111/j.1469-8986.1978.tb01349.x.

Mao, Y., Liang, Y., & Hu, Z. (2014). Accuracy rate of lie-detection in China: Estimate the validity of CQT on field cases. *Physiology and Behavior*, *140*, 104–110. https://doi.org/10.1016/j.physbeh.2014.11.063.

Matsuda, I., Ogawa, T., & Tsuneoka, M. (2019). Broadening the use of the concealed information test in the field. *Frontiers in Psychiatry*, *10*, 24. https://doi.org/10.3389/fpsyt.2019.00024.

Meijer, E. H., & von Koppen, P. J. (2008). Lie detectors and the law: The use of polygraph in Europe. In D. Canter & R. Zukauskiene (Eds.), *Psychology and the law: Bridging the gap* (pp. 31–50). Taylor and Francis.

Munsterberg, H. (1908). The traces of emotion. In *On the witness stand: Essays on psychology and crime* (pp. 111–134). The McClure Company.

National Research Council of the National Academy of Sciences. (2003). *The polygraph and lie detection*. The National Academies Press.

New Mexico Rule of Evidence 11–707. (2015). N.M. R. Evid. 11–707: Rule 11–707 - Polygraph Examinations. As amended by Supreme Court Order No. 15–8300-012, effective for all cases filed or pending on or after December 31, 2015.

Osumi, M. (2019). Japan's crime rate his postwar low, but child abuse, domestic violence and offenses by elderly on rise. The Japan Times. Retrieved from https://www.japantimes.co.jp/news/2019/11/29/national/crime-legal/japans-crime-rate-hits-postwar-low-report-shows-rise-child-abuse-domestic-violence-offenses-elderly/#.X0VgLS3MxTY

Philippe, R des B, (2020). Loi modifiant le Code d'instruction criminelle en ce qui conceme l'utilisation du polygraphe (1), Belgish Staatsblad, February 21, 2020, Montieur Belge, p. 10239–10240.

Podlesny, J. A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations? *Crime Laboratory Digest*, *20*(3), 57–61.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional and legal issues surrounding application and acceptance of polygraph evidence. *Utah Law Review*. *1986*(1):29–74.

Raskin, D. C., & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology*, *15*(2), 126–136. https://doi.org/10.1111/j.1469-8986.1978.tb01348.x.

Raskin, D. C., & Honts, C. R. (2002). The comparison question test. In M. Kleiner (Ed.), *The handbook of polygraph testing* (pp. 1–48). Academic.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (pp. 565–582). West Group.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (2014). *Credibility assessment: Scientific research and applications*. Academic Press.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology*, *51*(1), 4–13. https://doi.org/10.1037/0022-006X.51.1.4.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, *57*(3), 221–237. https://doi.org/10.1037/h0087427.

Salgado, J. F. (2018). Transforming the area under the normal curve (AUC) into Cohen's d, Pearson's rpb, odds-ratio, and natural log odds-ratio: Two conversion tables. *The European Journal of Psychology Applied to Legal Contexts*, *10*(1), 35–47. https://doi.org/10.5093/ejpalc2018a5.

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.

Shadish, W. R., & Sweeney, R. B. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology*, *59*(6), 883–893. https://doi.org/10.1037/0022-006x.59.6.883.

Trovillo, P. V. (1939a). A history of lie detection. *The Journal of Criminal Law and Criminology*, *29*, 848–881.

Trovillo, P. V. (1939b). A history of lie detection (concluded). *The Journal of Criminal Law and Criminology*, *30*, 104–119. https://doi.org/10.2307/1136392.

U.S. v. Scheffer. 523 U.S. 303 (1998).

U.S. vs. Jamaico Tennison. (2016) No. 15-cr-00212 MCA, U.S. District Court for the District of New Mexico, 3-day suppression hearing December 2015–February 2016. Transcripts of the suppression hearing and the Judge's order ultimately suppressing the post-polygraph confession can be downloaded from: https://www.dropbox.com/sh/7lqrvp7u80ka4vs/AADeshbM0SALgrX4mp4gwRnKa?dl=0

United States Bureau of Justice Statistics. (2019). FAQ Detail: What is the probability of conviction for felony defendants? Retrieved from https://www.bjs.gov/index.cfm?ty=qa&iid=403

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Wiley.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Science*, *10*(4), 141–142. https://doi.org/10.1016/j.tics.2006.02.003.

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, B., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*(3), 253–265. https://doi.org/10.1007/s10979-007-9103-y.

Warner, T. D., & Mitchell, J. A. (2002). Cyclooxygenase-3 (COX-3): Filling in the gaps toward a COX continuum? *PNAS*, *99*(21), 13371–13373. https://doi.org/10.1073/pnas.222543099.

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776.

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*(3), 155–167. https://doi.org/10.1037/1076-898x.8.3.155.

What Works Clearinghouse. (2008). Evidence standard for reviewing studies: Version 1.0. Retrieved from https://files.eric.ed.gov/fulltext/ED511668.pdf

Widacki, J. (2007). Polygraph examinations in Poland. *European Polygraph*, *1*(1), 24–34.

Zhang, X. (2011). The evolution of polygraph testing in the People's Republic of China. *Polygraph*, *40*(3), 181–193.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Honts CR, Thurber S, Handler M. A comprehensive meta-analysis of the comparison question polygraph test. *Appl Cognit Psychol*. 2021;35:411–427. https://doi.org/10.1002/acp.3779