JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2013
doi: 10.1111/1556-4029.12066
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

Frank Horvath,[1] *Ph.D.; Jamie McCloughan,*[2] *B.S.; Dan Weatherman,*[3] *M.S.; and Stanley Slowik,*[4] *M.B.A.*

The Accuracy of Auditors' and Layered Voice Analysis (LVA) Operators' Judgments of Truth and Deception During Police Questioning*

ABSTRACT: The purpose of this study was to determine if auditors could identify truthful and deceptive persons in a sample (n = 74) of audio recordings used to assess the effectiveness of layered voice analysis (LVA). The LVA employs an automated algorithm to detect deception, but it was not effective here. There were 31 truthful and 43 deceptive persons in the sample and two LVA operators averaged 48% correct decisions on truth-tellers and 25% on deceivers. Subsequent to the LVA analysis the recordings were audited by three interviewers, each independently rendering a decision of truthful or deceptive and indicating their confidence. Auditors' judgments averaged 68% correct decisions on truth-tellers and 71% on deceivers. Auditors' detection rates, generally, exceeded chance and there was significantly ($p < 0.05$) greater confidence on correct than incorrect judgments of deceivers but not on truth-tellers. These results suggest that the success reported for LVA analysis may be due to operator's judgment.

KEYWORDS: forensic science, forensic credibility assessment, layered voice analysis (LVA), lie detection, detection of deception, detecting deception in the voice

Since the 1970s, devices said to be useful for detecting deception in the human voice have been widely marketed in the United States and other countries. At least 15 such devices have appeared in the marketplace and it has been reported that many law enforcement agencies now use such equipment for forensic and other purposes. Most of these devices are functionally equivalent to the first "voice stress analyzer" sold in the United States, the psychological stress evaluator (PSE) (1). This and other brand name devices are said to detect a low frequency "microtremor" in the vocal spectrum that is inversely related to "stress." As a person experiences stress, such as that which might result from lying about involvement in a criminal offense, the micro-tremor dissipates and that change in pattern, in turn, signals "deception" (2,3). However, the persistent claims of promoters have not been supported by the scientific evidence; independent studies have not shown a relationship between stress as indicated by voice stress analyzers and deception, whether in the laboratory (1–4) or in real-life conditions (5).

In more recent years, another approach to voice-based "lie detection" has appeared. This method, developed by Nemesysco (Natania, Israel), is referred to as layered voice analysis (LVA) (6). The LVA analyzes portions of the vocal spectrum using proprietary algorithms. According to Harnsberger et al. (6), who reported one of the major assessments of the LVA, Nemesysco says their device relies "upon a 'voice frequency' analysis involving the application of '8000 mathematical algorithms' to '129 voice frequencies' that are affected by 'psychological versus physiological body reactions to the stress of telling lies' " (6, p. 643). Although that description does not correspond to any particular class of analysis devices employed by speech scientists, promoters of the LVA, nevertheless, claim it is useful in detecting deception and, moreover, that it displays "deception" automatically, requiring no operator input for decision making. The scientific research, however, has not been supportive of these claims. In the study by Harnsberger et al., for example, only chance-level detection of "lies" was reported for the LVA (6). Moreover, high false positive rates were the norm across all sets of speech materials; half of the unstressed and truthful samples were classified as exhibiting stress and deception, respectively (6).

More recent studies (5,7) used the LVA and one of the microtremor devices to detect the lies of persons who had been arrested for a variety of serious crimes. Following arrest these persons were questioned about recent drug usage in accordance with the federal Arrestee Drug Abuse and Monitoring (ADAM) program. The arrestees' verbal statements about drug usage were evaluated by "experts" who used the LVA and another voice device to determine the arrestees' veracity. Because the ADAM program requires extensive drug testing of arrestees, those test outcomes served as "ground truth" against which the results of the voice analysis devices were compared. In this study, the

[1]Michigan State University, Professor Emeritus, 108 Columbia Club Dr.-W, Blythewood, SC 29016.
[2]Michigan State Police, 103 James Street, Grayling, MI 49738.
[3]National Center for Credibility Assessment, 7540 Pickens Avenue, Ft. Jackson, SC 29207.
[4]Stanley M. Slowik, Inc., 28164 Tresine Drive, Evergreen, CO 80439.

*An earlier version of this paper was presented at the Annual Meeting of the American Academy of Forensic Sciences, February 16–21, 2009, in Denver, CO.

Received 11 Aug. 2011; and in revised form 29 Feb. 2012; accepted 10 Mar. 2012.

voice devices failed to identify correctly the arrestees' deception. Only about 15% of the arrestees who recently used drugs but reported that they had not were identified as being deceptive (5,7). This result complements previous research findings, showing that the voice analyzers are as ineffective in a real world setting as they are in laboratory-based studies.

The results in these two empirical studies take on even greater weight when considered in light of the National Research Council's (NRC) survey of the scientific research on lie detection (8). With respect to voice analysis, the NRC concluded that there was no scientific basis for the use of voice stress analyzers or similar voice measurement instruments for the detection of deception. This general conclusion and the specific findings in the scientific research on voice devices make a compelling case: voice-based lie detection using the commercially marketed devices has no scientific support (8,9).

Given this general conclusion, how then to explain the testimonials from police officers and others who praise voice analysis devices, such as the LVA, as highly effective lie detectors? Harnsberger et al. (6) speculated that as the device is not capable of discriminating between truthful and deceptive utterances, the reported field success of the LVA system may be due to the skill of the operator rather than the system output. That is, the operator may pick up cues directly from the ongoing interview rather than the LVA output.

It is the case that in field settings those who use the LVA do indeed directly interact with those who undergo testing, whether criminal suspects, informants, or witnesses. LVA operators, of necessity, carry out an interview with examinees prior to or during the testing when the device is operational. The LVA operators, therefore, may rely on observations of an examinee's demeanor and other factors to judge truthfulness irrespective of the LVA output; the device serves as a prop whose presence provides psychological encouragement for an examinee to be more forthcoming than would otherwise be the case, an instance of what is referred to as the bogus pipeline (7,10). LVA operators, therefore, wrongly attribute their success to the capability of the device rather than to their skills and judgment based on other information.

That LVA operators might gain useful diagnostic information from their personal interaction with examinees is quite plausible (11–13), although there is uncertainty regarding the specific cues that might be of value (14). However, one source of information that is always available: what an examinee states and how it is stated. In other words, the content of the examinee's statements, the tone of voice, the syntax, and so forth are all cues which the operator has available. The question arises then, is it possible, as Harnsberger et al. (6) speculated, that LVA operators might make use of what they hear irrespective of what the instrument indicates?

There is no report in the literature in which that question has been addressed. However, in one of the earliest validation studies of voice analysis Kubis (15) included an assessment of the accuracy of the PSE in comparison with observers of the testing while it was in progress. In this rather elaborate series of laboratory-based mock crimes, observers who saw and heard the PSE operator and the subjects were more accurate in detecting the subjects' deception than the PSE-based outcomes, which were at chance levels. The circumstances which precipitated this study provided an opportunity to carry out with the LVA a partial replication of the Kubis research, but with real-life subjects. Here, the voice samples of persons being questioned by police polygraph examiners for suspected involvement in criminal offenses were analyzed. These samples had been evaluated by two trained LVA operators and that corpus of data made it possible to compare the judgments made in LVA analysis to those independently rendered by persons who only audited the same data. The primary research question then was would auditors be able to judge truthfulness and deception as well as persons who analyzed the same voice samples using the LVA instrumentation? Expressed in another way, is there reason to believe that the favorable reports of operators of the LVA might be due to their reliance on vocal information not derived from the "signal" produced by the device?

Method

In 2004, the Michigan State Police (MSP), in its continuing effort to enhance its investigative mission, sought to examine if the LVA would provide an effective means of lie detection. To do so, the second author, a polygraph examiner with the MSP, designed a study in which the LVA would be used to analyze audio data collected from MSP, field-based polygraph examinations. For that purpose high-quality audio recordings of the pretest interview portion of the polygraph examinations were made. These recordings were collected using a Panasonic digital recorder (Model RR-US360; Panasonic Corporate Accessibility Program, Secaucus, NJ) in high-quality mode.

Over a period of 19 months, from January 2005 to July 2006, audio recordings of pretest interviews were collected from 210 polygraph examinations conducted by MSP examiners who agreed to record their interviews, but did not play any other role in the research. All persons who were interviewed signed a consent form indicating their approval of the audio recording and of the use of that material for research purposes.

For purposes of the MSP research, two criteria were established to ensure that the recorded interviews were useful for carrying out LVA analysis. First, the interview had to be derived from a polygraph examination in which the examiner had determined that the examinee had been either deceptive or truthful. All other outcomes, for example, inconclusive, incomplete, etc., were excluded. Second, the examiner's judgment had to be confirmed by one of two statistical classification algorithms, either Polyscore, version 1.0.0.1, or the objective scoring system (OSS), version 2, both of which are widely applied in the polygraph testing community (16–19). As expected, in some cases in which an examinee was reported deceptive a confession was made by the examinee after being advised of the outcome of the polygraph examination. In these instances, of course, the examinee acknowledged his or her role in the offense under investigation and there was confirmation of the testing outcome. Because it was understood that such confirmation would not be uniformly available, the presence of such information was not made a criterion for selection.

A review of the collected recorded interviews ($N = 210$) was carried out by a data facilitator, an MSP employee who had no other role in any aspect of the study. As a result of that review the facilitator determined that there were 103 interviews that met the two selection criteria. To decrease the amount of time necessary to do LVA analysis, the facilitator selected from the recorded interviews the first 75, which he confirmed met the two selection criteria. The facilitator then created a master list for each of the interviews on which he preserved information about the polygraph examination outcome (i.e., truthful or deceptive), the statistical algorithm result and other data necessary to identify uniquely each interview in a file created with the

Statistical Package for the Social Sciences (SPSS; IBM Corporation, Armonk, NY), version 12.0.0. He then forwarded each of the 75 audio recordings to the second author for further processing as required for LVA analysis.

LVA Analysis

In preparation for the MSP's interests, the MSP and V, LLC, the firm that markets the LVA in the United States, agreed that two MSP employees would undergo the standard Level 1, 40-h training program offered at the marketer's training facility (Waupun, WI). Each of these persons was certified as having satisfactorily completed all of the required training, LVA 6.50, Training Phase I, and returned to their position with the MSP prior to involvement in this project.

One of these trained LVA operators, the second author, who was blind to details about the cases, processed the audio data from each of the 75 interviews forwarded to him by the facilitator. This processing was done so that the audio data in each interview pertaining to all of the examinees' statements and responses to questions about the incident under investigation were included. The audio portions related to the examinees' responses to educational and other background information were excluded. In this way, LVA analysis would be carried out only on the audio data pertinent to the interviewee's denial of involvement and explanation of activities, if any, related to the offense under investigation. The audio processing also ensured that each LVA operator analyzed exactly the same information and it reduced the processing time required of the operators.

After all audio recordings were processed, they were analyzed with the LVA using the "Off-line Mode" to generate a report based on the LVA's internal algorithm processing. Then, both LVA operators, working independently, used the LVA algorithm reports to render their decisions of truthful, deceptive, or inconclusive. They did not have access to or make use of the raw audio files for their decisions. In their analyses, each LVA operator took into consideration the algorithm's outcome and the nature of the vocal segment that was processed. An inconclusive judgment indicated that the LVA operator was unable to render a definitive outcome, a result which LVA operators are taught to use if it is likely that external conditions (e.g., high emotionality or a suspected psychological disorder) are influencing the readings. In this project, the LVA operators also reported an inconclusive outcome in those instances in which the LVA's internal algorithm reported conflicting results. For example, if an interviewee's verbal response such as "I didn't point the gun" showed "deception" but another statement such as "I didn't have a gun" was shown to be truthful, the operator reported an inconclusive result if the conflict was not resolved in further analysis. After each LVA operator rendered a judgment, the outcome was recorded on a prepared form and then sent to the data facilitator to be entered into the SPSS data file.

In the processing of the audio files it was learned that one of the 75 was not of sufficient clarity and it was eliminated from the LVA analyses. In the remaining sample of 74 there were 31 persons who had been reported by the polygraph examiner as truthful and 43 who had been reported deceptive. In both cases, of course, the outcome had been confirmed, as required by the selection criteria, by the polygraph examination result and either the Polyscore or the OSS-2 computer scoring algorithm.

The LVA operators produced inconclusive results, on average, in 36% of their calls and when those cases were excluded, that is, when they both reported either truthful or deceptive outcomes, their interrater agreement (Kappa) was 0.93. It was 0.58 when inconclusive calls were included.

In the 74 recordings used for analysis, 57 (77%) were of male interviewees and 17 (23%) were those of female interviewees. According to MSP documentation, the sample used in this study approximated the representation of males and females in the general population of polygraph examinees, 74% and 26%, respectively.

The Polygraph Examination Pretest Interviews

All of the polygraph examinations in this study were carried out by three state-licensed, experienced MSP polygraph examiners. These examinations were conducted consistent with the comparison question technique (CQT), the most well known and widely used, but nevertheless controversial polygraph testing procedure (8). The CQT has been described in detail in other sources (8,20). Briefly stated, however, it consists of a pretest interview followed by the collection and analysis of polygraphic (physiological) data gathered during a series of tests in which relevant (crime-related items), irrelevant (buffer items), and comparison questions (items included to permit the "scoring" or physiological responses to relevant items) are asked.

The pretest interview is a period of time prior to the testing phase, that is, the collection of physiological data, in which the examiner collects background, demographic and related information from the examinee. These include such things as employment, educational background, health record, and so forth. The examiner and examinee also discuss the nature of the investigation and the examinee's knowledge of and involvement in the offense. They also discuss the specific test questions, which will be asked during the subsequent testing phase. The examinee is asked and in response states whether or not specific acts carried out in the offense were done by him or her. Thus, there is a record of the examinee's denial of involvement and his or her explanation of activities regarding the matter under investigation.

Inspection of the recorded interviews showed that in some instances the questioning followed the general format of the Structured Pretest Interview (SPI) (13) which, in a non-polygraph-related context, is known as the Behavioral Analysis Interview (BAI), a widely used police interviewing method for determining who the police may wish to question more intensively (11,12). In other instances, the questioning was unstructured and was interviewer-determined in format. In all cases, however, the questioning pertained to whether or not the interviewee was responsible for or involved in the particular criminal act or acts under investigation.

Whether consistent with the SPI or not, the pretest interview is nonaccusatory and its length and specific content varies depending on the complexity of the case and the information which the examinee might provide. Because the interviews in this project were field-derived, occurring in a naturalistic setting, they involved different types of criminal offenses coming to the attention of the MSP Polygraph Unit: 34% of them involved theft investigations, 26% sex crimes, and 14% child abuse and molestation. The remainder of the cases dealt with a variety of other offenses. This variation in offense types, of course, also led to variation in the length of the interviews, the way they were conducted and the particular issues that were discussed. In all cases, however, the examinee's vocal expressions denying involvement in the offense under investigation were deemed to be sufficient for LVA analysis.

Auditing of Audio Tapes

A little over a year after the MSP personnel had completed their LVA analyses and had filed an internal report of the findings, the first author requested permission to have access to the collected audio recordings. The MSP granted that permission. Personal identifying information was removed from the recordings by MSP personnel and each separate file (pretest interview) was tagged with a unique identifying number. MSP personnel then converted the original digital files to a Waveform audio file format (Microsoft WAV; Microsoft Corporation, Redmond, WA); these WAV files were then copied to digital media (DVD) which were provided to the first author.

A master list of each unique WAV-file number was maintained by the MSP. The polygraph outcome in each case, the LVA operators' findings and all other information regarding the investigations and the audio files were held in confidence until auditing of all files was completed.

Three trained and experienced interviewers were recruited for this project. Two of these persons were highly experienced (each, over 20 years) in the use and evaluation of the SPI (and the BAI); both had been employed by the firm that developed the BAI interview protocol. The third auditor was less experienced (6 years of using it), but still very familiar with that process. The three auditors were instructed to listen to each of the 74 audio files and render two judgments regarding each. First, after listening to the interview each auditor was asked to render a decision of "truthful" or "deceptive." (Auditors were asked to render only dichotomous decisions, that is, not to use "inconclusive" judgments if it could be avoided.) In addition, each auditor also was instructed to indicate the degree of confidence in the decision, ranging from "1" (no confidence) to "10" (almost certain). Statistical analyses were carried out using as dependent variables the auditors' truth/deception decisions and their confidence scores. In all statistical analyses, a 0.05 rejection region was used.

Results

As described earlier, the LVA operators' judgments were rendered prior to the auditors' judgments and neither those operators nor the auditors were aware of judgments of the other group. It was of interest though to compare the outcomes of the two groups. This is done in two ways in this section. First, the LVA decisions and the judgments rendered by the auditors are assessed relative to the ground truth criterion, the outcome of a polygraph examination in each of the 74 cases. Since each examination was also supported by the classification decision of a computerized scoring algorithm the examination outcome was free of whatever nonpolygraphic information the examiner may have had access to.

A second way to consider the LVA and auditors' decisions is to assess them relative to the "ground truth" established in those instances in which a confession was made following the polygraph examination. There were 18 such instances; unfortunately in all of these cases the confession only implicated the interviewee as the "guilty" (deceptive) person; there were no instances in which the confession also exonerated another person or persons in the same investigation.

LVA and Auditors' Judgments: All Cases

In Table 1, the findings for both the LVA operators and the auditors are shown for the 31 interviewees who were truthful.

TABLE 1—*Judgments of auditors and LVA operators on audio files of truthful persons (*n = *31).*

Evaluation Mode	Correct *n* (%)	Incorrect *n* (%)	Inconclusive *n* (%)	χ^2 (df)	$p \leq$
Auditor 1	26 (84)	5 (16)	0 (0)	8.3 (1)	0.01
Auditor 2	12 (39)	19 (61)	0 (0)	0.8 (1)	n.s.
Auditor 3	25 (81)	6 (19)	0 (0)	6.5 (1)	0.01
Auditor mean%	(68)	(32)	(0)	—	—
LVA operator 1	16 (52)	4 (13)	11 (35)	3.9 (2)	n.s.
LVA operator 2	14 (45)	1 (03)	16 (52)	9.4 (2)	0.01
LVA mean%	(48)	(08)	(44)	—	—

As indicated, the auditors' correct decisions ranged between 39% and 84%; on average, they were correct in 68% of their judgments. They did not render any inconclusive judgments. LVA operator 1 had an accuracy of 52%, but 35% of the calls were inconclusive. The other operator's accuracy was 45% with 52% inconclusive judgments. On average, the LVA operators were correct in 48% of their calls and 44% of their decisions were inconclusive.

Considered individually and assuming that the likelihood of being correct was at chance levels, statistical testing using the chi-square statistic was carried out. These tests showed that two of the three auditors, as indicated in Table 1, produced accuracy on truthful interviewees that exceeded chance expectancy. For auditor 1, $\chi^2(1) = 8.3$, $p \leq 0.01$ and for auditor 3, $\chi^2(1) = 6.5$, $p \leq 0.01$. This was not true for either of the LVA operators; for operator 1, $\chi^2(2) = 3.9$, $p > 0.05$. LVA operator 2 had an outcome distribution significantly different from chance [$\chi^2(2) = 9.4$, $p \leq 0.01$], but only because there was a large number of inconclusive calls.

Table 2 displays the results for the LVA and auditors' judgments on the 43 persons who were deceptive. In that table, it can be seen that the auditors' correct judgments ranged between 58% and 81%, with no inconclusive calls; they averaged 71% correct. Both auditor 2 and auditor 3 had an accuracy that exceeded chance; for the former $\chi^2(1) = 9.5$, $p < 0.01$ and for the latter $\chi^2(1) = 5.5$, $p < 0.01$. LVA operator 1 was correct in 28% of his judgments and operator 2, 21%; they averaged 25% correct calls. As indicated in Table 2, neither of them produced an accuracy on deceptive interviewees that was greater than chance [for operator 1, $\chi^2(2) = 1.7$, $p > 0.05$; for operator 2, $\chi^2(2) = 3.9$, $p > 0.05$]. As was also the case in calls on truthful persons, there was a high percentage of inconclusive decisions, averaging 31%.

Calculation of the correlation between the pairings of the auditors' judgments on all cases showed that the Kappa statistic was 0.16 for auditors 1 and 2; they agreed in only 31% of their decisions. The Kappa statistic was 0.24 for auditors 2 and 3,

TABLE 2—*Judgments of auditors and LVA operators on audio files of deceptive persons (*n = *43).*

Evaluation Mode	Correct *n* (%)	Incorrect *n* (%)	Inconclusive *n* (%)	χ^2 (df)	$p \leq$
Auditor 1	25 (58)	18 (42)	0 (0)	0.5 (1)	n.s.
Auditor 2	35 (81)	8 (19)	0 (0)	9.5 (1)	0.01
Auditor 3	32 (74)	11 (26)	0 (0)	5.5 (1)	0.01
Auditor mean%	(71)	(29)	—	—	—
LVA operator 1	12 (28)	23 (53)	8 (19)	1.7 (2)	n.s.
LVA operator 2	9 (21)	16 (37)	18 (42)	3.9 (2)	n.s.
LVA mean%	(25)	(45)	(31)	—	—

TABLE 3—*Judgments of auditors and LVA operators on audio files of the confession-confirmed and the other deceptive examinees.*

	Confession Confirmed (n = 18)		Not Confession Confirmed (n = 25)			
Evaluation Mode	Correct n (%)	Incorrect n (%)	Correct n (%)	Incorrect n (%)	χ^2 (df)	$p \leq$
Auditor 1	11 (61)	7 (39)	14 (56)	11 (44)	0.00 (1)	n.s.
Auditor 2	14 (78)	4 (22)	21 (84)	4 (16)	0.01 (1)	n.s.
Auditor 3	13 (72)	5 (28)	19 (76)	6 (24)	0.00 (1)	n.s.
Auditor mean%	(70)	(30)	(72)	(28)	—	—
LVA 1	9 (50)	9 (50)	14 (56)	11 (44)	0.00 (1)	n.s.
LVA 2	6 (33)	12 (66)	10 (40)	15 (60)	0.01 (1)	n.s.
LVA mean%	(42)	(58)	(48)	(52)	—	—

with agreement on 50% of their calls, and 0.32 for auditor 1 and auditor 3, who agreed in 62% of their judgments.

Effect of Ground-Truth Criteria

There were 18 instances in which the interviewees' deception had been confirmed by a confession, perhaps a more certain indicator of "ground truth" than a polygraph outcome even when confirmed by a statistical classifier. It was of interest therefore to examine the results when these 18 cases were considered in comparison with the 25 others involving deceptive interviewees. Those findings, treating LVA's inconclusive outcomes as errors, are shown in Table 3. As can be seen in that table, the accuracy of each of the three auditors in the 18 confession-confirmed instances was greater than that for either of the two LVA operators. The auditors' accuracy ranged between 61% and 78% and their mean accuracy was 70%. LVA operator 1 had an accuracy of 50%; the other, 33%, and they averaged 42% correct judgments. In those cases that were not confirmed by confession, the accuracy of the auditor who had the lowest accuracy matched that of the LVA operator, who had the highest accuracy, 56%. On average, the three auditors' accuracy was 72%; the average for the LVA operators was 48%.

Statistical testing of the data shown in Table 3 was carried out to determine if the accuracy in those instances when the interviews were confirmed by confession differed from that when the interviews were not so confirmed. Chi-square tests, shown in Table 3, indicated that neither the auditors nor the LVA operators had an accuracy that statistically differed between the confession-confirmed and the other cases. In other words, the presumed certainty of the confession criterion as opposed to the less certain polygraph-testing outcome did not affect the accuracy of either the auditors or the LVA operators.

Effect of the SPI

There were six interviewees who were questioned in a way similar to the SPI protocol mentioned earlier. Although this sample size is too small to warrant confidence in the findings the issue was explored because application of the protocol is common for some forensic purposes. Two of the auditors (#1 and #3 in all tables) had extensive experience with the SPI protocol whereas auditor 2 was less experienced in using it, although he was quite familiar with its application. Inspection of the outcomes in the six cases in which the SPI was used showed that the two experienced auditors were correct in 100% of their calls. Auditor 2 was correct in two of his six calls (33%); one of these was on a truthful examinee and one on a deceptive person. His errors all occurred in instances of truthful interviewees. In these same six cases, one LVA operator (#1 in the tables) was correct on four truthful persons (67%) and incorrect on one truthful and one deceptive interviewee. The other LVA operator made three correct calls, all on truthful interviewees; he was incorrect on two truthful and one deceptive person.

Auditors' Confidence

When the auditors rendered their decisions in each case they also indicated the degree of confidence in their judgment on a 10-point scale with a "1" indicating "extremely low" to "10," "almost certain." Statistical analysis of those data across all 74 cases showed that two of the auditors were more confident in their correct judgments than in those that were incorrect. These data are displayed in Table 4. For auditor 1, the mean confidence score on correct decisions was 6.4 (SD = 2.1) and on incorrect judgments it was 4.9 (SD = 2.3); this difference was statistically significant, $t(72) = 2.7$, $p \leq 0.01$. Auditor 2 had a mean value of 6.5 (SD = 2.4) on his correct judgments and 6.0 (SD = 1.9) on those that were incorrect, $t(72) = 0.9$, $p > 0.05$. For auditor 3 the mean confidence score on correct judgments was 5.9 (SD = 2.2) and on incorrect calls it was 4.6 (SD = 1.5); this difference was significant. $t(72) = 2.4$, $p \leq 0.01$.

Because of the differences in confidence scores on correct and incorrect judgments it was of interest to explore them further. To do so, the confidence scores were separately analyzed for each category of interviewee, truthful and deceptive. Those data are displayed in Table 5, where it can be seen that the mean confidence scores on deceptive examinees were 6.1, 6.9, and 6.4 on correct judgments, respectively, for the three auditors 1, 2, and 3, whereas they were 4.8, 4.5 and 4.4 on incorrect judgments; each of these auditor' differences was statistically significant. [For auditor 1, 2, and 3 respectively: $t(41) = 1.9$, $p \leq 0.05$; $t(41) = 2.7$, $p \leq 0.01$; $t(41) = 2.6$, $p \leq 0.01$]. The confidence scores on truthful interviewees are also shown in Table 5. As is indicated there, the confidence scores did not significantly differ on correct versus incorrect decisions for any of the auditors. Thus, all three auditors were more confident in correct than incorrect decisions only when they rendered judgments on deceptive interviewees.

TABLE 4—*Mean confidence scores of auditors in correct and incorrect calls.*

	Auditor's Judgment			
Auditor	Correct Mean (SD)	Incorrect Mean (SD)	$t(72)$ =	$p \leq$
1	6.4 (2.1)	4.9 (2.3)	2.7	0.01
2	6.5 (2.4)	6.0 (1.9)	0.9	n.s.
3	5.9 (2.2)	4.6 (1.5)	2.4	0.01

TABLE 5—*Mean confidence scores of auditors in correct and incorrect calls on both truthful and deceptive interviewees.*

	Truthful (n = 31)				Deceptive (n = 43)			
Auditor	Correct Mean (SD)	Incorrect Mean (SD)	t (29) =	$p \leq$	Correct Mean (SD)	Incorrect Mean (SD)	t (41) =	$p \leq$
1	6.7 (2.2)	5.4 (2.4)	1.2	n.s.	6.1 (1.9)	4.8 (2.4)	1.9	0.05
2	5.3 (2.0)	6.7 (1.7)	1.9	n.s.	6.9 (2.4)	4.5 (1.4)	2.7	0.01
3	5.4 (1.7)	4.8 (1.9)	0.6	n.s.	6.4 (2.4)	4.4 (1.3)	2.6	0.01

Discussion

The findings in this field assessment of the LVA's value in detecting deception do not provide any reason for optimism. Here, the LVA operators produced correct calls of deception, on average, only 25% of the time when deception was verified by the polygraph examination result; when deception was not present, when persons were truthful according to the polygraph examination outcome, the LVA operators were correct only 49% of the time. When the "guilty" persons had confessed their involvement in the matter under investigation and thus had acknowledged their deception, the accuracy of LVA analysis averaged only 48%. There was no instance in which the LVA produced correct decisions beyond chance. These results are remarkably similar to those reported in the field-based study by Damphousse et al. (7). In that study, involving police arrestees who lied about drug usage, correct detection of deceptive persons averaged only 15%. Moreover, the findings here are also in line with research reported in laboratory-based assessments, such as that done by Harnsberger et al. (6). They reported that their statistical analyses suggested only chance-level performance for the LVA and "high false positive rates were the norm across all sets of speech materials…roughly half of the unstressed and truthful samples were classified by the LVA as exhibiting stress and deception, respectively. A device that is, in fact, sensitive to these states should not falsely detect them if the procedures employed actually failed to elicit them" (6, p. 648). In other words, whether in the field or in the laboratory, the available research does not show the LVA to have any ability to detect deception or, on the other hand, to identify truthfulness.

In addition to its inability to produce accurate evaluations of truth and deception, the LVA showed a relatively high rate of inconclusive decisions, averaging 35%. Although inconclusive results may not be considered errors in field settings in that they merely indicate a need for other investigative alternatives, they are, nevertheless, still a concern. Even if it were assumed that the inconclusive outcomes in this study were due, in part, to the operators' inexperience with the LVA, the accuracy rate was still unacceptable when the LVA signaled a definitive outcome of deception or truthfulness.

In their analyses, the LVA operators here had access only to the output of the device. That is, they were blind to the raw audio file. This might have contributed to their high inconclusive rates. Because they were unable to rely on outside information to reconcile conflicting LVA results, unlike operators in real life, they may have been disadvantaged. What is notable about this is that it would be expected in a situation in which a device for detecting deception is ineffective. Given such a device, conflicting outcomes would be anticipated. Two or more actually truthful or, for that matter, deceptive statements, made by the same person would not necessarily be indicated accurately or consistently. Because the operators here did not have any way to reconcile such conflicts, such as by auditing the raw audio files, they chose to report inconclusive outcomes. In real life, LVA operators have access to outside information and they may resolve conflicting LVA output by making use of that information. The high inconclusive rates here, then, appear to reinforce research findings that show the LVA algorithm-based output to be incapable of detecting deception.

In spite of what might be concluded from these and other findings on the LVA as well as other voice analysis devices, there are known differences in the vocal behavior of those who are truthful and those who are deceptive (21–24). These differences can generally be categorized as cues related to time (e.g., changes in speech length, latency of response), frequency (e.g., pitch) and intensity (e.g., amplitude of response). Although this exploratory project was not designed to determine which vocal cues were attended to, it is clear that the auditors classified the interviewees with considerable success, much better, of course, than was done with reliance on the LVA signal. This would certainly suggest, as Harnsberger et al. (6) speculated, that the reported success by field practitioners comes not from the value of the LVA, but rather from operators' ability to "read" the cues inherent in an interviewees' behavior. It must be kept in mind that here the auditors had access only to what an interviewee stated and, in general, how he or she expressed it, that is, their tone of voice, assertiveness, directness, naturalness, and so forth. The auditors had no personal interaction with an interviewee. They were not able to control what was asked and they were unable to observe how an interviewee reacted to the interviewers' questions. Yet, they were surprisingly accurate in their judgments. In addition, they showed greater confidence when they correctly assessed deceptive interviewees but not when they judged those who were truthful. These findings would suggest that there are indeed diagnostic cues to deception in verbal behaviors. This would appear to be more likely when there is systematic observation such as in the SPI and BAI (11–13).

In field-based research on deception the ground-truth criterion is a difficult and complex methodological problem. Here the outcome of a CQT polygraph examination confirmed by a computerized scoring algorithm was the primary criterion of interest. However, such examination outcomes are not ground truth in the sense that they indicate with absolute certainty truthfulness or deception (8). Nevertheless, even given that less than perfect criterion, the fact is that the LVA analyses did not correspond with polygraph testing outcomes. Yet, auditors of the same material that was also subjected to LVA analyses were considerably more accurate, suggesting that there was diagnostic information in the material that the LVA was incapable of accurately processing.

In addition to the polygraph examination outcome what would appear to be a more certain ground truth criterion, a confession, was also employed. However, this criterion, which was

not used in the selection of the sample, was available in only about 25% of the cases. In addition, it was only useful for confirming deception, not truthfulness. Although the use of confessions as ground truth can be problematic in some circumstances, they have been and continue to be the gold standard in field research on deception, especially that dealing with polygraphy. One reason for this is that research in which direct comparisons between confession-based and other ground truth criteria has been done has not revealed any effect on outcomes (25,26).

In spite of the common use of confessions to establish ground truth it is to be noted that there is a growing literature showing that under certain conditions, and even in very serious crimes, false confessions occur, sometimes leading to wrongful court convictions (27). However, there is considerable dispute and uncertainty regarding this issue. At this time the best that can be said is that although confessions may be a very useful criterion they do not always provide certainty of ground truth; they represent merely the best that is practically obtainable in real-life cases (27–29).

There is only limited research in which the accuracy of observers of real-life interviews as used in this project was assessed (11–13). That research, however, differed from this study in that observation of both verbal and nonverbal behaviors was made. Here, the auditors had access only to what the interviewees expressed verbally, either in response to questions or spontaneously. The body of research on indications of deception in verbal behaviors (21–24) shows that deception is related to such things as briefer utterances, simpler vocabulary, and less complex syntax. In addition, deception also seems to be associated with paralinguistic features such as delaying tactics, dysfluencies, and other speech disturbances (21–24). The empirical evidence also shows that deceptive persons use language differently than truthful persons (30). Unfortunately, it was not possible in this project to determine specifically which features of the vocal segments the auditors attended to. It may be that their use of cues varied as a function of their training and experiences. That might account for their rather low interrater agreement. In fact, an encouraging finding is that the two auditors (#1 and #3) who had somewhat similar training and experiences were in greater agreement with each other than with the other auditor. This would suggest that training auditors in a consistent way to attend to cues of known empirical importance might enhance the possibility for detecting deception in situations where direct interaction and visual observation are not possible. Although there is some research on this topic there has been little involving real-life, unstructured, crime-focused interviews (31). The findings here indicate that there may well be reason to continue to pursue that effort for forensic and other purposes.

It is important to note that the audio materials used in this study were drawn from the pretest interview portion of a polygraph examination. Such a setting might differ in important ways from other police questioning situations. Here, the interviewees understood the context of the questioning. They knew their verbal responses and statements during the interview, even though given freely, were to be followed by polygraph testing, the known purpose of which was to serve as a check on their truthfulness. It would be assumed that under such circumstances there would be, at the least, heightened anxiety, leading perhaps to more, or different, verbal content than what might occur in another context. Whether or not that hypothesis is true is not certain, although there is some evidence that real-life, high-stakes situations such as in this project, do improve the accuracy of deception detection (12,32).

References

1. Horvath F. Detecting deception: the promise and the reality of voice stress analysis. J Forensic Sci 1982;27:340–51.
2. Hollien H, Harnsberger J. Voice stress analyzer instrumentation evaluation. Final Report, Counterintelligence Field Agency, CIFA contract-FA-4814-04-0011. Gainesville, FL: IASCP, University of Florida, 2006.
3. Horvath F. An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. J Appl Psychol 1978;63:338–44.
4. Hollien H, Harnsberger J, Martin C, Hollien K. Evaluation of the NITV CVSA. J Forensic Sci 2008;53:183–93.
5. Damphousse K. Voice stress analysis: only 15 percent of lies about drug use detected in field test. Washington, DC: National Institute of Justice, U.S. Department of Justice, NIJ Journal, 2009;259.
6. Harnsberger J, Hollien H, Martin C, Hollien K. Stress and deception in speech: evaluating layered voice analysis. J Forensic Sci 2009;54: 642–50.
7. Damphousse K, Pointon L, Upchurch D, Moore R. Assessing the validity of voice stress analysis tools in a jail setting. Oklahoma City, OK: Oklahoma Department of Mental Health and Substance Abuse Services, 2007; Award No.: 2005-IJ-CX-0047 from the National Institute of Justice, U.S. Department of Justice.
8. National Research Council. The polygraph and lie detection. Washington, DC: The National Academies Press, 2003.
9. Eriksson A, Lacerda F. Charlatanry in forensic speech science: a problem to be taken seriously. Int J Speech Lang Law 2007;14:169–93.
10. Jones E. The bogus pipeline: a new paradigm for measuring affect and attitude. Psychol Bull 1971;76:349–64.
11. Horvath F, Jayne B, Buckley J. Differentiation of truthful and deceptive criminal suspects in behavior analysis interviews. J Forensic Sci 1993;39:793–807.
12. Horvath F, Blair J, Buckley J. The behavioural analysis interview: clarifying the practice, theory and understanding of its use and effectiveness. Int J Pol Sci and Mgt 2008;10:101–18.
13. Horvath F. Verbal and nonverbal clues to truthfulness and deception during polygraph examinations. J Pol Sci and Adm 1973;1:138–52.
14. Vrij A, Mann S, Fisher R. An empirical test of the behavioural analysis interview. Law Hum Behav 2006;30:329–45.
15. Kubis J. Comparison of voice analysis and polygraph as lie detection procedures. Aberdeen Proving Ground, MD: U.S. Army Land Warfare Laboratory, 1973; Contract DAADO5-72-C-0217 from the U.S. Army Land Warfare Laboratory.
16. Webb A, Handler M, Krapohl D, Kircher J. A comparison of the objective scoring system and probability analysis. Polygraph 2008;37:250–5.
17. Olsen D, Harris J, Chiu W. The development of a physiological detection of deception scoring algorithm. Psychophysiology 1994;31:S11 [abstract].
18. Blackwell N. Polyscore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examinations from actual criminal investigations. Ft. McClellan, AL: Department of Defense Polygraph Institute, 1998; Report No.: DoDP197-R-006. Available from Defense Technical Information Center, AD Number A355504/PAA, 1998.
19. Olsen D, Harris J, Capps M, Ansley N. Computerized polygraph scoring system. J Forensic Sci 1997;42:61–71.
20. Reid J, Inbau F. Truth and deception: the polygraph ("lie-detector") technique, 2nd edn. Baltimore, MD: Williams & Wilkins Company, 1977.
21. DePaulo B, Lindsay J, Malone B, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. Psychol Bull 2003;129:74–118.
22. Rockwell P, Buller D, Burgoon J. The voice of deceit: refining and expanding vocal cues to deception. Commun Res Rpts 1997;14:451–9.
23. Bond G, Lee A. Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. App Cog Psychol 2005;19:313–29.
24. Burgoon J, Qin T. The dynamic nature of deceptive verbal communication. J of Lang Soc Psychol 2006;25:76–96.
25. Horvath F. The effect of selected variables on interpretation of polygraph records. J App Psychol 1977;62:127–36.
26. Krapohl D, Shull K, Ryan A. Does the confession criterion in case selection inflate polygraph accuracy estimates? Forensic Sci Commun 2002, http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july 2002/index.htm/krapohl.htm (accessed December 29, 2012).

27. Warden R, Drizin S, editors. True stories of false confessions. Evanston, IL: Northwestern University Press, 2009.
28. Cassell P. Protecting the innocent from false confessions and lost confessions from Miranda. J Crim Law and Crimin 1998;88:497–556.
29. Huff C, Rattner A, Sagarin E. Guilty until proven innocent: wrongful conviction and public policy. Crime and Delinq 1986;32:518–44.
30. Vrij A. Detecting lies and deceit: the psychology of lying and the implications for professional practice. New York, NY: John Wiley & Sons, Ltd., 2000.
31. Frank M, Feeley T. To catch a liar: challenges for research in lie detection training. J App Comm Res 2003;31:58–75.
32. Mann S, Vrij A, Bull R. Detecting true lies: police officers' ability to detect suspects' lies. J App Psychol 2004;89:137–49.

Additional information and reprint requests:
Frank Horvath, Ph.D.
Michigan State University
Professor Emeritus
108 Columbia Club Dr.-W
Blythewood, SC 29016
E-mail: horvathf@bellsouth.net