Scoring Respiration When Using Directed Lie Comparison Questions

Charles R. Honts and Mark Handler

Abstract

Recent research reports that respiration responses to directed lie comparison questions may not result in the expected shorter line excursion with innocent participants compared with responses to target stimuli. However, the implications for those physiological findings for numerical scoring are unexplored. We examined the impact of the use of the directed lie on numerical scores with new analyses of two existing data sets. We examined a set of 25 confirmed field cases from Honts and Raskin (1988) where directed lie and probable lie comparison questions were contrasted within subjects. We then examined data from 250 participants in an experiment (Honts & Reavy, 2009) that explores differences between examinations with directed or probable lie comparison questions. Our analyses failed to reveal any significant effects of directed lies on either Utah or Objective Scoring System, Version 2 numerical scores. Results showed that numerical scores differ significantly for guilty and innocent examinees using both probable and directed lie comparison questions. Results indicate the potential that examiners who use directed lie comparison questions may simply score them using certain standard numerical criteria. Continued research and interest in the directed lie comparison question is recommended.

Keywords: directed lie, probable lie, respiration, numerical scoring

One of the diagnostically useful physiological measures collected during polygraph testing is respiration (Kircher & Raskin, 1988; Nelson, Krapohl & Handler, 2008). The recorded respiratory waveform reflects chest and abdominal movement associated with breathing. In numerical scoring, examiners assign scores by making qualitative inferences of changes through pattern recognition, quantitatively by making reference to computer measurements displayed in their software, or through some combination of qualitative and quantitative indices.

All test data analysis models validated for diagnostic testing (American Polygraph Association, 2011) include some features from which examiners estimate changes in respiration (e.g. Bell, Raskin, Honts, & Kircher, 1999; Nelson, Handler, Shaw, Gougler, Blalock, Russell, Cushman & Oelrich, 2011; Swinford, 1999). Validated computer scoring models use computer measurements to calculate respiratory line length or respiratory excursion (Kircher & Raskin, 1988; Nelson, Krapohl & Handler, 2007).

In comparison question polygraph tests (CQT), comparison question responses and relevant question responses are compared for differential reactivity (Bell, Raskin, Honts & Kircher, 1999; Handler & Nelson, 2007; Krapohl, 2001; Senter, Weatherman, Krapohl, & Horvath, 2010). Presumably the degree of difference between the critical questions reflects underlying mental process reflecting

Author Note:
Correspondence concerning this article should be addressed to Charles R. Honts, Ph. D., Department of Psychology, Boise State University, 1910 University Drive, MS-1715, Boise, ID 83725-1715. E-mail: chonts@boisestate.edu.

The views expressed in this article are those of the authors and do not necessarily represent the views of the American Polygraph Association.

that one or the other of the classes of critical questions has generated more mental effort and that was reflected in the autonomic physiology (Honts, 2014; Vrij & Gannis,2014). A substantial body of scientific literature indicates that with traditional probable lie comparison questions this differential mental effort is expressed as a reduction in respiratory activity (Bell, et al., 1999). Numerical scores are assigned to polygraph examination data through the comparison of these respiratory activity changes in response to relevant and comparison test stimuli.

The Directed Lie Comparison (DLC) question variant of the CQT has been in use for three decades (Menges, 2004). Laboratory and field studies indicate high levels of criterion accuracy using the DLC variant (Honts & Alloway, 2002; American Polygraph Association, 2011; Barland, 1981; Department of Defense Polygraph Institute [DoDPI], 1997; DoDPI, 1998; Honts & Raskin, 1988; Honts & Reavy, 2009; Horowitz, Kircher, Honts & Raskin, 1997; Kircher, Packard, Bell & Bernhardt, 2001; Nelson & Handler, 2011; Nelson, Handler & Morgan, 2012; Nelson, Handler, Blalock, & Hernandez, 2012; Reed, 1994).

The few studies directly comparing the DLC variant of the comparison question to the Probable Lie Comparison (PLC) variant indicate no significant difference in criterion validity (Honts & Raskin, 1988; Honts & Reavy, 2009; Horowitz, Kircher, Honts, & Raskin, 1997; Kircher, Packard, Bell, & Bernhardt, 2001). However, two studies (Horowitz, et al., 1997; Kircher, et al., 2010) reported an unexpected finding in the respiration responses of truthful subjects when the DLC variant was used. Those two studies failed to find significant respiratory differential reactivity in objectively measured respiratory line length. In their review, Kircher and Raskin (2002) noted the differences between objective measures of respiratory line length with probable and directed variations of the CQT and called for additional research. However, to date, we know of no research that directly addresses potential respiration effects of comparison question type on field evaluations.

Current Controversy in Field Practice

The Directed Lie Screening Test, or DLST, (Handler, Nelson, & Blalock, 2008) is a variant of the Test for Espionage and Sabotage (DODPI 1997, 1998). The DLST has been presented at international, national and state polygraph association meetings and uses DLCs as the comparison questions. One question that sometimes arises during discussion of the DLCs is the concern over how to evaluate the respiration channel, whether with a computer algorithm or by hand scores. Because of the unexpected respiration results of Horowitz et al., (1997) and Kircher et al., (2001) examiners seek clarification for evaluating the respiration channel. Unfortunately, Horowitz et al., (1997) limited their evaluation of this unexpected pattern to analyses of computer-extracted measurements. They reported objective measurement differences scores between relevant and comparison questions for innocent subjects tested with DLCs were negative, indicating greater suppression to the relevant question than to the DLC. They stated that a review of the numerical scores was consistent with this finding but did not elaborate further.

Kircher et al., (2001) provided an extensive discussion of the unexpected findings in the quantitative (computer) data and the qualitative (hand score) data. The total numerical scores for truthful and deceptive DLC and PLC subjects were similar. No significant differences were found between the PLC and DLC subject in the computer outcomes for either the truthful or deceptive groups. The American Polygraph Association (2011) reported no significant differences in either criterion accuracy or total numerical scores for DLC and PLC variants of the examination technique described by the group of researchers from Utah.

However, when using computer measurements, Kircher et al. (2001) reported a significant interaction of guilty and question type for respiration excursion and DLC. Although deceptive DLC subjects responded in the expected manner, truthful DLC subjects did not. Both truthful and deceptive subjects showed a reduction in respiration excursion to the relevant questions as compared to the DLC questions. This unexpected effect was

also reflected in the numerical scores, but interestingly in Kircher et al. (2001) neither the PLC nor the DLC numerical scores were significantly correlated with the guilt criterion.

To explore these issues further we examined data from two studies. Honts and Raskin (1988) remains the only field validity study that directly compares probable lie and directed lie questions. However, Honts and Raskin only reported total score and did not report any analyses of individual physiological measures. Honts and Reavy (2009) is the largest controlled experiment ever conducted comparing probable lie and directed lie comparison question tests. We extracted respiratory channel scores from those two studies and subjected them to analysis for comparison question type differences.

Analysis of Field Data

Data Source 1

Honts and Raskin (1988) published a field study of the validity of the directed lie approach to creating comparison questions. Their subjects were 25 criminal suspects who were referred to them for polygraph examinations in forensic settings. Their referrals were from both prosecution and defense counsel. These 25 individuals were an exhaustive sample of the authors' confirmed examinations conducted between January 1983 and January 1987 where both DLC and PLC questions were used in the same examination. They considered cases to be confirmed if subsequent to the polygraph examination the subject confessed, or if some other suspect confessed and exonerated the subject, or if the accuser of the subject later retracted the accusations in a formal setting, such as court, or if physical evidence was developed that conclusively exonerated the subject. According to those confession criteria, 13 examinations were confirmed with actually innocent subjects and 12 were confirmed as being conducted with actually guilty subjects.

In Honts and Raskin (1988) each examination contained three relevant questions and three comparison questions arranged in the form that is now known as the Utah Zone, and which was first described in detail in Kircher and Raskin (1988). In Honts and Raskin comparison questions in the C1 and C3 positions were presented as PLCs. The comparison question in the C2 position was presented as a DLC. Across three presentations, questions rotated in position so that each comparison question was compared to each of the relevant questions. After all the identifying information was removed from the charts, they were blindly rescored by both examiners. Two scores were developed during the blind evaluation. The first score used the DLC in a standard application of the Utah Scoring Rules (Raskin & Hare, 1978). The second score did not use the DLC, but substituted the temporally closest PLC to make the scoring. Honts and Raskin (1988) report only the total scores for each subject based upon the blind rescoring of the examiner who was not the original examiner.

The original examiners in Honts and Raskin (1988) reached conclusive decisions in 24 of the 25. The inter-rater reliability of the two blind scorings was very high, $r = 0.92$. Effects of the use of the DLC were tested in several ways. In terms of overall discriminative values, the total scores and decisions including the DLC were more accurate than the total scores based only on PLC questions. However, Honts and Raskin reported that none of these differences reached statistical significance. Honts and Raskin did not report any analyses of the component scores.

New Analyses

The original score sheets from Honts and Raskin were maintained in the first author's files. We extracted the respiration component scores from those score sheets and subjected them to new analyses to look at effects of the DLC on the scoring of respiration. Since these charts were scored with the standard Utah rules many years before there was any suggestion that DLCs produced anomalous respiration responses, one would expect that if the respiration responses were anomalous that it would have an impact on the numerical scores generated when a DLC was used as a comparison.

Each subject's total respiration score for DLC comparisons was tested against each subject's total respiration scores for PLC comparisons at that relevant question with a mixed factor ANOVA. Comparison Question

Type (DLC v. PLC) was entered into the analysis as a within-subjects factor while Guilt (Innocent v. Guilty) was entered as a between-subjects factor. Modern concerns about the DLC producing anomalous respiration response would predict that the interaction between Guilt and Question Type should be significant. In our analysis, neither the main effect of Question Type nor the interaction of Guilt and Question type were statistically significant, $F(1, 23) = 0.010$, $p = 0.92$, *ns*, partial $\eta^2 > .001$ and $F(1, 23) = 0.37$, $p = 0.55$, *ns*, partial Eta square = .016, respectively.

Discussion of the Analysis of Field Data

The results from our analysis of the respiration scores from Honts and Raskin (1988) failed to find any indication that applying standard scoring rules to comparisons with DLC questions would produce anomalous or misleading results. These data were analyzed in a within subjects design with a powerful statistical test that would very likely have revealed any effects, if they existed. However, the data sample was relatively small and the data did represent the results from comparison to a single relevant question on each chart. To further explore this issue we looked to the data from a recent large laboratory study that directly compared directed lie and probable lie examinations.

Analysis of Laboratory Data

Data Source 2

Honts and Reavy (2009) conducted a large experiment designed to test for differences in validity between examinations run with DLC questions and examinations run with PLC questions. Honts and Reavy tested 250 (126 female, 124 male) participants who were recruited via help-wanted ads on craigslist.com and in a local alternative newspaper. Participants were paid an hourly wage of $15 for approximately 2 1/2 hours of participation in the study. Individuals who were currently pregnant, taking prescription medication for high blood pressure, a heart condition, or to treat a psychological disorder, or had previously taken a polygraph examination, were deemed ineligible for participation. Those who met the selection criteria were randomly assigned (see Honts & Reavy for details about the double-blind random assignment procedures) to one of eight experimental conditions in Guilt (Innocent v. Guilty) X Question Type (DLC v. PLC) by Between Chart Review (Review v. No Review). Each of the eight cells was made up of approximately 30 participants. Cell assignment varied from a low of 29 to a high of 34. Participants ranged in age from 18 to 65 years (Mode = 20, M = 30, SD = 10.5). Honts and Reavy found no significant differences in Objective Scoring System, Version2 (OSS2; Krapohl, 2002) total scores between tests conducted with PLC and DLC questions. No component score data were reported in Honts and Reavy (2009).

New Analyses

The Honts and Reavy (2009) data were evaluated with OSS2 and with independent scoring with the Utah Scoring System (Bell et al., 1999). We extracted the respiration scores from both of those scorings and subjected them to new analyses. We had useable data from 249 participants (the computerized data for one participant was lost due to file corruption).

OSS2

The OSS2 is a computer based analysis system that is designed to mimic the process of numerical scoring. One difference between the OSS2 and common numerical scoring procedures with respiration data is the way scores are assigned. OSS2 scores are assigned using a mathematical comparison of precise measurements to identify comparative reductions in respiration activity and to assign scores. Numerical scoring procedures involve the use of a pattern recognition approach to approximate the linear measurement of respiratory suppression. An important aspect of pattern recognition rules used in this study is a requirement that patterns be observed for a minimum of three respiration cycles if they are to be scored. The result is that numerical scoring tends to be more conservative in score assignment, with a modal score of zero (0) and corresponding sub-total scores that are closer to zero than the scores of the OSS2. Raskin and Kircher (2014) recently reported a large field study where the OSS2 had the highest criterion validity of the five computer-based methods they tested. The OSS2 respiration component scores from the Honts and Reavy data were subjected to a Guilt by Question

Type ANOVA. We collapsed across their review variable as it was of no interest for the questions raised here. The ANOVA of the OSS2 respiration component scores resulted in a significant main effect for Guilt, $F(1, 245) = 35.9$, $p < .001$, partial $\eta^2 = 0.13$. The average total OSS2 respiration score for Innocent participants was 0.40 ($SD = 8.89$) while the average for Guilty participants was -6.29 ($SD = 8.81$). The ANOVA also revealed a statistically significant, but small, main effect for Question Type, $F(1, 245) = 3.93$, $p = 0.049$, partial $\eta^2 = 0.016$. The average total respiration score for participants tested with PLC questions was -1.85 ($SD = 8.62$) while the average for participants tested with DLC questions was -4.04 ($SD = 10.10$). The interaction of Guilt and Question type did not approach significance, $F(1, 245) = 0.044$, $p = .83$, partial $\eta^2 > 0.001$. The lack of a significant interaction indicates that there was no detectable anomalous effect of the use of the DLC on the OSS2 scoring of the respiration responses of the innocent subjects.

Utah Numerical Scores

The Utah respiration component scores were also subjected to a Guilt by Question Type ANOVA. The ANOVA of the Utah respiration component scores resulted in a significant main effect for Guilt, $F(1, 245) = 15.21$, $p > .001$, partial $\eta^2 = 0.058$. The average total Utah respiration score for Innocent participants was 0.41 ($SD = 2.88$) while the average for Guilty participants was -0.91 ($SD = 2.51$). The ANOVA also revealed a statistically significant, but small, main effect for Question Type, $F(1, 245) = 5.45$, $p = 0.02$, partial $\eta^2 = 0.022$. The average total respiration score for participants tested with PLC questions was 0.15 ($SD = 2.74$) while the average for participants tested with DLC questions was -0.64 ($SD = 2.77$). The interaction of Guilt and Question type did not approach significance, $F(1, 245) = 0.047$, $p = .83$, partial $\eta^2 > 0.001$. The lack of a significant interaction indicates that there was no detectable anomalous effect of the use of the DLC on the Utah Numerical scoring of the respiration responses of the innocent subjects.

We also examined the correlation of the OSS2 and Utah respiration component scores with the Guilt criterion. The OSS2 respiration component scores significantly predicted the criterion, $r(249) = .35$, $p > 0.001$, as did the Utah scores, $r(249) = .22$, $p > 0.001$. The difference between those two correlations was not significant, $z = 1.4$, *ns*.

Discussion

In this study we examined the potential impact of the use of the directed lie on respiration component numerical scores with new analyses of two existing data sets. The two data sets included cases from field and laboratory examinations. For field cases, neither the main effect of Question Type nor the interaction of Guilt and Question type were statistically significant indicating that the DLC questions in that study did not produce anomalous respiration responses. With laboratory cases, our analyses failed to reveal a significant interaction between comparison question type and guilt in either Utah or OSS2 numerical scores. The analysis showed a significant main effect for guilt status for both PLC and DLC questions, along with significant correlation of scores from both PLC and DLC questions with the criterion state.

The lack of significant interactions between guilt/innocence and comparison question type in these two studies provides evidence that applying standard respiration scoring rules advocated by the Utah scoring system (Bell et. al, 1999) to examinations using DLC questions does not produce anomalous or misleading results. We are unable to explain the differences between our data and the data reported by Horowitz et al. (1997) and Kircher et al. (2001). However, one clear difference between the studies reported here and the previous studies concerns the examiners who conducted the examinations. Both examiners in Honts and Raskin (1988) were experienced field examiners. In Honts and Reavy (2009) a third of the examinations were conducted by an experienced examiner and the other examination were conducted by students trained by the experienced examiner. None of the examinations in Horowitz et al., or Kircher, et al., were conducted by experienced examiners. It may be that there was some important difference between how the examinations in these studies were conducted

that is not obvious at this point. Since there are conflicting results in the literature, additional work on this question is recommended. However, Horowitz et al. (1997) did not find significant effects on total numerical scores after standard numerical scoring despite reporting respiration differences for PLC and DLC questions. Although Kircher et al. (2010) reported respiration differences between DLC and PLC questions, neither produced scores that discriminated truth and deception at better than chance levels in that research. Overall, Kircher et al. failed to find differences in the criterion validity of the two techniques. Thus neither of those studies provided persuasive evidence that tests with DLC questions should not be scored with the normal rules. Moreover our results clearly suggest in the absence of data indicating otherwise, that experienced examiners who use DLC questions might effectively score respiration responses to DLCs using procedures similar to those of the Utah numerical scoring method.

References

American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph technique. *Polygraph*, 40, 193-305.

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

DoDPI Research Division Staff, (1997). A comparison of psychophysiological detection of deception accuracy rates obtained using the counterintelligence scope polygraph and the test for espionage and sabotage question formats. *Polygraph*, 26, 79-106.

DoDPI Research Division Staff. (1998). Psychophysiological detection of deception accuracy rates obtained using the test for espionage and sabotage. *Polygraph*, 27, 68-73.

Handler M., & Nelson R. (2007) Polygraph terms for the 21st century. *Polygraph*, 36, 157-166.

Handler, M. D., Nelson, R., & Blalock, B. (2008). A Focused Polygraph Technique for PCSOT and Law Enforcement Screening Programs. *Polygraph*, 37, 100-111.

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

Honts, C. R., & Alloway, W. (2007). Information does not affect the validity of a comparison question test. *Legal and Criminological Psychology*, 12, 311-312. (Available online in 2006).

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

Honts, C. R., & Reavy, R. (2009). Effects of Comparison Question Type and Between Test Stimulation on the Validity of Comparison Question Test. Final Progress Report on Contract No. W911Nf-07-1-0670, submitted to the Defense Academy of Credibility Assessment (DACA). Boise State University.

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

Kircher, J. C., Packard, T., Bell, B. G., & Bernhardt, P. C. (2010). Effects of prior demonstrations of polygraph accuracy on outcomes of probable-lie and directed-lie polygraph tests. *Polygraph* 39, 22–67.

Kircher, J. C., & Raskin, D.C. (2002). Computer methods for the psychophysiological detection of deception. In: Kleiner, M. (Ed.), *Handbook of Polygraph Testing*. Academic Press, London, pp. 287–326.

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

Krapohl, D. J. (2001). A brief rejoinder to Matte & Grove regarding "psychological set." *Polygraph*, 30, 203-205.

Krapohl, D. (2002). Short Report: An update for the Objective Scoring System. *Polygraph*, 31, 298-302.

Menges, P. (2004). Directed Lie Comparison Questions in Polygraph Examinations: History and Methodology. *Polygraph*, 33, 131-142.

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

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

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

Raskin, D. C., & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In, Raskin, D. C., Honts, C. R., & Kircher, J. C. *Credibility assessment: Scientific research and applications: First Edition* (pp. 63-130). Academic Press.

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, 109-117.

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

Vrij, A., & Gannis, G. (2014). Theory of detection of deception. In, Raskin, D. C., Honts, C. R., & Kircher, J. C. *Credibility assessment: Scientific research and applications: First Edition* (pp. 299-372). Academic Press.