Pneumograph Signal Processing and Feature Extraction

Raymond Nelson and Mark Handler

Abstract

Respiration line length, measured as excursion, was studied using N = 72 segments of data from 36 stimulus presentations sampled from comparison questions tests conducted during confirmed field investigations. Raw data were compared to signal processing models including low-pass filtering and interpolation of answer-movement artifacts. Measurements using raw data explained 2.4% of the criterion variance, while processed data explained 15.6% of the variance; a difference that was statistically significant (p = .018). After combining the data from thoracic and abdominal sensors, excursion measurements using the processed respiration data concurred with the binary case status for 69.4% of the sample segments, with a criterion coefficient that accounted for 18.5% of the variance in case status. Results using raw data concurred with the criterion state for 61.1% of the sample segments, and produced a criterion coefficient that accounted for 4.8% of the variance in case status. Scores using the filtered data concurred with the criterion status at a rate that was significantly greater than chance (p = .007), while results using the raw data were not significantly different than chance (p = .086). Sample data indicate that high sampling rates can introduce non-diagnostic noise that significantly reduce the diagnostic value of respiratory excursion measurements, while processing of raw pneumograph data can improve the accessibility of diagnostic information.

Introduction

Respiratory line length (RLL; Kircher & Raskin, 1988; Timm, 1982), measured as the linear sum of absolute changes in Y-axis excursion (Kircher, Kristjiansson, Gardner & Webb, 2005; Kircher & Raskin, 2002), is a commonly used scoring feature for breathing movement (i.e., pneumograph) data from comparison question test formats.[1] Respiratory suppression, or the relative degree of reduction in breathing movement activity, has become the respiration feature most often used in objective analysis of the time series data from comparison question test formats (Honts & Driscoll, 1987, 1988; Kircher & Raskin, 1988; Krapohl & McManus, 1999; Nelson, Krapohl & Handler, 2007; Raskin, Kircher, Honts & Horowitz, 1988). Relative strength of breathing movement suppression can be measured by comparing excursion measurements that occur in response to relevant and comparison question stimuli (Department of Defense, 2006; Harris, Horner & McQuarrie, 2000; Kircher et al., 2005).

Some differences can be observed in feature extraction and signal processing models reported in previous studies. The most notable of these differences are related to the sampling frequency. Kircher and Raskin (1988) described the digitization of analog

Raymond Nelson is a research specialist with the Lafayette Instrument Company, an experienced field polygraph examiner, psychotherapist, and elected member of the APA Board of Directors. Mr. Nelson is the developer of the open-source OSS-3 scoring algorithm, He is the author of several publications on various polygraph topics. The views and opinions expressed herein are those of the author and not the LIC or the APA.

Mark Handler is the research and information chairperson for the American Association of Police Polygraphists and the author of numerous publications on many aspects of the polygraph, including the physiological and psychological basis, testing procedures, test data analysis, interviewing, and law enforcement applicant selection.

[1] Y-axis excursion is the preferred method for measuring RLL because the hypotenuse method relies on both Y and X axes irrespective of the fact that X-axis values have nothing to do with physiology and are arbitrarily influenced by linear recording speed.

pneumograph data using 0.5 second sampling along with a 20-second measurement period. Conversion to a 60Hz sampling rate was described in another study (Kircher, Kristjiansson, Gardner & Webb, 2005), along with the use of a 10-second measurement period. In contrast, a 30Hz sampling was mentioned in a study by Harris, Horner and McQuarrie (2000), with no information about length of the measurement period. Additionally, Kircher et al. reported that pneumograph data were subject to a smoothing procedure prior to feature extraction, while Harris et al. described the similar use of a "defuzzification" procedure. Finally, Kircher and Raskin described the use of editing and interpolation to remove answer-distortion artifacts (i.e., subtle movement artifacts that occur when the examinee engages in the physical act of a verbal answer), while other studies make no mention of this procedure.

Most publications provide an absence of procedural and mathematical detail regarding the method for combining the scores or time-series data from the two respiratory sensors, though it appears likely that averaging is a common method (Harris, Horner & McQuarrie, 2005). However, Krapohl and McManus (1999) describe the retention of the stronger of thoracic and abdominal respiratory scores when the sign values concur, and setting the value to neutral (i.e., zero) when they do not concur.

Kircher and Raskin (1988) reported a point-biserial correlation (r_{pb}) of 0.55 for RLL scores and case status, with a multivariate weighting coefficient of .17. Raskin, Kircher, Honts and Horowitz (1988) reported r_{pb} = 0.39 for RLL scores in a replication study. Kircher, Kristjiansson, Gardner and Webb (2005) reported a point-biserial correlation of r_{pb} = .42 for thoracic respiration data, r_{pb} = .34 for abdominal respiration, and r_{pb} = .41 for the combined respiration data. Harris, Horner & McQuarrie (2000) did not report correlation statistics but reported an unweighted average concordance of 68.7% between RLL scores and case criterion state.

The hypothesis of interest was that objective/automated respiratory excursion measurements represent a valid and useable diagnostic feature to make statistical inferences about deceptive and truth-telling in response to verbal stimuli during comparison question testing. An additional hypothesis was that signal processing can improve the availability and usability of recorded diagnostic information in the time-series respiration data.

Design

A small sample of field examinations was obtained, consisting of confirmed tests conducted in the context of criminal investigations in a large metropolitan police agency. Veracity of deceptive cases was established by a combination of the examinee's own admission and physical evidence, while the veracity of truthful subjects was determined by a combination of a subsequent exonerating confession from another individual and physical evidence. Details of the examinee demographics, nature of the allegation or incident under investigation, and case-by-case confirmation was not made available, and numerical scores from the original examiner scores were also not available. Using six confirmed examinations that were determined to conform to established test administration protocols, a sample of 72 segments of pneumograph data was constructed, consisting of 36 stimulation presentations for which both thoracic and abdominal respiration data were recorded. Thirty-six segments of data were from examinees who were confirmed to be deceptive and 36 segments were from those who were later confirmed to be truthful.

All data were collected using LX4000 and LX5000 field polygraph instruments using both thoracic and abdominal breathing movement sensors. Pneumograph sensor design is that of two corrugated rubber tubes and beaded chains that encircle the test subject in a non-restrictive manner, to function as plethysmographic sensors that record changes in thoracic and abdominal volume and circumference. Pneumograph sensors are sealed to atmospheric pressure, and data are transformed to electrical signals via transducers that measure changes in the difference between atmospheric pressure and the pressure within the sealed thoracic and

abdominal sensors. Electrical signals are converted to digital values using 24-bit analog-to-digital conversion technology and then transmitted to a computer for recording and analysis. Sample data included 36 segments of data from thoracic respiration sensors and 36 segments from abdominal sensors.

All signal processing, interpolation, measurements, scores, and calculations were automated during this investigation. The threshold for statistical significance for this analysis was .05.

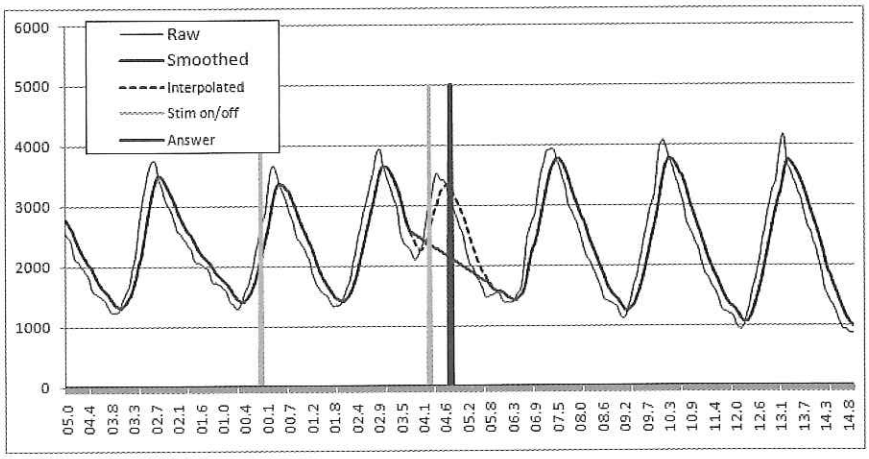
Analysis

Raw data for all segments were output to a text file format, consisting of stimulus event data and output from the analog-to-digital conversion process at a rate of 30Hz. Text files were converted into vectorized numerical data for each physiological sensor and the pneumograph data were retained for this analysis.

An algorithm was developed to automatically locate and remove answer-movement artifacts. Using the procedure described by Kircher and Raskin (1988), the linear space was interpolated for one preceding and one subsequent second surrounding each point of verbal answer. Figure 1 shows an example of an interpolated verbal answer segment from 5 seconds prior to stimulus onset to 15 seconds after stimulus onset. Also shown in Figure 1 is the raw pneumograph data.

A feature extraction algorithm was developed to measure RLL from stimulus onset until the end of a defined measurement period. RLL measurements were obtained by summing the absolute magnitude of Y-axis change for successive samples (Kircher & Raskin, 1988; Kircher, Kristjiansson, Gardner & Webb, 2005; Kircher & Raskin, 2002). Following Kircher et al. (2005), RLL measurements were taken from stimulus onset to 10 seconds following stimulus onset.

Figure 1. Interpolation of answer-movement artifact



Difference scores were calculated as the ratios of RLL measurements for relevant (R) and comparison (C) question stimuli. Ratios were calculated by dividing excursion measurements for each relevant stimulus by the excursion measurements for the comparison stimulus (Krapohl & McManus, 1999) using the stronger of the reactions at the comparison stimuli adjacent to each relevant stimulus (i.e., preceding or subsequent) following the procedure described by the Department of Defense (2006). Because the distribution of all possible R/C ratios will be asymmetrically distributed between zero

and infinity with a mean of 1, all ratios were converted to their natural logarithm to produce a normally shaped distribution with a mean of zero. Sign values were such that values above zero could be interpreted as corresponding with truth-telling, while values below zero could be interpreted as indicative of deception.

Point-biserial correlations were calculated using the logged R/C ratios and a binary value, coded as 1 or -1, representing the confirmed guilty or innocent case status, comparing each relevant question to the stronger adjacent comparison question (i.e., shorter excursion measurement). A coefficient of variation was calculated using the square of the point-biserial coefficients (r^2). This can be interpreted as a criterion coefficient, and can be understood of as the proportion of variance in binary case status that is explained by the measured feature. In other words, how much of a truthful or deceptive decision is actually accounted for by the measured pneumograph data.[2] Coefficients were calculated using unprocessed raw data, in addition to data that were processed using a first order Butterworth filter[3] designed to achieve a low-pass corner frequency of fc = 0.886Hz, and using a simple moving average smoothing filter involving a buffer of 0.5 seconds. The smoothing filter is a simple form of finite response filter with a low-pass corner frequency of .886Hz.[4]

Results

The Pearson product moment correlation for the 72 logged R/C ratios using raw and smoothed pneumograph was r = .772. The coefficient of variation indicated that 59.9% of the variance in logged R/C ratios was shared between the raw and smoothed data. The correlation statistic for the low-pass filter and moving average smoothing filter was r = .988, indicating that 97.7% of the variance in scores was shared between the two models. Because of the near-perfect correlation between these two signal processing models, only the raw and smoothing filter (i.e., moving average) data were retained for the remainder of the analysis.

The Pearson correlation statistic for the logged R/C ratios from 36 thoracic and 36 abdominal segments was r = .822 for the raw data, with a coefficient of variation indicating that 67.6% of the scored response variance was shared by the two sensors. Correlation for the logged R/C ratios using the filtered data was r = .830, indicated that 68.9% of the variance in scores was shared between the thoracic and abdominal sensors.

Logged R/C ratios using the raw data produced a criterion coefficient that explained 2.4% of the variance in case status, while the filtered data explained 15.6% of the variance. A two-sample t-test based on a bootstrap of 1000 re-sampled sets of the 72 data segments indicated that the difference was statistically significant (p = .018).

The criterion coefficient (r^2) for the thoracic respiration data indicated that the raw data accounted for 6.7% of the variance in case criterion state while the filtered data explained 22.3% of the criterion variance. Coefficients for the abdominal respiration data indicated that raw data explained 0.2% of the criterion variance of the sample data while filtered data accounted for 9.5% of the variance in case criterion state. Results are shown in Table 1.

[2] Criterion coefficients provide a better measurement of scoring feature validity compared to the simple proportion of concurrence or percent of correct scores because these simpler measures neglect to account for the likelihood that correct or incorrect scores can occur due to random chance.

[3] The first order Butterworth low-pass filter consists of a mathematical procedure in which each Output Value = Xcoeff + previous Xcoeff + Yconstant * Ycoeff, where Xcoeff = Input Value / Xconstant, and Ycoeff = the previous Output Value. Xconstant = 0.1174704212 and Yconstant = 0.8297443748 for the 30Hz sampling rate.

[4] fc = (0.443 / Number of Point) * fc, where fc = corner frequency and fs = sampling frequency.

Table 1. Point-biserial coefficients (r_{pb}) and [criterion coefficients (r^2)] for logged R/C ratios from thoracic, abdominal and combined sensors

N = 36 for all cells	Thoracic	Abdominal	Averaging	Combined
Raw	0.258 [.067]	.047 [.002]	.218 [.048]	.218 [.048]
Filtered	0.472 [.223]*	.308 [.095]	.433 [.188]*	.430 [.185]*
* statistically significant				

A Monte Carlo Bootstrap of the criterion coefficients using the filtered respiration data showed that the mean difference in criterion coefficient was significant ($p < .001$) for logged R/C ratios using the thoracic and abdominal data. R/C ratios using the thoracic sensor were significantly greater than chance ($p = .026$), while ratios using the filtered data from the abdominal sensor did not exceed chance ($p = .126$). Ratios using the raw data did not exceed chance expectations for either thoracic ($p = .140$) or abdominal ($p = .233$) respiration data. A two-way ANOVA showed that the interaction of criterion state and sensor location was not significant [F 1,68 = 1.3, ($p = .258$)].

To achieve a single respiratory response score using the two breathing-movement sensors, logged R/C ratios from thoracic and abdominal respiration sensors were combined using two different procedures. One procedure was a modification of that described by Krapohl and McManus (1999) in which the score was retained for whichever sensor produced a stronger absolute value whenever the sign values concurred for the logged R/C ratios for the thoracic and abdominal sensor data.[5] The second method combined the two logged R/C ratios by averaging. These results are also shown in Table 1.

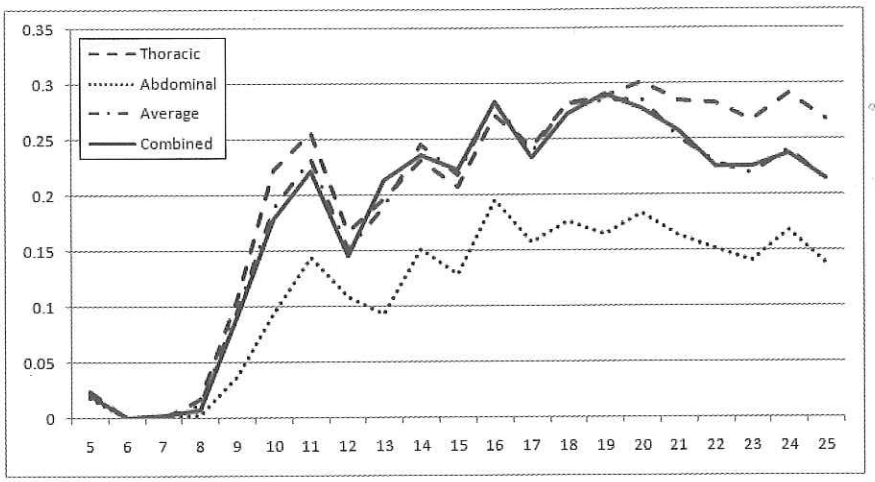
Using the filtered respiration data, the criterion coefficient for the N = 36 logged R/C ratios accounted for 18.8% of the variance in case status with the averaging model. Using the Krapohl and McManus (1999) method of combining the data from the two sensors, the coefficient explained 18.5% of the criterion variance. A Bootstrap t-test showed this difference was not significant ($p = .449$).

Sign values for RLL measurements based on filtered data concurred with the criterion state at a rate of 69.4% that was significantly greater than chance ($p = .007$), while sign values using raw data concurred with the criterion state at 61.1% which was not greater than chance ($p = .086$).

One ancillary analysis was conducted to further explore the effect of the measurement period on the criterion coefficient. Using the filtered data, RLL measurements and R/C ratios were calculated while varying the length of the measurement period at 1-second intervals from 5 to 25 seconds. Figure 2 shows a plot of the point-biserial correlation coefficients using the thoracic and abdominal scores in addition to the point-biserial coefficients when combining the two sensors by averaging and using a modification of the method described by Krapohl and McManus (1999). Coefficients showed peak diagnostic efficiency using measurement periods ranging from 10 seconds to 25 seconds. Criterion coefficients were weak using measurement periods shorter than 10 seconds.

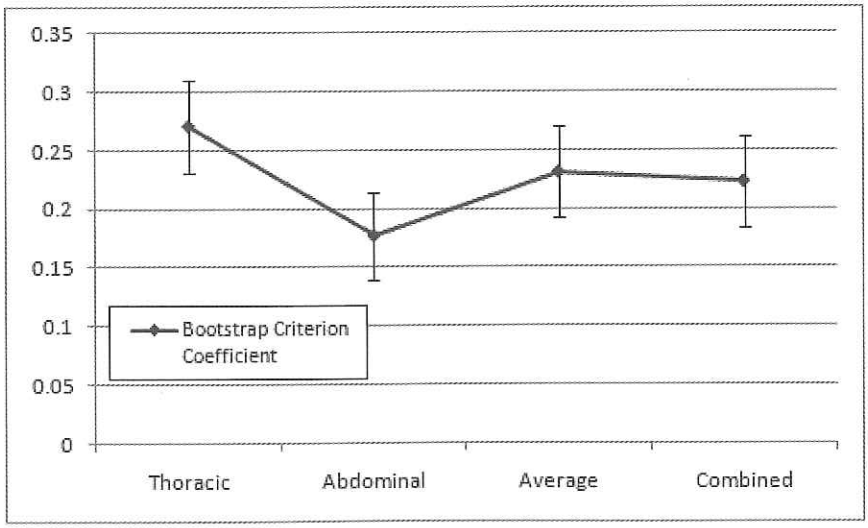
[5] Krapohl and McMannus (1999) converted R/C ratios to integer scores using a normed non-parametric transformation, whereas the present study uses a log transformation of R/C ratios. Another difference is that Krapohl and McManus set the score to 0 when the sign values from the scores of thoracic and abdominal sensors did not concur, while the present study used the thoracic pneumo score when the sign values were different.

Figure 2. Point-biserial correlations for logged R/C ratios using measurements from 5 to 25 seconds



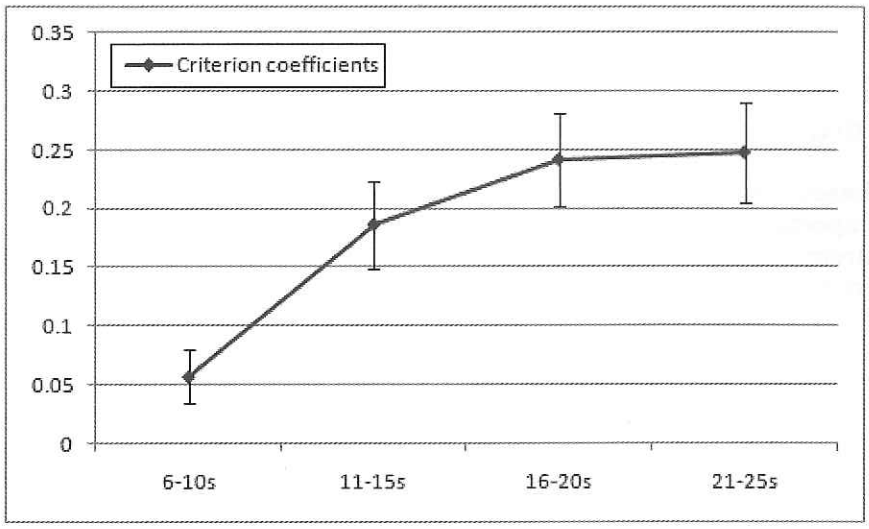
A Monte Carlo Bootstrap of the thoracic, abdominal, averaged and combined logged R/C ratios was calculated while varying the measurement period randomly from 10 to 25 seconds. A one-way ANOVA, using a Monte Carlo Bootstrap of the thoracic, abdominal, averaged and combined logged R/C ratios while varying the measurement period randomly from 10 to 25 seconds, showed that differences were significant [F (3,143) = 2.836, (p = .040)]. Post-hoc analysis showed that the difference was significant only for the thoracic and abdominal sensors [F (1,71) = 5.648, (p = .020)]. Figure 3 shows the mean plot for the Monte Carlo Bootstrap of the thoracic and abdominal sensors along with two methods of combining the sensor data.

Figure 3. Mean and 95% confidence ranges for criterion coefficients for respiration sensors



To better understand the potential effect of various measurement periods on excursion measurements, a second Monte Carlo Bootstrap ANOVA was completed to compare the results while randomly varying the measurement periods within four bins from 6 to 10 seconds, 10 to 15 seconds, 15 to 20 seconds, and 20 to 25 seconds while combining the filtered respiration data from the two sensors. Figure 4 shows the mean plot and 95% confidence intervals. Criterion coefficients differed significantly using different measurement periods [F (3,143) = 17.28, (p < .001)]. Post-hoc analysis showed no significant difference between measurements from 11 to 25 seconds [F (2,107) = 1.867, (p = .160)] and that significant differences were limited to measurements shorter than 10 seconds when compared to measurements from 11-25 seconds [F (1,71) = 23.55, (p < .001)].

Figure 4. Mean and 95% confidence ranges for criterion coefficients of binned measurement periods



Summary

Results of this study replicate and extend those of four earlier studies describing feature extraction and signal processing of pneumograph data from comparison question test formats (Harris, Horner & McQuarrie, 2000; Kircher, Kristjiansson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Raskin, Kircher, Honts & Horowitz, 1988). Difference scores, based on respiratory excursion measurements, produced criterion coefficients that were consistent with those reported in previous studies that used smoothing procedures or lower sampling frequencies. Results herein support the validity of the hypothesis that RLL, measured as the sum of absolute changes in Y-axis excursion for each successive sample, can be used in comparison question test paradigms to make inferences about deception and truth-telling at rates significantly greater than chance.

These results also support the hypothesis that signal processing,[6] in the form of low-pass filtering and interpolation of

[6] Signal processing alternatives include hardware filters, and digital signal processing methods that may be implemented in either device firmware or computer software.

answer-movement artifacts can improve the availability of diagnostic information in time-series respiration data. Filtered respiration data outperformed the raw data for this sample, accounting for a significantly greater portion of the criterion variance and resulted in respiratory excursion measurements that exceeded chance expectations. Respiratory excursion measurements using filtered data concurred with the case criterion at a rate that was significantly greater than chance, and was consistent with the results reported by Harris, Horner and McQuarrie (2000). Respiratory excursion measurements using the filtered data accounted for the variance in case criterion at a rate consistent with the multivariate weighting coefficient reported by Kircher and Raskin (1988), and the point-biserial coefficients reported by Kircher and Raskin (1988), Kircher, Kristjiansson, Gardner and Webb (2005), and those of Raskin, Kircher, Honts and Horowitz (1988).

Results based on raw data did not exceed chance expectations, suggesting that excursion measurements based on raw data may be non-diagnostic. The practical meaning of this is that high-frequency sampling and recording of pneumograph data may may offer little or no benefit when the data are to be aggregated through summation of absolute value of y-axis change, and that the diagnostic value high frequency recording data can be improved using common signal processing methods. In short: improving the visual appearance of the data through low-pass filtering can result in increased accuracy of excursion measurements and may potentially increase test accuracy.

Two different methods were compared for combining the R/C ratios from the thoracic and abdominal sensors, with nearly equivalent performance between the two methods: averaging, and retention of the stronger value. The practical implication of this is that field examiners may wish to continue the practice of combining the two using the simpler of the two methods – selecting the stronger of two scores – while retaining the thoracic value when the sign values do not concur. Additional studies should be conducted to further understand the optimal model for combining data from the two respiration sensors.

The point-biserial correlation using thoracic respiration data exceeded that using the abdominal sensor, and the difference was statistically significant. Of interest is that the coefficients based on the thoracic respiration sensor alone exceeded those from both the abdominal respiration sensor and the combination of the two. A similar result was described by Kircher, Kristjiansson, Gardner and Webb (2005), while Harris and Olsen (1994) described using only the thoracic pneumograph data.

Ancillary analysis of the influence of the measurement period on diagnostic efficiency showed that measurement periods shorter than 10 seconds are significantly less effective than those from 10 to 25 seconds. There were no statistically significant differences for measurement periods from 10 to 25 seconds, suggesting that the length of the measurement period may be a blunt issue.

This study is limited by the small sample size, absence of high-quality information about case demographics, and absence of details about case confirmation. Despite these limitations, the sample data appeared to be of average interpretable quality that is consistent with many examinations observed in field polygraph settings, and the results herein are consistent with those previously described in the published literature. An important consideration is that these results pertain to automated measurement and automated scoring. Although this information may be of secondary interest to manual scoring procedures, visual analysis of both online and printed test data may not be subject to high-frequency influence in the same way as automated measurement. A final consideration is that evidence at this time suggests that respiratory data from directed-lie comparison question tests may not hold the same diagnostic meaning as data from more traditional comparison question tests (Horowitz, Kircher, Honts & Raskin, 1997; Kircher, Kristjiansson, Gardner & Webb, 2005).

An additional limitation of the present study is the use of a hypothesis testing paradigm in the context of feature development, resulting in the potential that

real advantages may be overlooked and underutilized due to the lack of statistical significance. Although it would be generally unwise and ineffective to attempt to implement a signal discrimination model using a single diagnostic feature, statistical decision models can achieve significant results through the effective combination of features that are themselves less significant or not significant. Statistical model building involves more than the simple combination of valid criteria but requires the statistically optimal combination of a set of criteria that work together to enhance signal detection or signal discrimination.

Continued interest in respiratory excursion measurements is recommended among field examiners, researchers, and instrument manufacturers. Development efforts should continue to explore the use of signal processing methods to display and measure respiration data used in field polygraph instruments. Additional studies are recommended to further investigate the role of the measurement period in the effectiveness of respiratory excursion measurements, to further understand the potential advantages of different signal processing alternatives, and to further investigate the optimal use or combination of thoracic and abdominal respiratory sensor data. The alternatives to continued research would be to engage a static condition in which progress and improvement are absent, or to endorse hypothetical approaches without evidence, and to risk misleading both professional field examiners and those who make testing referrals with the hope of achieving high levels of decision accuracy. Finally, at a time when the use of evidence-based practices is increasingly emphasized in the medical, mental health, and forensic professions, any use of unvalidated signal processing and feature extraction models should be viewed with increasing caution.

References

Department of Defense (2006). Federal psychophysiological detection of deception examiner handbook. Reprinted in *Polygraph*, 40(1), 2-6.

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

Harris, J., Horner, A. & McQuarrie, D. (2000). An evaluation of the criteria taught by the department of defense polygraph institute for interpreting polygraph examinations. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272.

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

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., Kristjiansson, S. D., Gardner, M. K. & Webb, A. (2005). Human and computer decision-making in the psychophysiological detection of deception. *Polygraph*, 41(2), 77-126.

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(1), 22-66.

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

Kircher, J.C., & Raskin, D.C. (2002). Computer methods for the psychophysiological detection of deception. In M. Kleiner (Ed.), *Handbook of polygraph testing.* Academic Press.

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

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

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

Raskin, D., Kircher, J. C., Honts, C. R. & Horowitz, S.W. (1988). A study of the validity of polygraph examinations in criminal investigations. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040.

Timm, H. W. (1982). Analyzing deception from respiration patterns. *Journal of Police Science and Administration*, 10, 47-51.