Document Title:	**Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting**
Author(s):	**Kelly R. Damphousse ; Laura Pointon ; Deidra Upchurch ; Rebecca K. Moore**
Document No.:	**219031**
Date Received:	**June 2007**
Award Number:	**2005-IJ-CX-0047**

Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting

Kelly R. Damphousse
University of Oklahoma

Laura Pointon
University of Oklahoma

Deidra Upchurch
KayTen Research and Development

and

Rebecca K. Moore
Oklahoma Department of Mental Health and Substance Abuse Services

March 31, 2007

Acknowledgements: This research project was made possible through a grant from the National Institute of Justice (2005-IJ-CX-0047) to the Oklahoma Department of Mental Health and Substance Abuse Services. The report reflects the views of the research team and do not necessarily represent the official positions or policies of the United States Department of Justice or the Oklahoma Department of Mental Health and Substance Abuse Services. The research team gratefully acknowledges the assistance of John Whetsel, Oklahoma County Sheriff for allowing access to the Oklahoma County Jail. We also would like to thank Chief Bill Endler (CVSA) and Tom Winscher (V Worldwide) and the training staff at CVSA and LVA who were so helpful throughout the entire project. We also thank Christine Crossland and the external reviewers whose critical insights greatly improved this report. Finally, we acknowledge the participation of the research subjects, without whom the project would not have been possible.

Table of Contents

Page Number

Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting

Abstract

Manufacturers of Voice Stress Analysis (VSA) devices have suggested that their devices are able to measure deception with great accuracy, low cost, and little training. As a result, police departments across the country have purchased costly VSA computer programs with the intention of supplementing (or supplanting) the use of the polygraph at an estimated cost of more than $16,000,000. Previous VSA studies have been conducted using simulated deception in laboratory conditions. These earlier research projects suggest that VSA programs have the capacity to detect changes in vocal patterns as a result of induced stress. To date, however, no published research studies have demonstrated that VSA programs can distinguish between "general" stress and the stress related to being deceptive. The goal of this study was to test the validity and reliability of two popular VSA programs (LVA and CVSA) in a "real world" setting. Questions about recent drug use were asked of a random sample of arrestees in a county jail. Their responses and the VSA output were compared to a subsequent urinalysis to determine if the VSA programs could detect deception. Both VSA programs show poor *validity* - neither program efficiently determined who was being deceptive about recent drug use. The programs were not able to detect deception at a rate any better than chance. The data also suggest poor *reliability* for both VSA products when we compared expert and novice interpretations of the output. Correlations between novices and experts ranged from 0.11 to 0.52 (depending on the drug in question). Finally, we found that arrestees in this VSA study were much less likely to be deceptive about recent drug use than arrestees in a non-VSA research project that used the same protocol (i.e., the ADAM project). This finding provides support for the "bogus pipeline" effect.

Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting

Executive Summary

Purpose.

- Manufacturers of Voice Stress Analysis (VSA) devices have suggested that their devices are able to measure deception with great accuracy, low cost, and little training.
- The devices claim to measure the physiological effect experienced by subjects who try to deceive.
- Developers suggest that the effect of this stress on the vocal cords is measurable and deception can be observed as output using their product.
- There are no known tests of the ability of VSA to determine deception among arrestees.
- This project assesses the validity of two of the more popular VSA tools currently on the market - Layered Voice Analysis (LVA) and Computer Voice Stress Analyzer (CVSA) among a group of arrestees.
- Each of the VSA programs was tested to assess:

  (1) Validity – can the instruments correctly detect deception?

  (2) Reliability – How do novices and experts compare in interpretation of output?

  (3) The "bogus pipeline" effect – Do individuals deceive less likely when they think that the interviewer "knows" that they are being deceptive.

Method.

- We interviewed a random sample of arrestees in a county jail during the booking process to simulate the stress of a detention-related interview.
- We used the VSA programs to ask about recent drug use.
- Each survey was followed by a urinalysis to determine if the subject was being deceptive.
- The resulting data allowed for a comparison between "actual deception" and the VSA output.

<u>Data.</u>

- 319 arrestees completed the study.
    - 67.4% tested positive for at least one of the drugs.
    - 51.7% tested positive for marijuana.
    - 27.9% tested positive for cocaine.
    - 0.6% tested positive for opiates.
    - 5% tested positive for PCP.
    - 17.9% tested positive for methamphetamine.

<u>Results.</u>

1. Validity

- We compared urinalysis data to VSA output to determine validity (Table ES1).
- The percent of deceptive respondents correctly identified as being deceptive (the sensitivity rate) was very low (15%) for each of the five drugs.
    - The sensitivity rate for LVA averaged 21%
    - The sensitivity rate for compared with CVSA averaged 8%.
- The percent of non-deceptive respondents who were correctly classified as non-deceptive (the specificity rate) was much higher, averaging 92% across each of the drugs.
    - The specificity rate for LVA averaged 95%
    - The specificity rate for CVSA averaged 90%
- The ratio of “false positives” to “true positives” (False Positive Index) was disappointing for both VSA programs, averaging 9.4.
    - The False Positive Index for LVA averaged 5.0
    - The False Positive Index for CVSA averaged 14.1[1]
- <u>Summary. Both VSA programs show poor validity. Neither program efficiently determined who was being deceptive (poor sensitivity).</u>
    - CVSA was more likely than LVA to label a non-deceptive person as deceptive (lower specificity).
    - Only LVA showed a significant chi square statistic (for Opiates and PCP) but we urge caution interpreting these results given the small number of respondents who were deceptive (which resulted in an artificially high chi square) and the very low sensitivity rates across the board.

[1] We were not able to calculate the FPI for CVSA on two drugs (Opiates and PCP) because no deceptive respondents were correctly identified.

Table ES1. Interpretation of deception about drug use by all respondents for both VSA instruments, for CVSA, and for LVA.

Statistics for VSA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.012	2.532	2.614	4.615*	0.432	-
Sensitivity	7.8%	17.8%	20.0%	22.2%	7.1%	15.0%
Specificity	88.3%	90.2%	95.5%	94.7%	88.8%	91.5%
False Positive Index	5.5	3.25	14.0	8.0	16.0	9.4
Correctly Classified:	75.3%	79.7%	94.3%	92.6%	81.5%	84.7%

Statistics for CVSA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.065	0.988	0.129	0.233	1.284	-
Sensitivity	11.1%	22.7%	0%	0%	6.3%	8.0%
Specificity	91.4%	85.5%	93.9%	94.5%	82.8%	89.6%
False Positive Index	12.0	4.2	N/A	N/A	26.0	14.1
Correctly Classified:	86.5%	77.2%	92.8%	92.2%	75.5%	84.8%

Statistics for LVA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.208	2.925	8.484*	10.16*	0.368	-
Sensitivity	12.5%	13.0%	33.3%	40%	8.3%	21.4%
Specificity	92.1%	95.9%	97.3%	95.0%	95.6%	95.2%
False Positive Index	10.0	1.67	4.0	3.5	6.0	5.0
Correctly Classified:	87.3%	82.6%	96.0%	93.1%	88.4%	89.5%

* $p < 0.05$

2. Reliability.

- We examined the correlations between individual deception assessments by experts (trainers) and novices (interviewers with the normal one week of training).
- We expected to see correlations (measured using the Phi statistic and Cohen's Kappa) greater than 0.80.
- In Table ES2, we observe generally greater correlation between the CVSA experts and the novices than between the LVA experts and the novices.
- The correlation (Phi) between the CVSA experts and the novices range from 0.31 to 0.52 for cocaine, opiates, PCP, and methamphetamine but is not significant for marijuana.
- The correlation (Phi) is significant for marijuana (.35) but the correlations for cocaine, opiates, PCP, and methamphetamine were much less robust.

Table ES2. Correlations between expert examiners and novice examiners for each drug not including respondents who reported marijuana use but tested negative.

	CVSA Experts vs. Novices		LVA Experts vs. Novices	
Deceptive for…:	Phi	Cohen's Kappa	Phi	Cohen's Kappa
Marijuana	.160	.152	.353**	.351**
Cocaine	.389**	.389**	.109	.108
Opiates	.311**	.221**	.116	.104
PCP	.524**	.431**	.207*	.195*
Methamphetamine	.321**	.303**	.178*	.151*

* Coefficient is significant at the 0.05 level (2-tailed).
** Coefficient is significant at the 0.01 level (2-tailed).

- In Table ES3, we combined the individual assessments about deceptiveness for each drug for the CVSA sample.
- The novices made 844 deception assessments while the CVSA expert made 415 deception assessments.
  - The expert failed to "make a call" on roughly one-half of the output because of "over-modulation" problems.
  - The novices correctly identified 19.8% of the deceptive responses using the CVSA instrument compared to 19.4% for the CVSA expert.
  - The novices had a higher specificity rate (90% vs. 73%).
  - The novices had a lower False Positive Index (4.9 vs. 14.6).
- The novices seemed to do a better job of distinguishing among the non-deceptive respondents.

Table ES3. Summary table showing novice and expert interpretation of deception about all drug use by all respondents for CVSA.

Condition		**Deceptive Response**	**Non Deceptive Response**	**Total**
1. Deception Indicated	Novice	16 19.8%	78 10.2%	94
	Expert	7 19.4%	102 26.9%	109
2. No Deception Indicated	Novice	65 80.2%	685 89.8%	750
	Expert	29 80.6%	277 73.1%	306
Total	Novice	81	763	844
	Expert	36	379	415

Expert	**Novice**
Sensitivity= 19.4%	Sensitivity= 19.8%
Specificity = 73.1%	Specificity = 89.8%
False Positive Index = 14.57	False Positive Index = 4.875
Correctly Classified: 68.4%	Correctly Classified: 83.1%

- In Table ES4, we combined the individual assessments about deceptiveness for each drug for the LVA sample.
- The novices made 734 deception assessments while the LVA experts made 736 deception assessments.
  - The novices correctly identified 14.9% of the deceptive responses using the CVSA instrument compared to 4.4% for the LVA expert.
  - Both groups had about the same level of specificity.
  - The LVA experts had a False Positive Index five times greater than the novices, owing mostly to the low sensitivity score recorded by the experts.

Table ES4. Summary table showing novice and expert interpretation of deception about all drug use by all respondents for LVA.

Condition		Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	Novice	10 14.9%	32 4.8%	42
	Expert	3 4.4%	45 6.7%	48
2. No Deception Indicated	Novice	57 85.1%	629 94.3%	686
	Expert	65 95.6%	623 93.3%	688
Total	Novice	67	667	734
	Expert	68	668	736

Expert	**Novice**
Sensitivity= 4.4%	Sensitivity= 14.9%
Specificity = 93.3%	Specificity = 94.3%
False Positive Index = 15	False Positive Index = 3.2
Correctly Classified: 85.1%	Correctly Classified: 87.1%

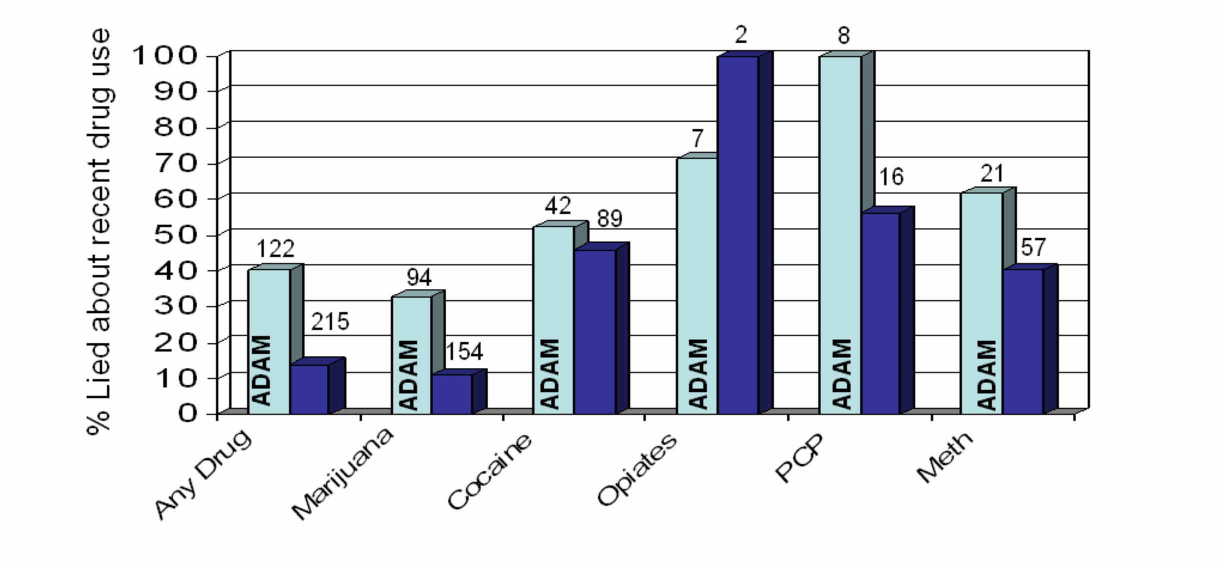
- Summary. The data suggest poor reliability for both VSA products when we compare experts and novice interpretations of the output.
    - The deception assessments by the novices were more highly correlated with the CVSA expert (except for marijuana) than with the LVA experts

3. The Bogus Pipeline Effect.

- In Figure ES1, we present a chart that compares the VSA 2006 data (the current project) with similar data (ADAM 2003) that were collected in the last quarter of 2003.
    - More than 3 times as many users of "any drug" were deceptive in the ADAM study (40%) compared with the VSA study (14%).
    - 33% of the marijuana users in ADAM study were deceptive compared to 11% in the VSA study.
    - 52% of the cocaine users in ADAM study were deceptive compared to 46% in the VSA study.
    - 100% of the PCP users in ADAM study were deceptive compared to 56% in the VSA study.

- 62% of the methamphetamine users in ADAM study were deceptive compared to 40% in the VSA study.
- There were too few opiate users in the VSA data to make a comparison possible.

Figure ES1. Comparison of "users" who were deceptive about recent drug use (ADAM 2003 and VSA 2006). Numerals reflect the number of respondents who tested positive for each drug.



- The difference between the percent of respondents who were deceptive in the VSA project (14%) compared to those who were deceptive in the ADAM project (40.2%) is statistically significant (Table ES5).

Table ES5. Comparing "Used Any Drug Recently" for ADAM 2003 and VSA 2006 Data.

	Admitted Any Drug Use	Did Not Admit Any Drug Use
ADAM-2003 (N=122)	73 59.8%	49 40.2%
VSA-2006 (N=215)	185 85.9%	30 14.0%

Chi Square = 30.0 (1), *p*<0.001

- Summary. The use of the VSA computer programs affects the likelihood of deceptive answers by arrestees.
  - Arrestees who thought their interviewers were using “lie detectors” were much less likely to be deceptive when reporting recent drug use.
  - Arrestees are more likely to be deceptive for “serious” drugs than they are for marijuana.

Project Narrative

1. Introduction and Project Background.

Correctly assessing the veracity of self-reports about deviant behavior (e.g., recent drug use) by suspects, respondents, or treatment clients can be very difficult. Doing so in a setting that might make the respondent more cautious (such as in a jail or during an interrogation) can create even greater difficulty. The inherent jeopardy increases the likelihood that respondents will not tell the truth. Research on deception concerning such information suggests that, even under guarantees of confidentiality and near anonymity, many individuals lie about recent drug use (Harrison and Hughes, 1997).

Law enforcement and treatment agencies require access to valid (i.e., "true") information to make correct decisions regarding investigations or successful treatment plans. For treatment officials, methodological "fixes" can be used to encourage greater validity (e.g., the promise of confidentiality or the use same sex interviewers). In addition, post-interview techniques can be used to validate the collected data. Law enforcement, on the other hand, tends to rely on more mechanical means (e.g., the polygraph) to detect deception. To date, there no deception detection tool has been shown capable of immediately determining if a subject is being deceptive.

Recent technological advances, however, suggest that there may be some promise in investigating the effectiveness of devices that measure stress-related physiological changes in vocal modulation patterns. Manufacturers of Voice Stress Analysis (VSA) devices suggest that subjects who try to deceive an interviewer will experience measurable stress (the same type of stress that is measured by a polygraph machine). Just as stress affects the heart rate, respiration rate, and galvanic skin response (i.e., the changes monitored in a polygraph examination), the effect of deception-related stress on the vocal cords is said to be measurable in such a way that deception can be observed. VSA devices are said to be able to measure the physiological effect of deception with (1) greater accuracy, (2) less training, (3) less invasiveness, (4) less cost, and (5) greater immediacy of findings than can be accomplished using a polygraph examination.

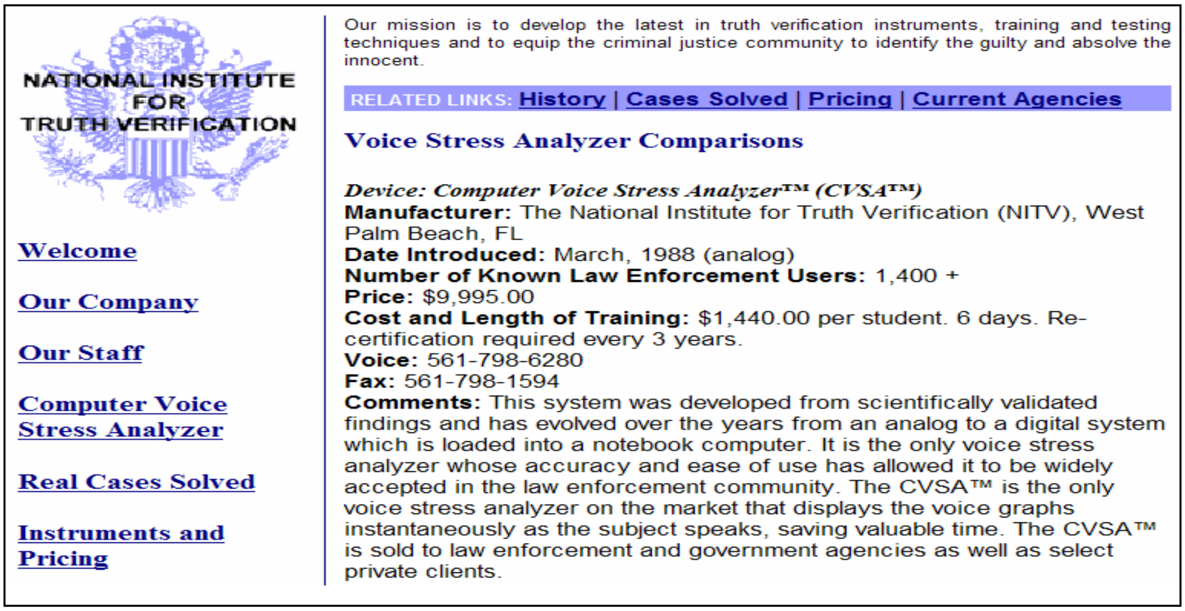
The purpose of this research project was to assess the validity of two of the more popular VSA tools that are currently on the market. The first program is called the "Computer Voice Stress Analyzer" (CVSA) and it is marketed by the National Institute for Truth Verification (http://www.cvsa1.com/index.php). It is one of the original VSA products and is probably the most widely used by law enforcement. The second program is called "Layered Voice Analysis" (LVA) and is distributed by V Worldwide (www.Vworldwide.com). Both systems are composed of computer software programs that analyze responses to questions in an attempt to determine deception. While each software package has decidedly different algorithms, they both have the capacity for the immediate examination of responses.

Voice stress analysis programs have been evaluated in laboratory settings in the past but no previously published research has tested the programs in the field. One place that both treatment and law enforcement officials are likely seek answers from a "wary" population is the jail setting. Police often interview arrestees about criminal behavior in the local jail. At the same time, jail health staff also interview recent arrestees during the booking process to determine any treatment needs or risk potential. The goal of this project, therefore, was to test the effectiveness of each VSA device to determine deception in a jail setting.

The project addressed this goal with four primary objectives. First, members of the research team received training from each of the vendors on the use of each VSA device. Second, the research team collected survey data about recent drug use from a sample of arrestees in the Oklahoma County jails. The arrestees were asked about recent drug use and the VSA programs were used to indicate the extent to which the responses were deceptive. Third, the research team collected urine data from each respondent in the sample. This allowed the research team to know which respondent was being deceptive about recent drug use. Fourth, the research team compared the correspondence between the indications of deception by the two VSA devices and the actual deception by the respondents.

The results of this analysis have important implications for law enforcement especially. As shown in the promotional material presented in Figure 1, the company that markets CVSA (NITV) reports that over 1,400 law enforcement agencies in the United States use its program.

Figure 1. Promotional Material from CVSA Website



Our mission is to develop the latest in truth verification instruments, training and testing techniques and to equip the criminal justice community to identify the guilty and absolve the innocent.

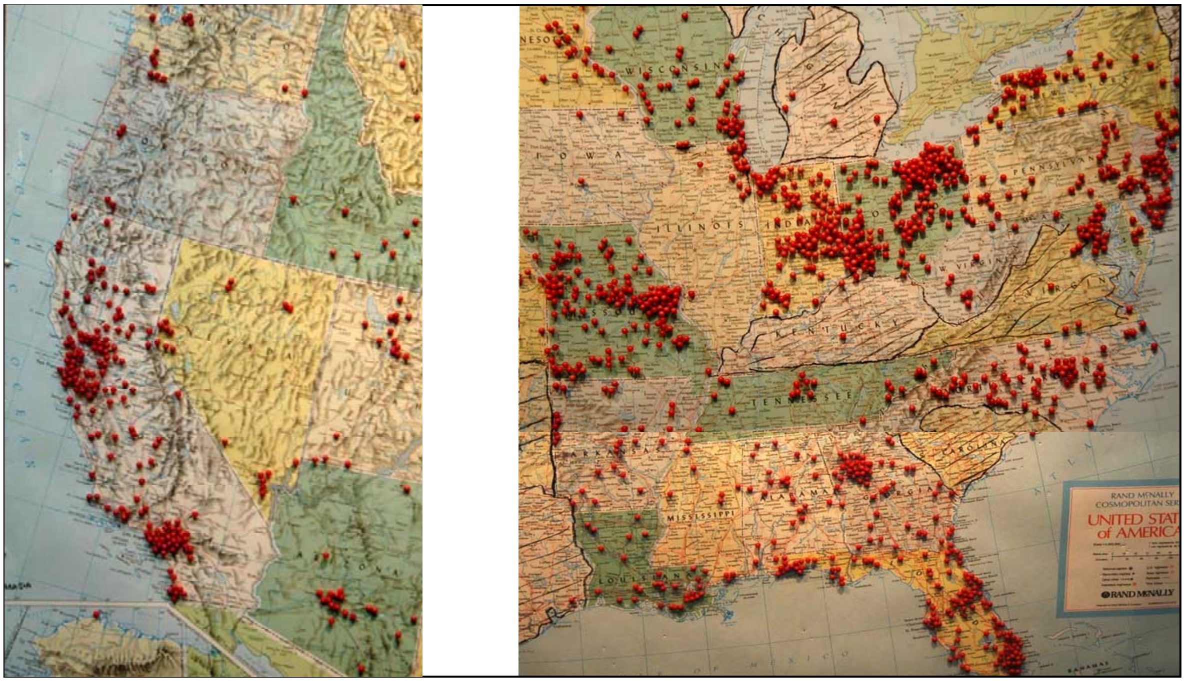
Voice Stress Analyzer Comparisons

Device: Computer Voice Stress Analyzer™ (CVSA™)
Manufacturer: The National Institute for Truth Verification (NITV), West Palm Beach, FL
Date Introduced: March, 1988 (analog)
Number of Known Law Enforcement Users: 1,400 +
Price: $9,995.00
Cost and Length of Training: $1,440.00 per student. 6 days. Re-certification required every 3 years.
Voice: 561-798-6280
Fax: 561-798-1594
Comments: This system was developed from scientifically validated findings and has evolved over the years from an analog to a digital system which is loaded into a notebook computer. It is the only voice stress analyzer whose accuracy and ease of use has allowed it to be widely accepted in the law enforcement community. The CVSA™ is the only voice stress analyzer on the market that displays the voice graphs instantaneously as the subject speaks, saving valuable time. The CVSA™ is sold to law enforcement and government agencies as well as select private clients.

Source: http://www.cvsa1.com/Products.htm (Accessed November 1, 2006).

In Figure 2, we present maps found on the CVSA website that show the distribution of agencies purported to use the CVSA product. Similar data are also presented in Table 1, where we show the number of agencies per state that NITV reports as CVSA users. The websites do not identify which departments are CVSA clients. These numbers do not include the more than 200 trained CVSA examiners in the US Military and Intelligence Agencies nor the hundreds of other agencies (both domestic and foreign) that NITV claims as CVSA users.

Figure 2. CVSA Client Map



Source: **http://www.cvsa1.com/clients.php**

It is worth noting here the cost associated with this program. The cost of the 6-day CVSA training program is $1, 440 plus at least $9,995 for the software and the laptop. Computer upgrades can increase the cost to almost $13,000. Keep in mind that these laptops are dedicated to VSA use only. They are not loaded with software that allows them to be general use computers. If each of the 1,400+ agencies sent only one person to be trained and that training cost $11,500, then the cumulative costs of the training alone has been more than

$16,000,000. Of course, that figure does not include the manpower cost (the cost of sending an officer to 6 days of training) nor the associated room and board costs. Clearly, law enforcement has invested heavily in CVSA.

Table 1. Number of Agencies Reportedly Using CVSA per State.

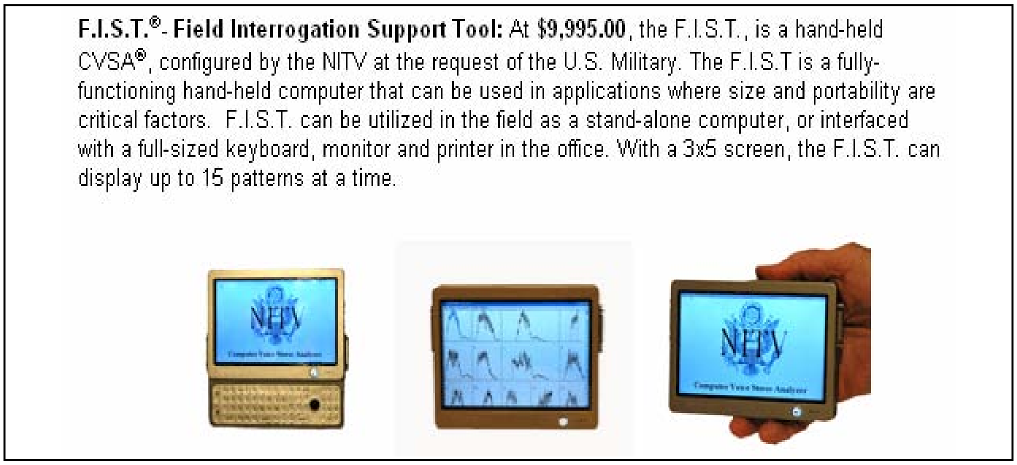
State	Number of Agencies	State	Number of Agencies
California	162	Washington	14
Ohio	161	Rhode Island	10
Florida	144	West Virginia	10
Missouri	138	Idaho	9
Indiana	101	Kansas	9
North Carolina	64	New Mexico	8
Georgia	63	Massachusetts	7
Wisconsin	56	Oregon	6
New York	53	Virginia	6
New Jersey	44	Alaska	5
Pennsylvania	43	Connecticut	4
Louisiana	36	Delaware	4
Tennessee	29	Iowa	4
Colorado	27	Kentucky	4
Illinois	27	Montana	4
Alabama	22	Wyoming	4
Maryland	22	Michigan	3
Minnesota	22	District Of Columbia	2
Utah	21	Hawaii	2
Arizona	20	South Dakota	2
Arkansas	19	Nebraska	1
Nevada	19	Texas	1
Mississippi	14		

Source: http://www.cvsa1.com/Agenciesusing.htm (Accessed December 12, 2006)

Training for LVA was even more expensive. V-Worldwide provided the training and support for LVA for a cost of $6,300. In addition, however, the cost for the software (Layered Voice Analysis Software License - Version 6.5) was $16,000 and the laptop also cost $2,500. LVA does not provide any data concerning the number of participating law enforcement agencies, but published documents state that the program is used by the Wisconsin Department of Corrections and several Wisconsin police departments (the training for LVA is in Wisconsin) along with "tests" in other agencies such as the Palm Beach police department, the State of New Mexico, and various "US government agencies that have to date not authorized the disclosure of their names" (http://ddssite.com/files/lva_article.pdf). Thus, cost is an important consideration in appreciating the effectiveness of VSA products.

In addition to cost, it is reasonable to consider issues related to privacy and potentially unethical (or at least unregulated) uses of voice stress analysis programs. As technological improvements affect the availability of this technology to the public, there are real concerns about how the program might be misused either by individuals or by organizations (e.g., employers). There are already several downloadable versions of voice stress analysis programs available on the internet. In addition, consumer products such as "lie detector" eyeglasses are now being advertised as devices that will allow users to detect deception by others with whom they interact (Johnson, 2004). As technology allows VSA devices to shrink (as shown in Figure 3), the opportunities for misuse are dramatically increased. Thus, it is vital that such programs be evaluated.

Figure 3. A Miniaturized Version of the CVSA program



F.I.S.T.®- Field Interrogation Support Tool: At **$9,995.00**, the F.I.S.T., is a hand-held CVSA®, configured by the NITV at the request of the U.S. Military. The F.I.S.T is a fully-functioning hand-held computer that can be used in applications where size and portability are critical factors. F.I.S.T. can be utilized in the field as a stand-alone computer, or interfaced with a full-sized keyboard, monitor and printer in the office. With a 3x5 screen, the F.I.S.T. can display up to 15 patterns at a time.

Source: http://www.cvsa1.com/F.I.S.T..htm (accessed on December 12, 2006).

Finally, we suggest that the findings in this report also have important ramifications for the Department of Homeland security in its mission to safeguard the country in the post 9/11 era. S. 1447 of the Aviation and Transportation Security Act as enacted by the U.S. Congress – Sec. 109 (7) (Enhanced Security Measures), for example, provided for the use of biometric technologies such as voice stress analysis to prevent a person who might pose a danger to air safety or security from boarding an aircraft. Clearly, if such technologies are to be implemented to ensure flight safety, then it is vital that empirical tests be conducted.

2. Review of relevant literature.

Since the publication of Lippold's (1971) seminal article on the physiological effect of stress on "micro-tremors," developers have attempted to capitalize on this relationship to create "lie detector" devices. Lippold (1971) measured the effect of voluntary muscle contraction on micro-tremors (small oscillations). In the field of deception detection, his work has been interpreted to mean that conscious efforts to deceive will result in measurable micro-tremors. These micro-tremors affect the vocal cords in such a way that careful examination of vocal patterns can reveal deception. In fact, most voice stress "lie detector" developers do not claim to be able to detect lies. Instead, they claim that their devices are able to detect these micro-tremors (or some such variant), which might be related to the stress of trying to conceal or to deceive. Thus, several devices have been developed and refined to measure changes in vocal response patterns when a subject is being deceptive. These software programs are generically referred to as Voice Stress Analysis (VSA) devices and various such devices have been on the market since the early 1970s. This literature review describes two VSA computer programs and subsequent validity tests of the programs.

a. VSA Computer Programs

Voice stress analyzers have a varied history of development that began in the 1970's with many individuals and companies competing in pursuit of the latest technology to introduce into the market. During this time manufactures of voice stress analyzers faced considerable resistance from the polygraph industry. In fact, several states have banned the use of VSA for the purposes of deception detection. Anti-VSA websites promoted by the groups such as the American Polygraph Association provide information about the lack of scientific support for the VSA and express concern about the proliferation of VSA programs (e.g., http://www.polygraph.org/voicestress.cfm). Concerns about the effectiveness of VSA have not appeared to dampen interest in the programs and the number of such programs continues to increase over time.

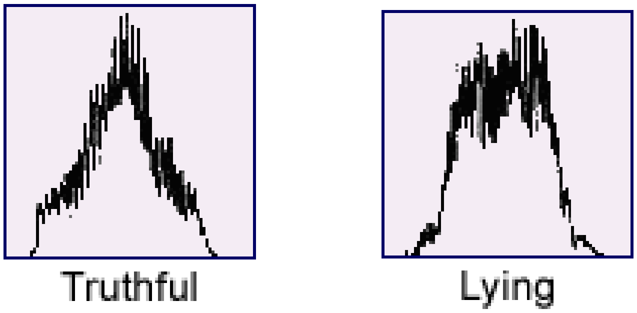
The software programs are based on assessments in changes in voice patterns caused by the stress associated with trying to hide deceptive responses. The human voice has two basic components, Amplitude Modulation (AM) and Frequency Modulation (FM). The AM sound is audible whereas the FM is not. Increased stress ultimately results in a loss of FM. CVSA captures a portrait of the change produced by the decrease in FM in the voice of a subject under stress. "The loss can be measured, the CVSA indicates the presence of stress and thus, depending on the application deception" (CVSA Manual).

There are two basic ways that VSAs work (Hopkins, Ratley, Benincasa, and Grieco, 2005). The first ("energy-based" VSA) follows closely the work of Lippold and others by detecting micro-muscle tremors that are elicited during a deceptive reply to a question. If a person is being deceptive, then he/she will experience stress. This stress will result in a measurable waveform pattern that indicates "deception." If the person is not being deceptive, then he/she will not experience stress, and his/her response will result in a measurable waveform pattern that indicates "non-deception." Commonly, energy-based VSAs feed "live" (or "on-line") audible responses into a computer program that filters the sound to create the waveform. Most programs can also perform analyses for taped (or "off-line") audible responses. The waveform for a non-deceptive response looks like an inverted V. The waveform for a deceptive response looks like an inverted U (flattened at the top) as exhibited in Figure 3.

The most well-known energy-based VSA product is CVSA, marketed by the National Institute for Truth Verification (http://www.cvsa1.com/index.php). The product known as

CVSA was introduced into the market in 1988 by the National Institute of Truth Verification. This product has undergone a number of changes and system upgrades; the current notebook version was introduced in 1997. CVSA manufacturers claim that the product detects discrete changes in the fundamental frequency of the human voice.

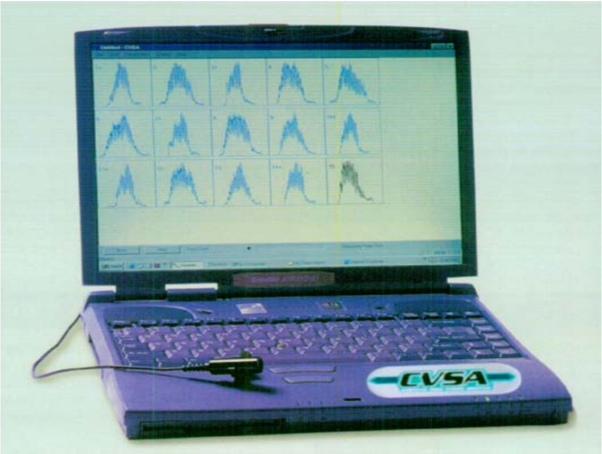
Figure 3. Results of an energy-based Voice Stress Analysis.



Source: http://www.nitv1.com/product.php

The CVSA marketing literature suggests that there are over 1,400 law enforcement users of the product. The software comes loaded onto a laptop computer and 6 days of training are required for certification to conduct deception analyses (see Figure 2).

Figure 4. CVSA laptop, microphone, and display.



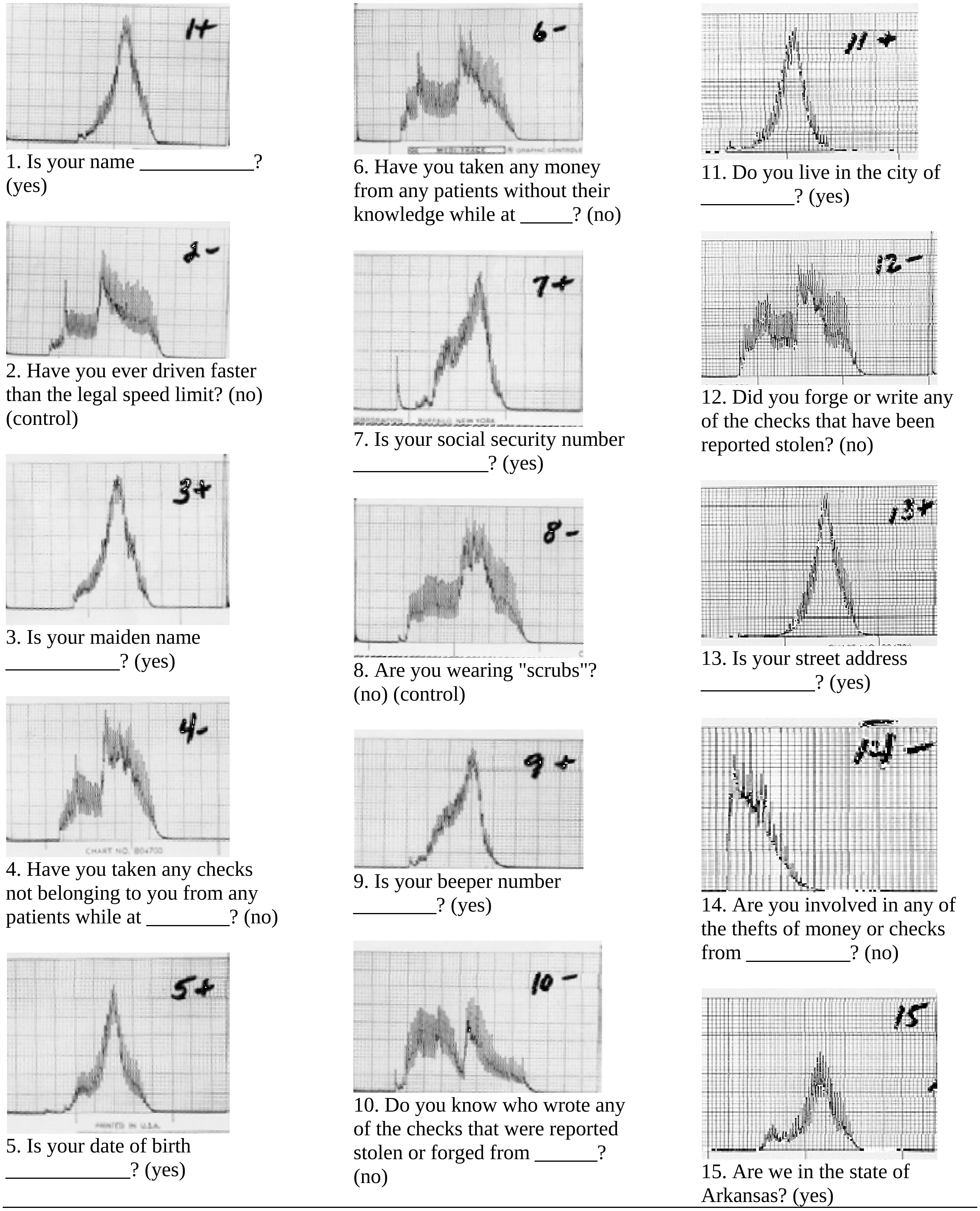
Source: CVSA brochure.

A detailed example of how CVSA charts are created and interpreted can be observed by considering the following charts from the Barling, Arkansas police department (http://campus.umr.edu/police/cvsa/news32c.htm). In the first chart (Figure 5), the respondent is asked for his name. The NDI response is expected. Then, the respondent is asked a control question on the second question - he is asked to lie to a statement about speeding. The distinctive DI pattern is established. Then, a series of alternating relevant and non-relevant questions are asked of the respondent. In this case, the CVSA program indicates that the respondent was being deceptive on questions related to the stealing of checks from his patients.

CVSA employs five test formats and the training staff emphasizes that question formulation is key to obtaining accurate results. CVSA also teaches and encourages examiners to use the NITV developed Defense Barrier Removal (DBR®) technique of interrogation and interviewing. This technique is used prior to the start of an exam in order to reduce any fears the subject may have and establish rapport between the examiner and the subject. "The successful examiner will deal with these fears by projecting an image of sincerity, empathy and impartiality within the first ten to fifteen minutes of the interview." (CVSA Manual: 89)

CVSA uses five test formats and the test format is dependant on the application being considered. Each format employs a series of relevant, irrelevant and/or control questions. For the purposes of this study only the General Series format was used. This is the format best suited for circumstances when there are more than two relevant questions or multiple issues. Relevant questions are used to obtain information that the examiner wants to know therefore must be direct and to the point. Irrelevant questions are used as buffers between relevant questions which are to be known truths and questions that are not stressful to the subject and are not related to the topic of discussion.

Figure 5. Examples of CVSA Charts from the Barling, Arkansas Police Department



1. Is your name ___________? (yes)

2. Have you ever driven faster than the legal speed limit? (no) (control)

3. Is your maiden name ___________? (yes)

4. Have you taken any checks not belonging to you from any patients while at ________? (no)

5. Is your date of birth ____________? (yes)

6. Have you taken any money from any patients without their knowledge while at _____? (no)

7. Is your social security number _____________? (yes)

8. Are you wearing "scrubs"? (no) (control)

9. Is your beeper number ________? (yes)

10. Do you know who wrote any of the checks that were reported stolen or forged from ______? (no)

11. Do you live in the city of _________? (yes)

12. Did you forge or write any of the checks that have been reported stolen? (no)

13. Is your street address ___________? (yes)

14. Are you involved in any of the thefts of money or checks from __________? (no)

15. Are we in the state of Arkansas? (yes)

Source: http://campus.umr.edu/police/cvsa/news32c.htm (Accessed December 2, 2006).

Control questions are known truths and have no relationship to the topic of discussion just like irrelevant questions. When the examiner asks the subject to deliberately lie about the known truth, however, the control question becomes a "known lie." The purpose of the control question is to provide the examiner with a comparison for the relevant responses. A lying subject's responses on relevant questions should indicate more stress than that of the control responses. If a subject is not being deceptive, their responses to control questions would appear to show more stress than the relevant questions. The test format must be followed strictly and any examiner could refer to the saved charts and recognize the test format.

As CVSA analyzes the subject's responses, it displays each voice pattern graphically in a chart format. The entire CVSA examination is carried out through a series of phases that include preparation, DBR, testing, post-test interview and re-test if necessary. The testing phase consists of asking the questions following the test format selected and obtaining chart one and completing this process again to obtain chart two. The re-test would consist of the re-questioning and obtaining a third chart. CVSA trainers teach examiners to never use the first chart to determine the final results and that there should not be a situation in which an examiner would need more than three charts. The recommended number of charts is two, although CVSA does recognize there are some instances when a third chart is necessary. Once an examiner has obtained the second chart a cold call is to be obtained. A cold call is an unbiased certified CVSA Examiner that is not involved with the case who reviews the charts and provides results in addition to the original examiner.

To be trained on the CVSA software user are required to attend a six day course covering fifty nine hours of intensive training provided by a certified CVSA trainer (see Table 2). The courses allow a maximum of 20 students and provide instruction on CVSA and DBR

Defense Barrier Removal techniques. The bulk of the training course is focused on three areas instrumentation, interviewing and interrogation as well as chart interpretation. Upon completion of the course the students are required to pass an exam before being certified and CVSA requires follow-up training to maintain certification.

Table 2. CVSA Course 59-Hour Course Outline

History of Lie Detection	1Hr
Instrumentation Students learn how to use CVSA program.	10 Hrs.
Physiology An examination of the body's reaction to stress/jeopardy ("fight" or "flight").	3 Hrs.
Interviewing and Interrogation The “art” of the interview.	15 Hrs.
Psychology Interaction of the physiological response to physiological stimuli and the physiological effects of question formulation and on deception detection.	4 Hrs.
Chart Interpretation Learning from the established criteria for an indication of deception.	16 Hrs.
Test Construction and Question Formulation Test application and construction (Zone of Comparison tests to control questions).	6 Hrs.
Covert Interviewing and Analysis Using CVSA by tape recorder or over the telephone.	4 Hrs.

Source: http://www.cvsa1.com/Training.htm (Accessed November 30, 2006)

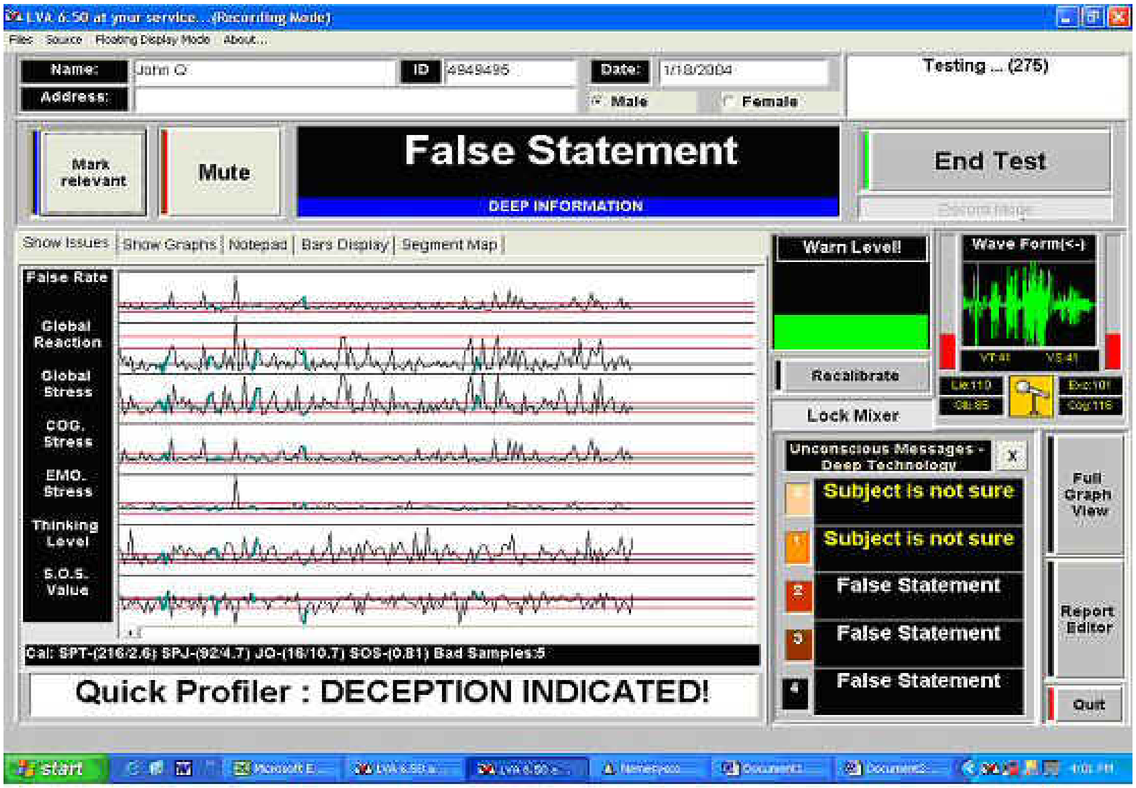
The second type of VSA (“frequency-based” VSA) relies on the same basic principles of energy-based VSA, but its focus is on how frequencies within a waveband are distributed. The LVA was developed in Israel by Amir Lieberman who applied mathematic algorithm science to voice frequencies. The manufacturer of the product reports that law enforcement, intelligence and security agencies have been using this technology for the past seven years. The LVA software claims to be based on 8,000 mathematical algorithms applied to 129 voice frequencies. Although it is reported to be "voice frequency"-based, the exact methodology applied has yet to be released for scientific review. The key difference between LVA and CVSA technology is that the former claims to pinpoint the cause of stress whereas CVSA existence of stress.

The LVA software provides a greater amount of information about the responses than energy-based VSA products – beyond a simple decision about Deception/Non-deception Indicated. For example, while an energy-based VSA will assess deception for a given response, a frequency-based VSA might assess uncertainty, anxiety, cognitive distress, avoidance/voice manipulation, or sexual arousal (in addition to deception). A screen shot of the output from one frequency-based VSA device is presented in Figure 6.

In this figure, we are presenting the Layered Voice Analysis system (LVA) marketed by V Worldwide (www.Vworldwide.com). Notice that along the left side of the screen, the output is measuring "False Rate," "Global Reaction," "Global Stress," and several other variables. It is not clear in the LVA literature exactly how these concepts are being measured. While the CVSA program claims to measure changes in micro-tremors, makers of the LVA product only report that the program operates on a number of algorithms that measure psychological patterns associated with deception. According to LVA, the mathematical equations are capable of distinguishing between stress resulting from excitement and other emotional stress.

The LVA program also claims to distinguish between concepts such as "confusion" and other cognitive or global stress. It also claims to distinguish between "acceptable levels" of stress compared stresses that is the result of deception. To detect deceptions, LVA detects tension, fear, and embarrassment while attempts are made to outwit the interviewer. LVA also claims to measure the subject's level of "thinking." According to V-Worldwide, LVA is very capable of detecting deception, boasting a nearly100% success rate when used by an expert (see Figure 7), although no studies are presented to support this claim.

Figure 6. Results of an energy-based Voice Stress Analysis.



Name: John Q
Address:
ID 4949495
Date: 1/18/2004
● Male ○ Female
Testing ... (275)

False Statement
DEEP INFORMATION

Cal: SPT-(216/2.6) SPJ-(92/4.7) JQ-(16/10.7) SOS-(0.81) Bad Samples:5

Quick Profiler : DECEPTION INDICATED!

Unconscious Messages - Deep Technology

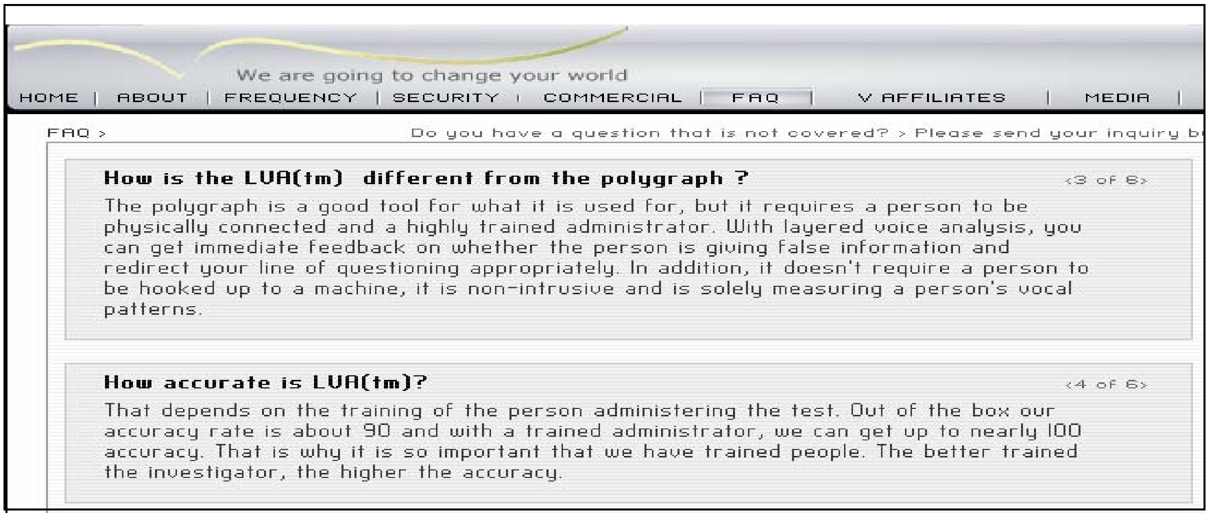
Subject is not sure
1 Subject is not sure
2 False Statement
3 False Statement
4 False Statement

Source: http://www.vsolutions.org/

In the same way that CVSA employs test formats, the LVA system provides two modes for the operation of the system. The user can choose the mode that best suits the situation depending on whether the subject is present to be interviewed or if there is a recorded statement to be analyzed. One mode of the LVA system is an initial analysis that can be further analyzed using the second mode if necessary. The format of interview, however, is conversational - with the interviewers using their training and the computer feedback to guide the interview. While

CVSA manufacturers state that question formulation is a crucial element of success, LVA does not require the interviewer to design the questioning in a particular sequence; rather the system is designed to analyze free flowing conversation.

Figure 7. Claims about LVA by V-Worldwide.


We are going to change your world
HOME | ABOUT | FREQUENCY | SECURITY | COMMERCIAL | FAQ | V AFFILIATES | MEDIA |
FAQ >
Do you have a question that is not covered? > Please send your inquiry b
<3 of 6>
<4 of 6>

How is the LVA(tm) different from the polygraph ?

The polygraph is a good tool for what it is used for, but it requires a person to be physically connected and a highly trained administrator. With layered voice analysis, you can get immediate feedback on whether the person is giving false information and redirect your line of questioning appropriately. In addition, it doesn't require a person to be hooked up to a machine, it is non-intrusive and is solely measuring a person's vocal patterns.

How accurate is LVA(tm)?

That depends on the training of the person administering the test. Out of the box our accuracy rate is about 90 and with a trained administrator, we can get up to nearly 100 accuracy. That is why it is so important that we have trained people. The better trained the investigator, the higher the accuracy.

Source: http://www.v-lva.com/newsite/site.html (Accessed December 12, 2006).

The LVA system records the initial interview as it is occurs and provides preliminary results during the on-line mode. That is, as the interview is taking place, the computer reports indications of deception (among other things). The off-line mode can then be used to take that same recorded statement from the online mode and analyze it further. To prepare the recording for the off-line mode, the examiner goes through a process of listening to the recording and sectioning the statements into segments according to relevance. This process is called segmenting. The examiner must segment the recording to isolate the relevant statements made by the subjects and minimize the background noise, irrelevant comments, and silence. The Off-Line Mode measures various aspects of the vocal patterns that can be analyzed by the examiner. A statistical analysis is also provided with this mode. A complete profile and report summarizing the overall emotional behavior of the subject is created during the Off-Line Mode. The manufacturers of LVA suggest that their product analyzes emotional states connected with memories and experiences and identifies truth and deception normally hidden by emotions. The implication is that information normally available only in the subconscious can be accessed by LVA. The LVA manufacturer claims that the off-line mode provides a psychological profile with in-depth analysis of the psychological states of the subject.

The training required for users of the LVA system is broken down into three levels of training that require the user to devote one week to each level of training. Level 1 training covers the on-line and off-line modes. At the conclusion of the course students should be able to conduct analysis using both modes. Level 2 training focuses on the investigative mode and introduction to the deep analysis on the off-line mode. Finally, Level 3 provides students with a complete understanding of the deep analysis in the off-line mode. The typical LVA user does not require all levels of training to operate the system in the field. For the purposes of this study, the novice examiners received the Level 1 training from LVA instructors and typical CVSA training course.

Summary**.** The development of voice analysis systems occurred not only as a response to the controversy surrounding the reliability of polygraph results but also in an effort to reduce the invasive nature of these examinations. The polygraph measures several different physiological reactions to stress and some consider this to be an intrusive procedure. CVSA and LVA systems do not rely on physiological measurements and subjects are not required to make physical contact with the operating systems. The only products required for examination (in addition to the software and computers) are external microphones. In addition, CVSA and LVA are both capable of conducting analyses on previously recorded statements. Therefore, both products provide mobility and covertness that the polygraph previously could not provide.

The featured software systems claim to have the following advantages in common. Each system provides the subject with a non-invasive exam and the examiner with the possibility for mobile analysis. The systems also provide immediate feedback and can be used with various types of media technologies. The manufactures of both systems intend for the use of their products to be limited to law enforcement and government agencies - neither system is available by the general public at this time. Both programs require rigorous training using a laptop before allowing users to operate their systems in the field.

As the previous descriptions of the two voice analysis systems may suggest, operation of these products in the field requires comprehensive training and practice. The manufacturers of both CVSA and LVA limit the availability of their products to law enforcement and government agencies. Similar limitations apply to the training courses provided by each of the companies. CVSA and LVA systems require users to attend training courses before obtaining the computer software. The training courses require students to pass a proficiency tests before being certified.

b. VSA Evaluations.

Unlike the polygraph, which has been rigorously tested for over 60 years (National Research Council, 2003), there have been relatively few published tests of the VSA theory and devices. When the theory itself has been tested, there has generally been solid support (Cestaro, 1996, Fuller, 1984). Smith (1977), for example, showed that acoustical indications of stress can be measured by examining filtered waveforms. More recently, researchers at the Air Force Research Laboratory (Hansen and Zhou, 1999; Haddad, Ratley, Walter, and Smith, 2002) showed that two VSA devices (Lantern and the Psychological Stress Evaluator – a precursor of

CVSA) were able to recognize stress in vocal patterns. They concluded that several features in an individual's speech pattern are different under stress and that these differences could be measured. They reported, however, that was not sufficient evidence to say that voice stress analyzers detect deception, which cannot easily be distinguished from the stress related to anger or fear (Haddad *et al*. 2002). On the other hand, recent research by Meyerhoff, Saviolakis, Koenig and Yourick (In Press) conducted a comparative text of the CVSA with blood pressure, plasma ACTH, salivary cortisol (these are medical indicators for stress). They found no relationship between these indicators and the CVSA predictions of stress.

What is less clear is the ability of the VSA devices to detect <u>deception-related</u> stress (as opposed to <u>general</u> stress). The marketing literature for both products described in this proposal suggests that the products have received scientific support. Indeed, both products have a large (and growing) following in the police community and they have received many anecdotal claims of success. Unfortunately, the scientific literature has not shown clear support. Most of the VSA device research that has been conducted consists of laboratory studies with relatively small sample sizes or with historical (off-line) tape recordings of people who were "known" to have been lying or who were under stress. Scientific studies of VSA are relatively rare (compared to studies of the polygraph). Those few studies have failed to show much support.

In fact, we were only able to find two studies (neither was peer-reviewed) that suggested that deception may be detectable using VSA (Bruck, n.d.; Hopkins, Ratley, Benincasa, and Grieco, 2005). The Hopkins *et al*. (2005) conference paper examined the off-line tape recordings of 56 criminal case interviews. They showed that the average accuracy of 5 VSA programs that evaluated the interviews for deception was about 68%. They also showed that accuracy increased somewhat when the users were more experienced. As a reference, since

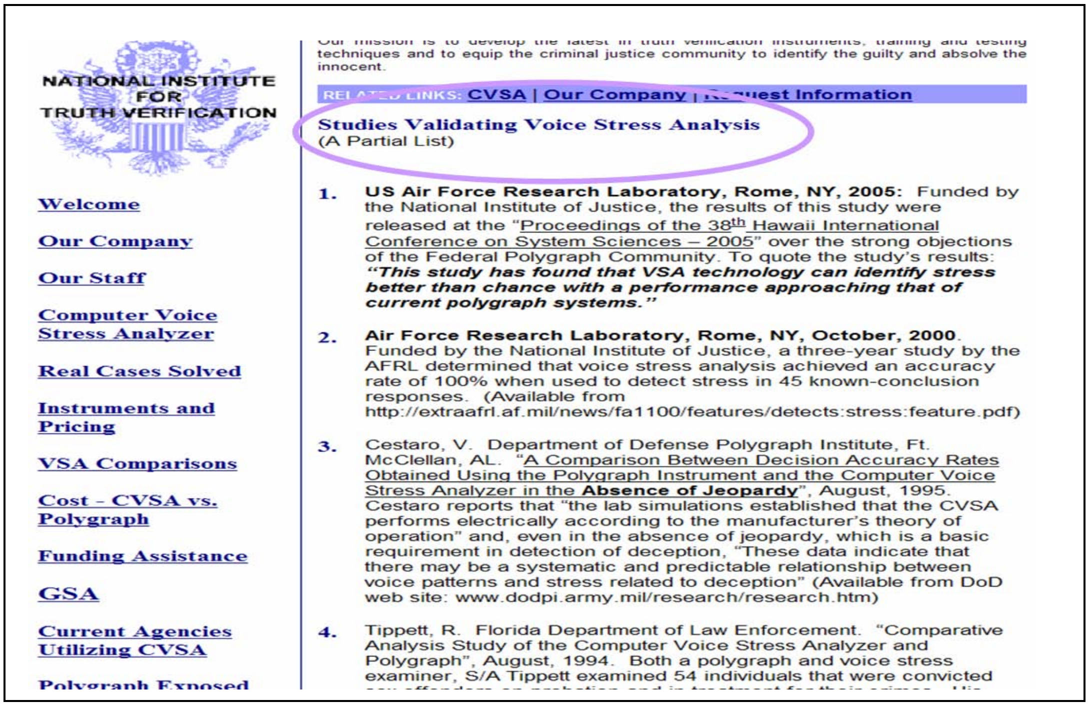
1980, research studies have shown polygraphs to be anywhere from 80-98% accurate (National Research Council, 2003).

Early researchers working in lab settings failed to find support for VSA theory or technology (Brenner, Branscomb, and Schwartz, 1979; Hollien, Geison, and Hicks, 1987; Horvath, 1978, 1979; Kubis, 1973; Lynch and Henry, 1979; Timm, 1983; Waln and Downey, 1987). Cestaro and Dollins (1996), for example, tested VSA for respondents who were concealing a number from a deck of cards. They were not able to find a relationship between deceptive answers and a variety of vocal components. Data from analysis of "pitch patterns" showed that the accuracy rate in the detection of deception was 37% - not significantly different from chance (Cestaro and Dollins, 1996; Cestaro, 1995; Janniro and Cestaro, 1996).

In another study, though, Cestaro (1996) found that the CVSA device did function "technically" as the manufacturer described – it measured changes in vocal patterns. When he compared the accuracy rates for deception detection by the polygraph and the CVSA in a mock crime scenario found that for CVSA, however, his results suggested an accuracy rate of only about 52%, again, not significantly different from chance (Cestaro, 1996). Cestaro and his colleagues state that the non-ambiguous decisions about declaring a deceptive response using CVSA were not significantly different from chance in any of their studies (Janniro and Cestaro, 1996).

Interestingly, NITV, the marketer of CVSA seems content to use these results as support for its product (see Figure 8). Indeed, some of the research papers listed on this site state clearly that research was unable to document the ability of CVSA to detect deception (apart from detecting "stress").

Figure 8. CVSA Website Providing Citations (But Not Links) to Research Studies that "Validate" CVSA.



Our mission is to develop the latest in truth verification instruments, training and testing techniques and to equip the criminal justice community to identify the guilty and absolve the innocent.

Studies Validating Voice Stress Analysis
(A Partial List)

1. **US Air Force Research Laboratory, Rome, NY, 2005:** Funded by the National Institute of Justice, the results of this study were released at the "Proceedings of the 38th Hawaii International Conference on System Sciences – 2005" over the strong objections of the Federal Polygraph Community. To quote the study's results: ***"This study has found that VSA technology can identify stress better than chance with a performance approaching that of current polygraph systems."***

2. **Air Force Research Laboratory, Rome, NY, October, 2000**. Funded by the National Institute of Justice, a three-year study by the AFRL determined that voice stress analysis achieved an accuracy rate of 100% when used to detect stress in 45 known-conclusion responses. (Available from http://extraafrl.af.mil/news/fa1100/features/detects:stress:feature.pdf)

3. Cestaro, V. Department of Defense Polygraph Institute, Ft. McClellan, AL. "A Comparison Between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer in the **Absence of Jeopardy**", August, 1995. Cestaro reports that "the lab simulations established that the CVSA performs electrically according to the manufacturer's theory of operation" and, even in the absence of jeopardy, which is a basic requirement in detection of deception, "These data indicate that there may be a systematic and predictable relationship between voice patterns and stress related to deception" (Available from DoD web site: www.dodpi.army.mil/research/research.htm)

4. Tippett, R. Florida Department of Law Enforcement. "Comparative Analysis Study of the Computer Voice Stress Analyzer and Polygraph", August, 1994. Both a polygraph and voice stress examiner, S/A Tippett examined 54 individuals that were convicted

Source: www.cvsa1.com (Accessed October 31, 2006).

A key caveat to these early studies, however, is that the deception in each of these studies was relatively minor and there was no "jeopardy" involved – those who deceived had nothing to lose by deceiving (or telling the truth, for that matter). If there is no jeopardy, then there is no stress. No stress means that the VSA technology may not have been tested appropriately (Barland, 2002). Janniro and Cestaro (1996) attempted to simulate jeopardy in a subsequent test of CVSA validity by asking half of a group of students to commit a "mock" theft while the other half had no knowledge of the crime. CVSA examiners conducted and scored the exams in accordance with NITV procedures. Results were similarly disappointing – CVSA evaluators made correct decisions about deception on only about 50% of the cases (no different from chance).

Still, it is not clear that even "simulated" stress in a lab setting causes enough stress for VSA to work. In field studies where the induced stress is more realistic (Brown, Senter, and Ryan, 2003), CVSA had deception assessment rates similar to the laboratory settings (i.e., no greater than chance). For example, the US Department of Defense Polygraph Institute (DoDPI) studied VSA technology (CVSA) in the field at Walter Reed Hospital. The DoDPI is responsible for studying new theories and technologies about deception detection. In a blind review of CVSA charts based on interviews with US soldiers, DoDPI were not able to correlate CVSA stress scores with other physiological stress scores (Meyerhoff, Saviolakis, Koenig, and Yourick, 2000).

Since CVSA supporters suggest that laboratory studies are not an adequate way to test the product, some researchers have tried to test CVSA "in the field" with limited success. Palmatier (2000), for example, examined 50 cases where suspects were confirmed through investigation as being deceptive. These cases were compared to 50 cases where the suspects were "confirmed" to have been "truthful." He tested CVSA using voice data recordings that were made during polygraph examinations. The results suggested that CVSA examiners were not able to distinguish those who were being deceptive from those who were telling the truth in a law enforcement setting (see also Palmatier 1999; n.d.).

In spite of this lack of scientific evidence, the use of VSA technology continues to grow and to receive high profile attention. CVSA claims, for example, that over 1,400 police agencies use their device. In the aftermath of 9/11, interest in VSA grew dramatically, especially concerning airport security. One indication of the growing presence of VSA in the law enforcement, security, and business world is the genesis of an interesting counter-industry – how to fool a VSA examination (e.g., http://www.passapolygraph.com/). As the name of this website suggests, most of these websites were originally designed to help users pass polygraph examinations. Now, customers coming to these sites can pay $35-$80 for a manual that will help them overcome the VSA assessment. Clearly, there is a need for a field-based experiment that incorporates jeopardy and the ability to compare assessed deception with known deception. That is the purpose of this study.

3. Research design and Analysis Strategy.

According to voice stress analysis proponents, previous research projects that fail to show support for VSA technology are often associated with the polygraphy community. Indeed, there is much literature outlining the anxiety and mistrust between both deception detection camps. The CVSA website, for example, contains a link that points to a blog entitled "Pentagon Obstructionism" while another hyperlink was titled "DoD Conspiracy Exposed." The resulting links takes the reader to a 2005 article complaining about the decision by the US Department of Defense to bar the use of CVSA by the military because of the lack of evidence showing that it is valid. The irony of the two links is that they are on the same page that proclaims that CVSA has been classified as a "Restricted Crime Control Technology" by the US Government –suggesting that the government has subtly "endorsed" CVSA.

On the other hand, state polygraphy associations have successfully lobbied some legislatures to allow only trained polygraphers to use mechanical devices to detect deception – effectively outlawing VSA in those states (e.g., Kansas, Texas, and Virginia). Several websites (e.g. http://www.voicestress.org/) have been developed by people affiliated with the American Polygraph Association to showcase problems – such as false confessions – that have been associated with voice stress analysis.

That said, VSA advocates may be correct in their assessment that previous tests of VSA products have some methodological problems. The most common complaint is that the programs are popular because they are successful in the field so they should be tested in the field. There is much anecdotal evidence from police officers who have successfully used VSA to obtain a confession from a suspect. VSA proponents suggest that all validity tests of the program should be conducted in real world settings. On a related note, critics of VSA research complain that mock deception and fake scenarios fail to create the necessary amount of "jeopardy" that will stimulate a "deception detection" response. If the subject does not worry about getting caught (since the experiment has no real consequences), then it will be more difficult for the software to work. Similarly, if the subject is only "pretending" to lie, then the software will not be able to detect the deception.

The goal of this study was to create a methodology that would address these criticisms. Our solution was to test the programs in a criminal justice setting where police interviews commonly occur and to ask about relevant criminal behavior that the respondents would be intrinsically motivated to hide. Thus, we interviewed arrestees in the Oklahoma County jail about their recent illicit drug use. Answers by the respondents were compared to the results of a urinalysis to determine the extent to which they were being deceptive. Then, their "actual deceptiveness" was compared to the extent to which deception was indicated by the VSA programs.

The methodology and sampling protocol for this study were derived from the preexisting methodology and sampling techniques employed in the NIJ-funded Arrestee Drug Abuse Monitoring (ADAM) program that operated in Oklahoma County from 2000 to 2004 (for more information, see ADAM, 2003). The rationale behind the adoption of the ADAM methodology for this project was based on many factors. First, the existing ADAM site research team and jail staff and administration were immediately available and cooperative. Second, the population targeted by the ADAM protocol matched very closely the "real world" emphasis we intended to match. Third, the use of urine specimens provided a key method of getting "grounded truth," a necessary requirement to validate voice analysis systems.

The sample for this study was collected at the Oklahoma County Detention Center in Oklahoma City, Oklahoma during the months of February and March 2006. The voluntary and confidential interviews were conducted only with arrestees who had been in the detention facility for fewer than 48 hours. Using the ADAM probability-based sampling plan, the total number of men arrested within Oklahoma County during this period (regardless of charge) composed the sampling frame for the VSA study. This method of sampling was devised to allow the selection of arrestees during the time of day with the highest volume of arrests and to allow random selection of arrestees who were booked during the remaining hours in that 24 hour day.

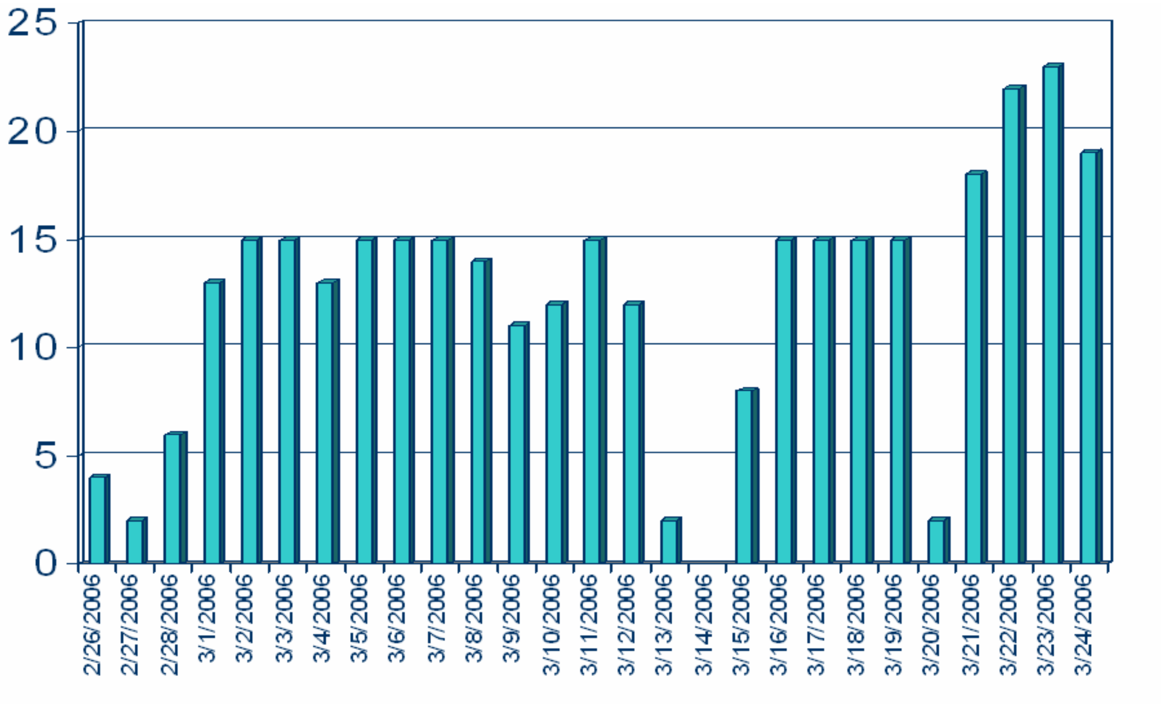
The sample selected during the high volume time was referred to as “flow” and the arrestees from the remaining hours of the day were referred to as “stock.” These terms were traditionally used because interviews in the ADAM project were conducted during the peak booking time (flow). Thus, flow samples were composed of people who “flowed” into the jail during the 8-hour period that the research team was collecting data. The “stock” referred to the sample of people who were booked during the 16-hour period that the research team was not collecting data - during the times of the day when booking rates were normally low. The sample was drawn proportionately from both stock and flow throughout each data collection period to reflect the distribution of arrests each day. The site had previously been given a target number of interviews it was expected to complete each day, so we used that figure. In Oklahoma City, the sampling plan required the selection of seven “stock males” and a minimum of five “flow” males per day.

The original ADAM program participants included both adult males and females. While we originally intended to collect data from both males and females in the VSA study as well, circumstances at the jail precluded this possibility. We were grateful to have been given a relatively private room to conduct the VSA interviews (as opposed to conducting the interviews in the hallway as was done in the original ADAM project) but the small space only allowed one interviewer to work at a time. Since there are about 5 times as many male arrestees booked each day, we decided to maximize our sample size by only collecting data from males. There are few indications in the literature that the CVSA or LVA software systems work differently for male and female subjects. Future studies would do well to consider collecting data from female arrestees as well.

The data collection for this study was conducted by former ADAM staff of the Oklahoma City site. The staff included coordinators and interviewers who were very experienced with the original sampling procedure and with interviewing a difficult population (recent arrestees). In addition, the interviewers were also trained in the operation of the CVSA and LVA systems. Before data collection begins, the research team traveled to the training headquarters for LVA (Waupun, WI) and CVSA (Palm Beach, FL) to receive the 6-day training program offered by each vendor. This important step should not be over-looked. The research team, on the one hand, represents novices who would be using the software with relatively little training. This allowed us to test the extent to which the devices can be used by relatively unskilled people. On the other hand, it is vital that each of the VSAs be tested as they were designed to be used. Thus, the research team endeavored to incorporate all aspects of the particular devices into the research protocol. During training, the team worked closely with the VSA trainers to make sure that the research protocol matched the use protocol of the VSA devices. In both cases, the training team knew the purpose of our participation in their training program. The team participated in CVSA training November 14-19, 2005 and LVA training February 6-13, 2006. Following training, the researchers practiced use of both computer programs in Oklahoma until data collection began.

On February 26, the research team entered the jail for the first time to practice using the equipment in the jail facility. Our original plan was to collect data using both LVA and CVSA at the same time but we discovered that our data collection room (a small office just outside of booking) was too small to allow for two interviews to be conducted at the same time. As a result, we collected data using the CVSA program for the first 12 days and then using the LVA program for the second 12 days (see Figure 9).

Figure 9. Number of urine samples collected each day during course of the project.



We successfully tested our protocol again on February 27 and then began data collection in earnest on February 28 and concluded March 24, 2006. We only include data collected from February 28 through March 24 in our analyses. We took one day off between the CVSA data collection (February 28-March 13, 2006) and the LVA data collection (March 15-March 24, 2006). Only two surveys were collected on March 20 because one of the interviewers became ill.

Just as was required with the ADAM protocol, the data collection team had access to booking data to help select the sample and collect basic data (e.g., demographics and charge information) for each respondent. The staff was provided access to an interview room adjacent to the booking area to complete the interviews. Staff also coordinated security personnel to escort and safeguard the interviewers while in the facility. Security staff were briefed about the study and trained to not interfere with the project by disclosing too much about the project to the arrestees or by standing too close to the arrestees during the interview (see Appendix A).

The original ADAM instruments were used as the basis for the data collection in this study. We created a version of the original "face sheet" that containing official data about arrestee (e.g. age, race, instant offense – see Appendix B). We also used the same "recent" drug questions that were asked in the original ADAM survey (see Appendices C and D). The process for completing the interviews differed slightly depending on the type of voice stress analysis system being tested. The CVSA program required that direct questions that resulted in yes or no answers be asked. The LVA program, on the other hand, required that the questions (and the responses) be more open-ended and "conversational." We would have preferred to run both protocols the same but the LVA staff recommended using a less formal approach because of the nature of their software. In both cases, the question formulation and test format used to coincide with each respective VSA system was pre-approved by an official representing CVSA (NITV) and LVA (V-Worldwide).

In addition to recordings made by the computers (graphically by CVSA and audibly by LVA), the interviewers also coded each subject's responses to each on a paper survey. We had originally planned to tape record the interviews but in a last minute decision, the jail administration would not allow it. After the first few days of data collection with CVSA, the research team sent copies of the output files to the CVSA expert to make certain that the data was being collected correctly. Other than some problems with over-modulation, we were assured that all was well and we proceeded to collect the data. Because there is so much processing work that must be done before the LVA output could be analyzed, we were not able to send out copies of the output during data collection. Fortunately, with few exceptions, data on the majority of the LVA respondents were collected successfully.

The interviews were conducted under the terms of ODMHSAS IRB requirements. Before each interview began, the research team explained the interview procedure with the arrestee. All special instructions were discussed with the arrestee prior to the start of the survey. The subjects were informed that the interview could not be used to harm them in any way during their stay in the facility or at a later date. The arrestees were also told that their participation was voluntary (see informed consent form in Appendix E). They were told that their responses would be recorded to test the how well voice stress analysis determines stress. We did not specifically state that the software was being used to test for indications of deception, but our sense was that most inmates knew that we were testing for deception. Many commented about the computer and procedures, for example, and referred to the procedure as a “lie detector test.” Each interview took between 15-30 minutes to complete.

The arrestees were asked to respond to questions about recent drug use: marijuana use in the previous 30 days and cocaine, heroin, methamphetamine, and PCP use in the previous 72 hours. Since marijuana metabolizes at a much slower rate than the other four drugs, we used a

30-day "window" to assess "recent" drug use. Therefore, if the arrestees "honestly" reported no recent drug use in the respective time frames, then their urine test should be negative.

We recognize the potentially related validity threats of "failed memory" (i.e., individuals forgetting that they smoked marijuana in the past 30 days) and "telescoping" (i.e., individuals thinking that they had smoked marijuana in the past 30 days when it was actually more than 30 days). That said, we believe that both possibilities were rather unlikely. To counter this possibility, however, we shared the questions with the respondent in advance of the testing so that they would have time to think about their response.

After the completion of the interview, the subjects were asked to complete the data collection process by supplying urine specimens. If the subject agreed, he was escorted to a private restroom where he was asked to fill a specimen bottle with his urine. The interviewers were trained to recognize and respond to respondents who returned specimen bottles returned with water or diluted urine. All urine samples contained sufficient urine to be able to perform a drug test. Those arrestees who completed the interview and provided a urine sample were given a candy bar as an incentive.

As with the original ADAM project, we generated a set of three uniquely numbered labels for each arrestee. One label was placed on the face sheet, a corresponding label was placed on the survey, and the third label was placed on the urine bottle. This ID number was also typed into the computer program for each arrestee's VSA data. At the completion of each data collection day, all information sheets, interview answer sheets, urine samples, and computer files were removed from the facility. Every two days, the urine bottles were taken to Compliance Resource Group (CRG) Laboratories for testing. The drug testing company did not perform a gas chromatography-mass spectrometry (GC/MS) test (thought to be the gold standard) to confirm the positive findings. The cost for each test would have been $25 (over $5,000 for the project). Instead, the company used the EMIT immunoassay screening test. The following cutoff levels were used by CRG per DHHS guidelines: Amphetamine 1000 ng/ml, Cocaine 300 ng/ml, THC (marijuana) 50 ng/ml, PCP 25 ng/ml, and Opiates 2000 ng/ml.

Any sample that tested positive for amphetamine was subsequently screened for methamphetamine as well. Excel spreadsheets that reported the results were emailed to the project director at the end of each week. Data from the surveys and the face sheets were entered into an SPSS dataset. Results from the urinalyses were also merged into the survey data set (matching on ID number).

The VSA data collected using each of the software systems in this study were sent to certified examiners from CVSA and LVA for their analysis. These examiners were referred to as the "expert" examiners in this study. The expert analysts were individuals with several years of law enforcement experience. They were also trainers for each of the respective VSA programs, so they were intimately familiar with the software and its output. In comparison, the "novice" examiners each had 5 years of survey interview experience but no law enforcement (i.e., interrogation) experience.

When the project was completed, the interviewers ("novice" examiners) and the expert examiners began to assess the computer output for each arrestee to determine if deception was indicated (DI) or not (NDI). Both sets of examiners analyzed the output for each subject to determine if software had detected that the subject was being deceptive about recent drug use. In addition, the "novice" analysts also entered into the data a prediction of whether the arrestee would test positive for a particular drug. Both sets of assessments were done completely blinded. That is, neither the novice nor the expert examiners were privy to the results of the urine samples or the results of the other examiners. To this day, neither the novices nor the experts know which of the arrestees tested positive and which did not. To guard against the potential impact of non-VSA related cues that might tip-off the novice analysts about potential deception, the interpretation analysis did not begin until two months after the end of data collection. Therefore, the novices were not able to associate a given ID number in the data with a specific individual who may have reported no drug use but appeared to be intoxicated.

When all the data sources were received by the project director, the data were all merged into one large master SPSS file. As a result, the master data set contains (1) demographic information obtained from the official booking records, (2) responses to survey questions about recent drug use, (3) the results of a urinalysis test on the five drugs, (4) variables recording "deception" or "no deception" on each of the drugs, and (5) decisions by novice and expert analysts regarding the indication of deception. Analysis of the VSA deception data was completed by the end of August and data cleaning was completed in September.

There are two key ingredients that are required to perform a proper test of a VSA device. First, the researchers need to know if the response is truly deceptive. This is referred to in the literature as "ground truth." Second, there must be "jeopardy" involved in the response. The respondent must have something to hide and a reason to hide it. Collecting data according to the procedures described in this section addresses both issues. We collected survey and urinalysis data from a group of arrestees who are in the midst of being booked into a county jail. The survey asked the respondents if they had "recently" used each of five different drugs. The urinalysis ("ground truth") will be used as the criterion to know if the respondent was being deceptive during the survey.

The more difficult question is about jeopardy. Previous researchers who have conducted this kind of research have been justly critiqued because they have not been able to simulate jeopardy (Barland, 2002). We believe that jeopardy exists in this proposed methodology because the study participants are a "wary" population (arrestees) in a high-risk environment (a jail). Previous arrestee studies suggest that respondents are commonly deceptive about recent drug use. Comparisons between the urine sample and the self-report of recent drug use in the 2000-2004 ADAM data, for example, suggest that a relatively high proportion of the respondents lied about recent drug use, even though they knew that a urine test would follow the survey and they had been guaranteed confidentiality (see Fendrich and Xu, 1994; Hser, 1997; Lu, Taylor, and Riley, 2000; Yacoubian, 2000).

Mieczkowski, Barzelay, Gropper, and Wish (1991), for example, compared cocaine urinalysis results with self-reports from a sample of arrestees. Among those arrestees who tested positive for cocaine, three out of four denied using cocaine use in the previous 48 hours. Another study (Harrison, 1995) compared self-reported drug use to drug tests among arrestees in 1989 and found that only about one half of those who tested positive had reported using the drugs within the previous three days. In general, the more "serious" the drug is, the more likely arrestees are to lie about recent use. Clearly, arrestees who were interviewed in a jail setting engaged in a great deal of deception regarding recent drug use. This suggests that they are experiencing jeopardy.

b. Analysis Strategy.

After data collection had been completed, we classified all relevant responses according to the matrix in Table 3. The first step for completion of the matrix was to examine the VSA data to determine if the analysts had noted that the relevant responses for each arrestee indicated deception or not. This was done for both novice and expert analysts.

While the devices provide almost instantaneous deception feedback on the responses, there is a great deal of interpretation and preparation that needs to be done to isolate the responses by the arrestee from the questions by the interviewer before converting the data into an SPSS data set. This work takes about one hour per survey. These assessments were made "blinded." That is, the individuals making the assessment about deception did not know if the response was actually deceptive.

Table 3. Matrix comparing correct and incorrect classification of response.

VSA Classification	**Condition 1** **Deceptive Response** Used and said "no"	**Condition 2** **Non Deceptive Response** Used and said "yes"	**Condition 3** **Deceptive Response** No use and said "yes"	**Condition 4** **Non Deceptive Response** No use and said "no"
1. Deception Indicated	A. Correct - True Positive	B. Incorrect - False Negative	C. Correct - True Positive	D. Incorrect - False Negative
2. No Deception Indicated	E. Incorrect - False Positive	F. Correct - True Negative	G. Incorrect - False Positive	H. Correct - True Negative

The second step for matrix completion was to examine the merged survey and urine data to determine if the respondents were being deceptive about recent drug use. Once we completed these two steps, we were able to complete the matrix by determining if each arrestee was being deceptive and if the software detected the deception. For example, if an arrestee tested positive for cocaine but reported not having used the drug in the previous 3 days, he was classified in Condition 1 (Deception Indicated: Used but said "no"). If an arrestee tested positive for cocaine and reported having used the drug in the previous 3 days, he was classified in the second condition (No Deception Indicated: Used and said “yes”). In rare instances, a respondent tested negative for a drug but reported having used the drug. Such respondents were classified in Condition 3 (Deception Indicated: No use but said yes”). Finally, if a person tested negative for cocaine and reported not having used the drug in the previous 3 days (non-deceptive), he was classified in Condition 4 (No Deception: No use and said “no”). This was done for each of the five drugs. If VSA works as predicted, then we would expect to find a statistically greater percentage of respondents in cells A, F, C, and H than in cells E, B, G, D (respectively).

To simplify the matrices described in Table 3, however, the research team used the strategy developed by the National Research Council to assess the relationship between the *sensitivity* and *specificity* of the VSA devices (National Research Council, 2003). This is a popular technique that is used to test the validity of the polygraph. By *sensitivity*, we refer to the ability to correctly identify deception. Sensitivity is expressed as a percentage, where we compare the number of correctly identified deceivers divided by the number of deceivers in the sample (multiplied by 100). A perfectly sensitive device would correctly detect 100% of all deceptions (there are no false negatives). Say we have a sample of 200 people and 100 of them were being deceptive. If our device correctly identified 75 as being deceptive, then we would have a sensitivity score of 75%. In comparison, a device that had a sensitivity score of 40% is less sensitive than the device with a sensitivity score of 75% (it only correctly identified 40% of those who were being deceptive).

By *specificity*, we mean the percentage of non-deceptive respondents who are correctly identified as non-deceptive. A perfectly specific device would correctly detect 100% of all non-deceptions (there are no false positives). Say we have a sample of 200 people and 100 of them were being non-deceptive. If our device correctly identified 55 of those individuals as being non-deceptive, then we would have a specificity rate of 55%. This suggests that 45% of the non-deceptive respondents were incorrectly identified as being deceptive. Thus, the device is not very specific. That is, while the device might be measuring some sort of stress, it is not measuring deception-related stress.

With these concepts in mind, we can simplify the matrix in Table 3 with a 2x2 table as shown in Table 4. In this new matrix, we can calculate four key indices. The first two indices are the most important since they suggest the extent to which the VSA programs work as expected. *Sensitivity* is the proportion of all the "truly deceptive" cases (A+C) that are correctly identified as "deceptive" (A). This value (A/[A+C]) is also known as the conditional value of the true-positive proportion. The higher the sensitivity, the more likely that a deceptive response will be indicated as deceptive. *Specificity* is the proportion of all the "truly non-deceptive" cases (B+D) that are correctly identified as "non-deceptive" (D). This value (D/[B+D]) is also known as the conditional value of the true-negative proportion. The higher the sensitivity, the less likely that a non-deceptive response will be indicated as deceptive. It is important to compare sensitivity and specificity at the same time since an overly sensitive instrument (for example, one that indicates all responses as deceptive) is not very useful. Our goal is to assess the joint sensitivity and the specificity of the two VSA programs.

At the same time, we can also assess two other values using the matrix in Table 4. *False negative probability* is the proportion of all the truly deceptive cases (A+C) that are incorrectly identified as non-deceptive (C). This value (C/[A+C]) is also known as the conditional value of a false-negative proportion. *False positive probability* is the proportion of truly non-deceptive cases (B+D) that are incorrectly identified as deceptive (B). This value (B/[B+D]) is also known as the conditional value of the false-negative proportion. Once these numbers are calculated, we can then calculate a *False Positive Index* (FPI), which is the ratio of "false positives" to "true positives" (B/A). The higher this value is, the more likely it is that the devise will incorrectly implicate a respondent.

Table 4. Matrix comparing sensitivity and specificity.

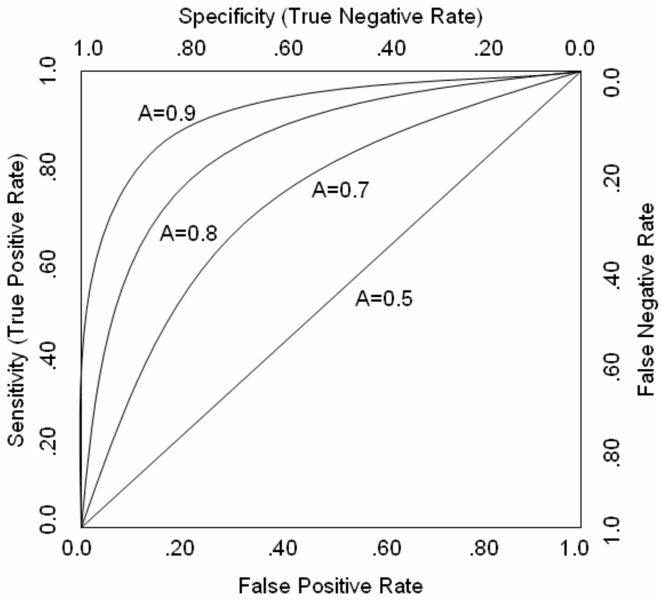
VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	A. True Positive	B. False Positive	A + B
2. No Deception Indicated	C. False Negative	D. True Negative	C + D
Total Sample	A + C	B + D	A+B+C+D

Source: National Research Council, 2003

Once sensitivity and specificity are calculated, then one is able to plot both scores on a Receiver Operating Characteristics (ROC) chart as shown in Figure 10 (National Research Council, 2003). We show the ROC chart for illustrative purposes only to show the joint dependency of sensitivity and specificity. The ROC chart helps to show the importance of considering both the sensitivity and the specificity of indications of deception.

The diagonal line represents an "accuracy index" of 0.50 (no better than chance). The curved lines represent accuracy indices of 0.70 (correct 70 percent of the time), 0.80, and 0.90. The upper left hand corner of the chart represents a test that is perfectly accurate (specificity and sensitivity are both equal to 1.0). An instrument with a sensitivity score of 0.40 and a specificity score of 0.60 (fairly specific but not very sensitive) would fall on the diagonal line (representing about 50% accuracy) - the device would be no more accurate than just guessing. An instrument with a specificity score of 0.80 and a sensitivity score of 0.60 (fairly sensitive and fairly specific) would be better than 70% accurate.

Figure 10. Plotting sensitivity and specificity on a ROC chart.



4. Findings

We were concerned that the use of the laptop and the specter of the voice stress analysis would make arrestees less willing to participate in the study than they were during the ADAM study. In the past, the research team had an average of about 90% of arrestees agree to the interview and about 95% of those who were interviewed agreed to supply a urine sample as well. For the VSA project, we randomly selected 356 male arrestees into our sample. Of those, 331 (93%) agreed to an interview and 319 (90%) of the interviewed arrestees agreed to provide a urine sample. This refusal rate is very similar to the original ADAM project except that a greater proportion of interviewed arrestees (96%) agreed to provide a urine sample. The refusal rate did not vary significantly for each VSA protocol. Indeed, there are no significant differences between the CVSA and LVA sample for age, percent who were deceptive about any drugs, or percent who tested positive for any drug (see Appendix H).

In Table 5, we show the first (most serious) charges of the arrestees selected into the sample. Unfortunately, the booking sheet that we used to enter much of the data was often missing the instant offense, so about one-third of the charge information was missing. These numbers compare favorably to the ADAM 2003 data, however, where 15.7% of the lead offenses were violent crimes, 27.9% of the lead offenses were substance-related crimes, and 13.3% were property offenses. This suggests that many of the “missing” cases in the VSA data would likely have fallen in the “Other” category – it is very likely that the missing cases would have been classified as an unspecified warrant in the VSA data (a large miscellaneous category in the ADAM data). That said, it appears that our VSA sample looks very similar to the earlier ADAM samples.

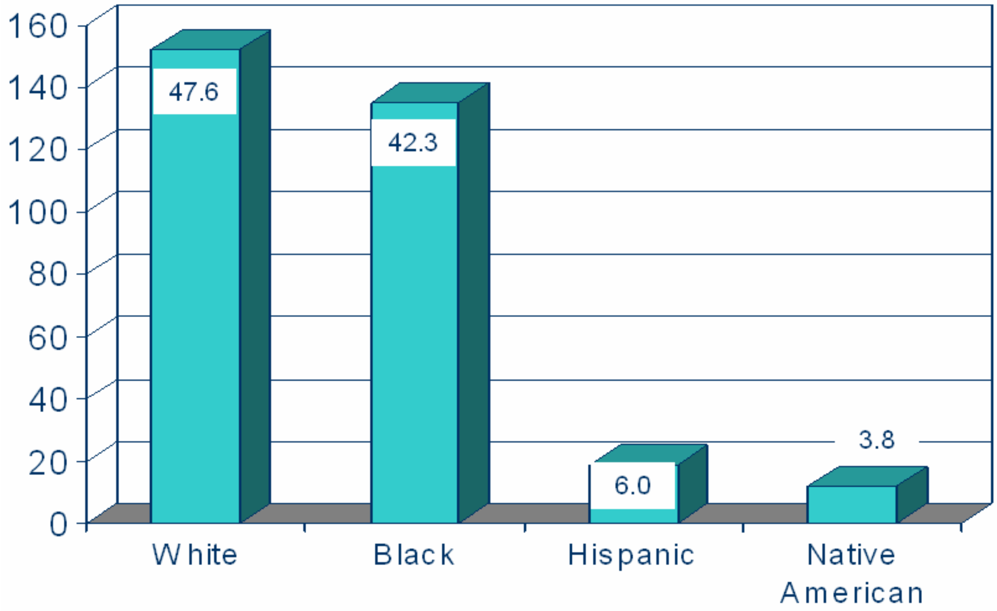
In Figure 11, we show the race/ethnicity distribution for the VSA sample for the sample. We note that we had almost as many African American arrestees (42.3%) as white arrestees (47.6%) and that there were about 6% Hispanic and 4% Native American arrestees in the sample.

Table 5. First listed charge on facesheet.

First Charge	**Number of Arrestees**	**Percent**
1. Violent Crime	**45**	**14.1**
Aggravated assault	6	1.9
Manslaughter – negligent	1	.3
Murder/homicide	1	.3
Robbery	2	.6
Weapons	7	2.2
Domestic violence	23	7.1
Child abuse	2	.6
Sex offense	3	.9
2. Substance Related	**90**	**28.2**
DWI/DUI	25	7.8
Drug possession	43	13.7
Drug sale	21	6.5
Possession of alcohol	1	.3
3. Property	**51**	**16.0**
Arson	1	.3
Burglary	11	3.4
Forgery	11	3.4
Fraud	9	2.8
Larceny/theft	6	1.9
Stolen property	11	3.7
Stolen vehicle	2	.6
4. Other	**23**	**7.2**
Flight/escape	5	1.6
Obstruction of justice	2	.6
Other	4	1.2
Traffic-related	12	3.7
5. Missing	**110**	**34.5**
Total	**319**	**100**

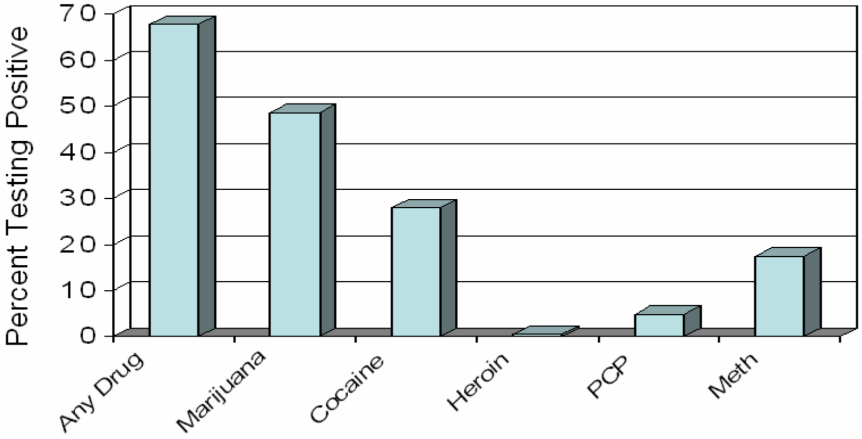
These figures compare favorably with the 2003 ADAM data except that the VSA sample has relatively more African American arrestees and fewer Native American arrestees. The average age of the respondents was 33.8 years (SD=12.3) with the youngest being 18 and the oldest being 78. About 53% of the sample was composed of arrestees who were in the "stock" population (n=172) while 47% came from the flow population (n=150). The minimum number of hours that an arrestee had been incarcerated was one hour (n=33) and the longest time a person had been incarcerated was 38 hours (n=1). The average number of hours that arrestees had spent in jail before being interviewed was 9.3 hours (SD=7.69).

Figure 11. Race/ethnicity distribution for the VSA sample (n=319)



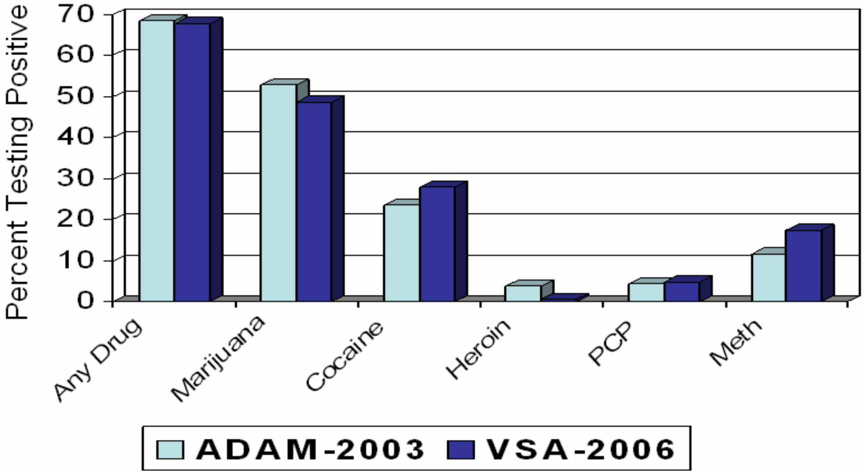
Of the 319 arrestees who provided urine samples, many tested positive for at least one drug. As shown in Figure 12, 67.4% tested positive for at least one of the drugs, while 51.7% tested positive for marijuana, 27.9% tested positive for cocaine, 0.6% tested positive for opiates, 5% tested positive for PCP, and 17.9% tested positive for methamphetamine.

Figure 12. Percent of VSA arrestees testing positive for drugs (n=319).



In Figure 13, we show comparison data from the 2003 ADAM project to compare similarities in the drug use patterns. The VSA data contains slightly more recent cocaine and methamphetamine users and fewer opiate users, but overall, the data look very similar for the two samples.

Figure 13. Comparing percent of ADAM (n=178) and VSA (n=319) arrestees testing positive for drugs.



Having described the VSA data and its favorable comparison to the 2003 ADAM data, we now turn to our analysis of the validity and the reliability of CVSA and LVA software for detecting deception. In the following sections, we present data that show the extent to which the VSA programs are able to detect deception (paying special attention to the distinction between sensitivity, specificity, and the false positive index). Finally, we address an interesting methodological by-product of this research protocol: the bogus pipeline effect.

a. Evaluating Validity of VSA Software for Deception Detection

In this section, we compare the deceptiveness of the answers given by the respondents to the indication of deception made by each computer program. The first step was to determine if the VSA programs indicated if the respondents were being deceptive. The research team recorded indications of deception (DI) or no deception (NDI) for each of the five relevant drug questions for each respondent using each VSA program. After the research team recorded each instrument's prediction about deception, we used the survey and urinalysis report to determine if the respondent had answered truthfully or had been deceptive. Each respondent was asked if he had "recently" used marijuana, cocaine, heroin, PCP, and methamphetamine. If he said he had not recently used the drug but tested positive for it, then his response was coded as "Deceptive." Likewise, if he said he had used the drug but then tested negative for the drug, his response was coded as "Deceptive." If the respondent said that he had not recently used the drug and tested negative for the drug, then his response was coded as "Non Deceptive." Finally, if he said he had recently used the drug and he tested positive for the drug, then his response was coded as "Non Deceptive."

In the following sets of tables, we assess the ability of the CVSA and LVA instruments to detect deception about recent drug use among an arrestee population. For each type of drug, we test the validity among only those who tested positive for the drug (the user population) and then among all the respondents. Then, we show the results separately for each VSA instrument.

In Table 6, we show the comparisons between deceptive responses and indications of deception for marijuana among those who tested positive for marijuana. Among the marijuana users, 11.2% provided a deceptive response (i.e., they said that they had not used marijuana in the past 30 days) but almost 9 in 10 respondents told the truth about recent use. The good news is that the sensitivity score is very high. Among the 135 respondents who were not being deceptive, 131 (97%) were not classified as indicating deception by the VSA programs. The False Positive Index (the ratio of <u>incorrect</u> indications of deception to <u>correct</u> indications of deception) is 2.0. Unfortunately, the VSA programs only indicated deception for two of the 17 respondents who were deceptive about recent marijuana use. This results in a sensitivity score of 11.8%.

Table 6. Interpretation of deception about marijuana by marijuana users.

VSA Classification	**Deceptive Response**	**Non Deceptive Response**	**Total**
1. Deception Indicated	2 11.8%	4 3.0%	6
2. No Deception Indicated	15 88.2%	131 97.0%	146
Total Sample	17 11.2%	135 88.8%	152

Chi Square = 3.085 (*p* = 0.079)
Sensitivity = 11.8%
Specificity = 97.3%
False Positive Index = 2.0
Correctly Classified: 87.5%

The problem with assessing the VSA programs while only examining the "users" is that all the respondents in the table are "guilty" of the act. It is more meaningful, therefore, to examine the results for all respondents in the study (including those who have not used the drugs). We present those data in Table 7.

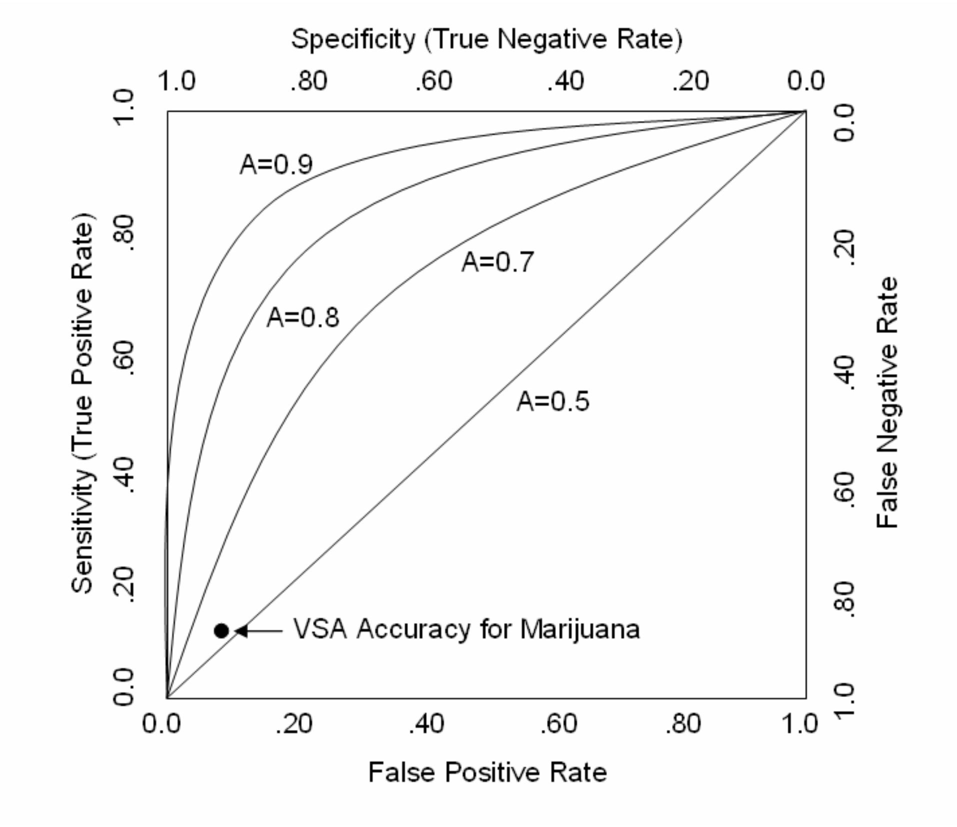
Table 7. Interpretation of deception about marijuana by all respondents.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	4 7.8%	22 8.3%	26
2. No Deception Indicated	47 92.2%	234 91.7%	290
Total Sample	51 16.1%	265 83.9%	316

Chi Square = 0.012 (*p* = 0.913)
Sensitivity = 7.8%
Specificity = 88.3%
False Positive Index = 5.5
Correctly Classified: 75.3%

There were 316 respondents for whom we have data on the deceptiveness of their response and the extent to which the VSA instruments indicated deception. Only about three quarters of the cases were correctly classified as deceptive or not deceptive. Among the 316 subjects for whom we have data, 51 (16.1%) provided a deceptive response while 83.9% were not deceptive. Only four of the respondents who were deceptive about marijuana use indicated deception using the VSA programs. This is an extremely low sensitivity rate (7.8%). On the other hand, 22 (8.3%) of those who were not being deceptive were identified by the VSA programs as being deceptive. This results in a False Positive Index of 5.5 (for every one true positive, we observe 5.5 false positives). This suggests that the VSA programs are not very useful for determining deception about recent marijuana use among an arrestee population. Indeed, if we plot the sensitivity and specificity of the results in a ROC chart, we see that the accuracy of the programs is only just above 50% - just above the diagonal (see Figure 14).

Figure 14. Plotting sensitivity and specificity for marijuana on a ROC chart.



In order to be certain that variations among the two VSA programs are not masking their ability accurately predict recent marijuana use, we conducted separate analyses for each VSA instrument. In Table 8, we show the accuracy results for CVSA (n=167) and in Table 9 we show the results for LVA (n=149). The data suggest that both programs do not predict recent marijuana use very accurately although LVA seems to have a relatively higher accuracy rate than CVSA.

Table 8. Interpretation of deception about marijuana by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 3.6%	12 8.6%	13
2. No Deception Indicated	27 96.4%	127 91.4%	154
Total Sample	28 16.8%	139 83.2%	167

Chi Square = 0.832 (*p* = 0.362)
Sensitivity = 3.6%
Specificity = 91.4%
False Positive Index = 12
Correctly Classified: 76.7%

Table 9. Interpretation of deception about marijuana by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	3 13.0%	10 7.9%	13
2. No Deception Indicated	20 87.0%	116 92.1%	136
Total Sample	23 15.4%	126 84.6%	149

Chi Square = 0.637 (*p* = 0.425)
Sensitivity = 13.0%
Specificity = 92.1%
False Positive Index = 3.3
Correctly Classified: 79.9%

The CVSA instrument, for example, only correctly identified one of the 28 deceptive responses about recent marijuana use while it indicated that 12 of the non-deceptive respondents were being deceptive. This resulted in a False Positive Index of 12. The LVA instrument, on the other hand, correctly indicated that three of the 23 deceptive respondents were being deceptive - a sensitivity rate of 13% and it was more specific than the CVSA program (only 10 of 126 non-deceptive respondents were identified as deceptive). Still, the relatively high False Positive Index (3.3) suggests that even LVA is does not very accurately predict deception about recent marijuana use.

There are three possible explanations about the accuracy problems shown in Tables 7-9. The most obvious, of course, is that the programs are not accurate. That is, they do not accurately distinguish between those who are being deceptive or those who are telling the truth. A second explanation, however, has to do with a problem we discussed earlier – telescoping. Since the time window for marijuana is 30 days, it may be that many respondents had used marijuana more than 30 days previous to the data collection and, as a result, may have been mistaken (as opposed to being deceptive). In fact, many of those who were coded as being “deceptive” did not test positive for marijuana. Although one might argue that VSA programs ought to be able to identify these deceptive responses as well, we are not certain if we are observing true deception (purposeful lying) or simple forgetfulness. To take both possibilities into account, we re-analyzed the data by removing those who were coded as deceptive only because they had reported drug use but did not test positive. The results for the CVSA program are presented in Table 10 and for the LVA program in Table 11.

The data suggest that CVSA is more sensitive (11.1% compared to 3.6%) when we only consider those who were lying about having recently used marijuana (as opposed to including those who were "deceptive" as defined by reporting recent marijuana use but not testing positive for marijuana.

Table 10. Interpretation of deception about marijuana by all respondents for CVSA not including individuals who said "yes" but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 11.1%	12 8.6%	13
2. No Deception Indicated	8 88.9%	127 91.4%	135
Total Sample	9 6.1%	139 93.9%	148

Chi Square = 0.065 (*p* = 0.799)
Sensitivity = 11.1%
Specificity = 91.4%
False Positive Index = 12
Correctly Classified: 86.5%

Table 11. Interpretation of deception about marijuana by all respondents for LVA not including individuals who said "yes" but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 12.5%	10 7.9%	11
2. No Deception Indicated	7 87.5%	116 92.1%	123
Total Sample	8 6.0%	126 94.0%	134

Chi Square = 0.208 (*p* = 0.648)
Sensitivity = 12.5%
Specificity = 92.1%
False Positive Index = 10.0
Correctly Classified: 87.3%

Removing the deceptive respondents who tested negative from the LVA sample had negligible effect on LVA's sensitivity but tripled its False Positive Index (from 3.3 to 10). This is because LVA identified more respondents who were deceptive and tested negative for marijuana than CVSA. When those individuals were removed, the accuracy of LVA was more similar to CVSA.

A third explanation for low accuracy for identifying deception about recent marijuana use may have to do with jeopardy. Since arrestees are less likely to be deceptive about marijuana use than other drugs, we might be observing a problem where arrestees are not "worried enough" about marijuana use to register distress about deception. Testing the programs for drugs that have traditionally had higher deception rates (cocaine, methamphetamine, and PCP), may mitigate this problem. We turn to those analyses now.

In Table 12, we show the comparisons between deceptive responses and indications of deception for cocaine use among those who tested positive for cocaine. Among the 87 respondents who tested positive for cocaine, almost half (46%) provided a deceptive response (said that they had not used cocaine in the past 72 hours). Among the 47 respondents who were not deceptive, 42 (89%) were not classified as indicating deception by the VSA programs. The False Positive Index (0.63) is very low compared to marijuana. Unfortunately, the VSA programs only correctly indicated deception for eight of the 40 respondents who were deceptive about recent cocaine use. This resulted in a sensitivity score of 20%, much higher than we observed for similar data for marijuana. This suggests that the VSA programs may have greater accuracy for questions about drugs that have increased jeopardy (such as cocaine) than for drugs that have less social stigma attached (such as marijuana). Still, 20% sensitivity and 90%

specificity is not very accurate. Plotting these values on a ROC chart would show that the VSA programs are only about 70% accurate.

Table 12. Interpretation of deception about cocaine by cocaine users.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	8 20.0%	5 10.6%	13
2. No Deception Indicated	32 80.0%	42 89.4%	74
Total Sample	40 46.0%	47 54.0%	87

Chi Square = 1.490 (*p* = 0.222)
Sensitivity = 20.0%
Specificity = 89.4%
False Positive Index = 0.625
Correctly Classified: 57.2%

When we examined all of the respondents in the sample (Table 13), we observe far less accuracy by the VSA programs for predicting deception about recent cocaine use. Almost 80% of the 311 respondents were correctly classified as deceptive or not deceptive. Only about 15% (n=45) provided a deceptive response while 86% were not deceptive. Among the 45 respondents who were deceptive about recent cocaine use, 8 were correctly identified by the VSA programs. This is a very low sensitivity rate (17.8%). On the other hand, 26 (9.8%) of those who were not being deceptive were identified by the VSA programs as being deceptive. This results in a False Positive Index of 3.25 (for every one true positive, we observe 3.25 false positives). These data suggest that the VSA programs are not very useful for determining deception about recent cocaine use among an arrestee population. It appears that increased levels of jeopardy do not have much of an affect on the accuracy rate of the VSA programs.

Table 13. Interpretation of deception about cocaine by all respondents.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	8 17.8%	26 9.8%	34
2. No Deception Indicated	37 82.2%	240 90.2%	277
Total Sample	45 14.5%	266 85.5%	311

Chi Square = 2.532 (*p* = 0.112)
Sensitivity = 17.8%
Specificity = 90.2%
False Positive Index = 3.25
Correctly Classified: 79.7%

As we did with marijuana, we also re-analyzed the cocaine data separately for CVSA and LVA (see Tables 14 and 15). Neither program performed very well on this test, although the chi square for LVA approaches significance. While the CVSA program had much higher sensitivity (23% compared to 13% for LVA) it suffered from lower specificity (86% compared to 96% for LVA). CVSA exhibited a high False Positive Index score (4.2), suggesting that many non-deceivers were incorrectly being classified as deceptive. Only two “deceptive” respondents tested negative for cocaine, so unlike the case with marijuana, we do not present separate tables removing those cases. The findings are almost exactly the same and we are less concerned with the problem of forgetfulness with cocaine (since this is a shorter time window) than we were about marijuana.

Table 14. Interpretation of deception about cocaine by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	5 22.7%	21 14.5%	26
2. No Deception Indicated	17 77.3%	124 85.5%	141
Total Sample	22 13.2%	145 86.8%	167

Chi Square = 0.988 (*p* = 0.320)
Sensitivity = 22.7%
Specificity = 85.5%
False Positive Index = 4.2
Correctly Classified: 77.2%

Table 15. Interpretation of deception about cocaine by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	3 13.0%	5 4.1%	8
2. No Deception Indicated	20 87.0%	116 95.9%	136
Total Sample	23 16.0%	121 84.0%	144

Chi Square = 2.925 (*p* = 0.087)
Sensitivity = 13.0%
Specificity = 95.9%
False Positive Index = 1.67
Correctly Classified: 82.6%

In Tables 16-19, we present the deception detection validity data for opiates. In this case, we asked respondents about recent use of heroin and then tested their urine for opiates. We recognize that drugs other than heroin can elicit a positive test for opiates in urine, but since only two respondents tested positive for opiates, we are less concerned about this problem than we might otherwise be.

Table 16. Interpretation of deception about opiates by opiate users.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	0 0%	0
2. No Deception Indicated	2 100%	0 0%	2
Total Sample	2 100%	0 0%	2

Chi Square = not calculated
Sensitivity = 0%
Specificity = 0%
False Positive Index = 0
Correctly Classified: 0%

As shown in Table 16, both respondents who tested positive for opiates reported no recent use of heroin. Neither respondent was identified by the VSA programs as being deceptive, but again, they might have been using something other than heroin and therefore were not actually being deceptive. Table 17 reveals an additional 3 respondents who were “deceptive” about recent heroin use (those who said that they had used heroin recently but tested positive). Only one such person was identified as being deceptive. In addition, though, 14 (4.5%) of the 311 non-deceptive respondents were incorrectly identified as being deceptive, resulting in a large False Positive Index of 14. As we see in Tables 18 and 19, however, most of the false positives shown in Table 17 were the result of the CVSA instrument, which incorrectly identified 10 non-deceptive respondents as deceptive.

Table 17. Interpretation of deception about opiates by all respondents.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 20.0%	14 4.5%	15
2. No Deception Indicated	4 80.0%	297 95.5%	301
Total Sample	5 1.6%	311 98.4%	316

Chi Square = 2.614 (*p* = 0.106)
Sensitivity = 20.0%
Specificity = 95.5%
False Positive Index = 14.0
Correctly Classified: 94.3%

Table 18. Interpretation of deception about opiates by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	10 6.1%	10
2. No Deception Indicated	2 100%	155 93.9%	157
Total Sample	2 1.2%	165 98.8%	167

Chi Square = 0.129 (*p* = 0.720)
Sensitivity = 0%
Specificity = 93.9%
False Positive Index = not calculated
Correctly Classified: 92.8%

Table 19. Interpretation of deception about opiates by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 33.3%	4 2.7%	5
2. No Deception Indicated	2 66.7%	142 97.3%	144
Total Sample	3 2.0%	146 98.0%	149

Chi Square = 8.484 (*p* = 0.004)
Sensitivity = 33.3%
Specificity = 97.3%
False Positive Index = 4
Correctly Classified: 96.0%

Interestingly, the chi square for Table 19 is significant, suggesting a strong relationship between the actual deception and indications of deception. This is largely due to the very high specificity (almost all non-deceptive respondents are correctly classified) and the relatively low number of respondents who were "deceptive." Since chi square is sensitive to cells with few respondents in them, we worry that this relationship might be over stated. Since we are not certain about deception in this case because of the question about opiates vs. heroin, we are cautious about assigning too much value to this finding.

In Tables 20 through 23, we present the validity results for deception detection for questions relating to recent PCP use. Only 16 respondents tested positive for PCP use but over half of those (56.3%) were deceptive about recent PCP use. The relationship between deceptiveness of responses and indications of deceptiveness is not significant even though the specificity is 100% (all non-deceptive responses are correctly identified). This lack of significance is likely tied to the small number of respondents in the cells (expected values lower than 5 tend to negatively affect chi square). That said, the sensitivity for tests of deception related to questions about PCP is very low. Only 22% of the deceptive users were identified by the VSA programs.

Table 20. Interpretation of deception about PCP by PCP users.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	2 22.2%	0 0%	2
2. No Deception Indicated	7 77.8%	7 100%	14
Total Sample	9 56.3%	7 43.8%	16

Chi Square = 1.778 (*p* = 0.182)
Sensitivity = 22.2%
Specificity = 100%
False Positive Index = 0
Correctly Classified: 56.3%

In Table 21, we present the data for the whole sample regarding deception about recent PCP use. No respondents said that they had used PCP and then tested negative for PCP. Among the 303 respondents who truthfully reported not using PCP recently, 16 (5.3%) were incorrectly identified as being deceptive. As a result, the VSA programs have a False Positive Index of 8 regarding recent PCP use. As shown in Tables 22 and 23, the number of false positives is about evenly distributed among the CVSA and LVA findings. That said, the relationship between deceptiveness of response and indications of deceptiveness is significant (chi square = 4.615).

Although the low sensitivity rate of the findings warrants concern, this finding suggests that VSA programs have some limited predictive power concerning questions about recent PCP use.

Table 21. Interpretation of deception about PCP by all respondents.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	2 22.2%	16 5.3%	18
2. No Deception Indicated	7 77.8%	287 94.7%	294
Total Sample	9 2.9%	303 97.1%	312

Chi Square = 4.615 (*p* = 0.032)
Sensitivity = 22.2%
Specificity = 94.7%
False Positive Index = 8
Correctly Classified: 92.6%

When we compare the two VSA instruments, the findings are very interesting. There is no significant relationship between actual deceptiveness and successful indication of deception for CVSA. In fact, CVSA did not indicate deception for any of the 4 recent PCP users who were being deceptive and it incorrectly identified 9 non-deceptive respondents as being deceptive. Since no deceptive respondent was correctly identified, it is impossible to calculate the False Positive Index but the sensitivity score is zero. On the other hand, there is a significant relationship between the extent of actual deceptiveness and indications of deceptiveness for the LVA instrument. While this is likely due to the small number of deceptive respondents (n=5), these results are interesting nonetheless. The specificity rate of 40% for PCP is the highest we found in the project and the specificity rate is as high as we observed for other drugs. The relatively high False Positive Index (3.5) is cause for concern along with a specificity rate of less than 50% is still cause for concern.

Table 22. Interpretation of deception about PCP by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	9 5.5%	9
2. No Deception Indicated	4 100%	154 94.5%	158
Total Sample	4 2.4%	163 97.6%	167

Chi Square = 0.233 (*p* = 0.629)
Sensitivity = 0%
Specificity = 94.5%
False Positive Index = Not calculated
Correctly Classified: 92.2%

Table 23. Interpretation of deception about PCP by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	2 40.0%	7 5.0%	9
2. No Deception Indicated	3 60.0%	133 95.0%	136
Total Sample	5 3.4%	140 96.6%	145

Chi Square = 10.158 (*p* = 0.001)
Sensitivity = 40%
Specificity = 95.0%
False Positive Index = 3.5
Correctly Classified: 93.1%

Finally, we present the validity data for methamphetamine in Tables 24 through 27. As shown in Table 24, among the 57 respondents who tested positive for methamphetamine, 22 (39%) reported that they had not used recently (were deceptive). The VSA instruments had a very low sensitivity score (9.1%) but they also had a very low False Positive Index (1.0). Only about 61% of the respondents were correctly classified, due mostly to the low sensitivity score (the specificity score was a respectable 94%).

Table 24. Interpretation of deception about methamphetamine by methamphetamine users.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	2 9.1%	2 5.7%	4
2. No Deception Indicated	20 90.9%	33 94.3%	53
Total Sample	22 38.6%	35 61.4	57

Chi Square = 0.236 (*p* = 0.627)
Sensitivity = 9.1%
Specificity = 94.3%
False Positive Index = 1.0
Correctly Classified: 61.4%

When we examined deceptiveness related to methamphetamine among all respondents in Table 25, we also observe a very low sensitivity pattern as we did among the methamphetamine users (7.1%). More alarming, however, is the large number of non-deceptive respondents (n=32) who were identified by the instruments as being deceptive about recent methamphetamine use. This resulted in a lower specificity score of 89% and a False Positive Index of 16. Clearly, VSA instruments did not accurately indicate deception about questions related to recent methamphetamine use.

Table 25. Interpretation of deception about methamphetamine by all respondents.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	2 7.1%	32 11.2%	34
2. No Deception Indicated	26 92.9%	254 88.8%	280
Total Sample	28 8.9%	286 91.1%	314

Chi Square = 0.432 (*p* = 0.511)
Sensitivity = 7.1%
Specificity = 88.8%
False Positive Index = 16.0
Correctly Classified: 81.5%

Table 26. Interpretation of deception about methamphetamine by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 6.3%	26 17.6%	27
2. No Deception Indicated	15 93.8%	125 82.8%	140
Total Sample	16 9.6%	151 90.4%	167

Chi Square = 1.284 (*p* = 0.257)
Sensitivity = 6.3%
Specificity = 82.8%
False Positive Index = 26.0
Correctly Classified: 75.5%

The findings reported in Tables 26 and a27 are similar as well. Neither VSA program performed well although the CVSA program had a False Positive Index of 26 (the highest we observed in this study). This resulted in a significantly lower specificity rate for CVSA than was observed for LVA, although both programs had a sensitivity rate lower than 9%.

Table 27. Interpretation of deception about methamphetamine by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 8.3%	6 4.4%	7
2. No Deception Indicated	11 91.7%	129 95.6%	140
Total Sample	12 8.2%	135 91.8%	147

Chi Square = 0.368 (*p* = 0.544)
Sensitivity = 8.3%
Specificity = 95.6%
False Positive Index = 6.0
Correctly Classified: 88.4%

For one final validity test, we prepared Tables 28-30 that compare the extent of deceptiveness with indications of deception for any drug. That is, if the respondent was deceptive about any of the five drugs, they were coded as being deceptive. If the VSA instruments indicated deception for any of the five drugs, then we coded that as an indication of deception. We consider this a "global" validity assessment that takes into account the possibility that the rapid succession of questions might have confused either the programs or the respondents. Thus, we think that this is a liberal test of the programs, since the VSA instruments should at least be able to measure some deception when the respondent was being deceptive about at least one of the drugs. Because so many respondents who tested negative for marijuana reported that they had recently used marijuana (thereby artificially inflating the number of "deceptive" respondents, we removed those respondents from the analysis. The tables including those 34 respondents are located in Appendix G.

The results of this "global" test are disappointing. Not including the 34 respondents who reported marijuana use but tested negative, Table 29 shows that 80 of the 283 respondents (28.3%) were deceptive about at least one of their answers related to recent drug use. The VSA programs were only able to correctly identify 25 of those deceptive respondents, resulting in a sensitivity rate of 31.3%. Unfortunately, the specificity rate (74.4%) was also very low, resulting in 52 respondents who were not deceptive on any of the drug questions being identified as deceptive. This resulted in a False Positive Index of 2.08, with only 62.2% of the cases being correctly classified.

Table 28. Interpretation of deception about any drug by all respondents removing those who reported marijuana use but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	25 31.3%	52 25.6%	77
2. No Deception Indicated	55 68.8%	151 74.4%	206
Total Sample	80 28.3%	203 71.7%	283

Chi Square = 0.920 (*p* = 0.338)
Sensitivity = 31.3%
Specificity = 74.4%
False Positive Index = 2.08
Correctly Classified: 62.2%

Table 29. Interpretation of deception about any drug by all respondents for CVSA removing those who reported marijuana use but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	14 35.0%	38 35.2%	52
2. No Deception Indicated	26 65.0%	70 64.8%	96
Total Sample	40 27.0%	108 73.0%	148

Chi Square = 0.000 (*p* = 0.983)
Sensitivity = 35.0%
Specificity = 64.8%
False Positive Index = 2.71
Correctly Classified: 56.8%

When we compare the two programs, we observe that the CVSA instrument had a sensitivity rate of 35% but a relatively poor specificity rate of 65%. This resulted in a False Positive Index of 2.71 with only 57% of the respondents being correctly classified. The LVA instrument fared slightly better, with a lower sensitivity rate (28%) but a higher specificity rate (85%). This resulted in a False Positive Index of 1.27, one of the lowest scores observed in this study.

We note that the deception rate was about the same for both groups, suggesting that likelihood of deception by the respondents was not adversely affected by the type of instrument being used by the research team. Actually, we found this with all of the tables that compared CVSA with LVA with the exception of cocaine, where a slightly greater percentage of LVA respondents (16%) were deceptive than CVSA respondents (13%).

Table 30. Interpretation of deception about any drug by all respondents for LVA removing those who reported marijuana use but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	11 27.5%	14 14.7%	25
2. No Deception Indicated	29 72.5%	81 85.3%	110
Total Sample	40 29.6%	95 70.4%	135

Chi Square = 3.039 (*p* = 0.081)
Sensitivity = 27.5%
Specificity = 85.3%
False Positive Index = 1.27
Correctly Classified: 68.2%

Summary. Our findings suggest very little support for the validity of either VSA instrument to detect deception about recent drug use by arrestees (see Table 31). For each of the five drugs, the average sensitivity rate for the VSA programs (the percent of deceptive respondents correctly identified as being deceptive) was very low (15%). The sensitivity rate was more than twice as high for LVA (21%) compared with CVSA (8%). The specificity rate (the percent of non-deceptive respondents who were correctly classified as non-deceptive) was much higher, averaging 92% across each of the drugs. Again, LVA performed slightly better (specificity = 95%) compared to CVSA (specificity = 90%). The False Positive Index (the ratio of “false positives” to “true positives”) was disappointing for both VSA programs, averaging 9.4. Again, LVA fared considerably better (average FPI = 5.0) in comparison with CVSA (average FPI = 14.1). Note, however, that we were not able to calculate the FPI for CVSA on two drugs (Opiates and PCP) because no deceptive respondent was correctly identified.

Table 31. Interpretation of deception about drug use by all respondents for both VSA instruments, for CVSA, and for LVA.

Statistics for VSA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.012	2.532	2.614	4.615*	0.432	-
Sensitivity	7.8%	17.8%	20.0%	22.2%	7.1%	15.0%
Specificity	88.3%	90.2%	95.5%	94.7%	88.8%	91.5%
False Positive Index	5.5	3.25	14.0	8.0	16.0	9.4
Correctly Classified:	75.3%	79.7%	94.3%	92.6%	81.5%	84.7%

* $p < 0.05$

Statistics for CVSA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.065	0.988	0.129	0.233	1.284	-
Sensitivity	11.1%	22.7%	0%	0%	6.3%	8.0%
Specificity	91.4%	85.5%	93.9%	94.5%	82.8%	89.6%
False Positive Index	12.0	4.2	N/A	N/A	26.0	14.1
Correctly Classified:	86.5%	77.2%	92.8%	92.2%	75.5%	84.8%

* $p < 0.05$

Statistics for LVA	Marijuana	Cocaine	Opiates	PCP	Meth	Average
Chi Square	0.208	2.925	8.484*	10.16*	0.368	-
Sensitivity	12.5%	13.0%	33.3%	40%	8.3%	21.4%
Specificity	92.1%	95.9%	97.3%	95.0%	95.6%	95.2%
False Positive Index	10.0	1.67	4.0	3.5	6.0	5.0
Correctly Classified:	87.3%	82.6%	96.0%	93.1%	88.4%	89.5%

* $p < 0.05$

In general, we observed that both programs were not able to efficiently determine who was being deceptive (poor sensitivity) and that CVSA was more likely to label a non-deceptive person as deceptive (lower specificity than LVA). Only LVA showed a significant chi square statistic (for Opiates and PCP). While these significant findings suggest support for LVA, we urge caution given the small number of respondents who were deceptive (which resulted in an artificially high chi square) and the very low sensitivity rates across the board.

b. Evaluating the Reliability of the VSA Programs.

There are two reliability issues at stake in VSA studies. The first has to do with the reliability of the program to consistently detect deception over time from person-to-person. Testing this kind of reliability test would require a situation where a subject was asked at least twice to be deceptive to test if the program could detect the deception in both cases. This situation would be next to impossible to re-create and we do not attempt to do so in this study.

A second type of reliability test, however, is referred to as inter-rater reliability. In this case, we are comparing the ability of different types of raters to evaluate the deceptiveness of respondents using the VSA software. Comparing novice analysts to each other may only be indicative of successful training, but comparing novice analysts to expert analysts allows us to assess the extent to which (1) relatively inexperienced analysts can quickly grasp the software and the interpretation of the output, and (2) expert analysts rely on intuition or exceptional interrogation skills that fall outside of the software purview to successfully detect deception. Thus, our key questions are (1) can the technology be easily learned and (2) does the technology require special interrogation “instinct” or skills to determine deceptive answers? In our study, the “novice” examiners worked together to determine if the VSA programs were indicating deception or not. At the same time, we asked trainers from each of the VSA vendors to assess our data independently of our own assessments. Each novice and expert assessment was done independently of each other (neither knew what the other had reported) and independent of any knowledge of the results of the urine test. In the following tables, we present the detailed results for the CVSA and the LVA experts along with the summary statistics for the novice examiners to aid comparison.

Our first test of the inter-rater reliability of the novices and VSA experts is to examine the correlations and kappa coefficients comparing individual deception assessments by the experts and the novices. Correlation statistics are a common (and intuitive) method of comparing inter-rater reliability, especially in the case of interval data (Fleiss, 1971; Shrout and Fleiss, 1979). Since the variables in these comparisons are *dichotomous*, the appropriate correlation coefficient is the *phi* (Yule, 1912). Since we are comparing only two raters, a second appropriate coefficient would be Cohen's Kappa (Cohen 1960; Cohen 1968). Cohen's Kappa calculates the inter-rater reliability by taking into account any agreement that might occur by chance alone. Kappa *normally* ranges from 0-1, with 1 representing perfect agreement between the two raters. Scores above 0.60 are considered "good" agreement and scores above 0.80 are considered "very good." The correlation (*phi*) and Kappa coefficients shown in Table 32 reflect very strong consistency with each other, suggesting that both are robust measures of agreement.

Table 32. Correlation (*phi*) and kappa coefficients comparing expert and novice examiners for each drug not including respondents who reported marijuana use but tested negative.

	CVSA Experts vs. Novices		**LVA Experts vs. Novices**	
Deceptive for…:	Phi	Cohen's Kappa	Phi	Cohen's Kappa
Marijuana	.160	.152	.353**	.351**
Cocaine	.389**	.389**	.109	.108
Opiates	.311**	.221**	.116	.104
PCP	.524**	.431**	.207*	.195*
Methamphetamine	.321**	.303**	.178*	.151*

* Coefficient is significant at the 0.05 level (2-tailed).
** Coefficient is significant at the 0.01 level (2-tailed).

We observe that there appears to be greater correlation between the CVSA experts and the novices than between the LVA experts and the novices. Interestingly, for the agreement between the CVSA experts and the novices for marijuana was not significant for either statistic while it was significant for the LVA experts. The correlation between the CVSA experts and the novices, however, for all other drugs is at least 0.31 (and as high as 0.52 for PCP). The correlations between LVA experts and the novices, however, for cocaine, opiates, PCP, and methamphetamine, were much less robust. In fact, only the correlations for PCP and methamphetamine were significant. As is shown in the tables that follow, this appears to be a reticence (or an inability) on the part of the LVA experts to identify deceptive respondents. That said, it is important to point out that inter-rater reliability assessments would expect to find much higher correlation and Kappa coefficients (at least r = .80) in order to be seen as "reliable." The correlations reported in Table 32 suggest that the programs do not have inter-rater reliability.

In Tables 33 and 34, we show the results of the assessments conducted by the CVSA and LVA expert examiners (respectively). It is important to note that the CVSA expert did not "make a call" on several cases because he said (1) the charts were inconclusive or (2) the charts showed evidence of over-modulation which made them unreadable. As a result, we only have expert data on about one-half of the CVSA data to compare. For the 74 respondents for which the CVSA expert assessed deception regarding the marijuana question, he correctly identified about one in four (25%) recent users who were being deceptive. While that compares favorably to the sensitivity score for the novice, his specificity score (77%) was much lower than the novices (91%), suggesting that the novices did a better job correctly identifying non-deceptive respondents. Again, the CVSA expert failed to assess so many responses (most of which were correctly classified by the novices as non-deceptive responses), making such comparisons difficult. That said, the correlation between the determination of deception for the novices and the CVSA expert was not significant (r = 0.16).

Table 33. Expert interpretation of deception about marijuana by all respondents for CVSA removing those who reported marijuana use but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 25.0%	16 22.9%	17
2. No Deception Indicated	3 75.0%	54 77.1%	57
Total Sample	4 5.4%	70 94.6%	74

Expert	**Novice (Table 10)**
Chi Square = 0.010 (*p* = 0.921) Sensitivity = 25.0% Specificity = 77.1% False Positive Index = 16 Correctly Classified: 74.3%	Chi Square = 0.065 (*p* = 0.799) Sensitivity = 11.1% Specificity = 91.4% False Positive Index = 12 Correctly Classified: 86.5%

Comparison of the deception assessments for the marijuana questions by the LVA experts and the novices, however, suggests relatively greater inter-rater reliability (Table 33). Both the novices and the LVA experts had a sensitivity score of 12.5% and a specificity score of over 92%. Both, unfortunately, had a relatively high False Positive Index (9 for the experts and 10 for the novices) and both correctly classified about 87% of the respondents. The correlation between the determination of deception for the novices and the LVA experts was significant but not large (r = 0.35).

Table 34. Expert interpretation of deception about marijuana by all respondents for LVA removing those who reported marijuana use but tested negative.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 12.5%	9 7.2%	10
2. No Deception Indicated	7 87.5%	116 92.8%	123
Total Sample	8 6.0%	125 94.0%	133

Expert	**Novice (Table 11)**
Chi Square = .304 (*p* = 0.582) Sensitivity = 12.5% Specificity = 92.8% False Positive Index = 9.0 Correctly Classified: 88.0%	Chi Square = 0.208 (*p* = 0.648) Sensitivity = 12.5% Specificity = 92.1% False Positive Index = 10.0 Correctly Classified: 87.3%

In Table 35 and 36, we present the expert/novice comparisons for cocaine. While the LVA experts and the novices were much more similar for marijuana, we find just the opposite for cocaine. Specifically, the sensitivity for the CVSA experts and novices are 25% and 23% respectively (Table 35). The major difference, however, is that the novices have high specificity rates (86% vs. 78%) and a lower False Positive Index (4.2 vs. 5.3). The correlation between the two sets of examiners was significant but small (r = 0.389), suggesting relatively low reliability between the two sets of raters.

The comparison between the LVA experts and the novices for the cocaine question was even less favorable (Table 36). The LVA experts had a much lower sensitivity score (4%) by only correctly identifying one of the deceptive cocaine users compared to a sensitivity score of 13% for the novices. Both the novices and the LVA experts had similar specificity scores but the

False Positive Index for the experts was more than 3 times greater than the scores for the novices due to the low sensitivity. The correlation between the two sets of examiners was insignificant (r = 0.109), suggesting low reliability.

Table 35. Expert interpretation of deception about cocaine by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	3 25.0%	16 22.5%	19
2. No Deception Indicated	9 75.0%	55 77.5%	64
Total Sample	12 14.5%	71 85.5%	83

Expert	**Novice (Table 14)**
Chi Square = 0.035 (*p* = 0.851)	Chi Square = 0.988 (*p* = 0.320)
Sensitivity = 25%	Sensitivity = 22.7%
Specificity = 77.5%	Specificity = 85.5%
False Positive Index = 5.3	False Positive Index = 4.2
Correctly Classified: 69.9%	Correctly Classified: 77.2%

Table 36. Expert interpretation of deception about cocaine by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 4.0%	7 5.7%	8
2. No Deception Indicated	24 96.0%	116 94.3%	140
Total Sample	25 16.9%	123 83.1%	148

Expert	**Novice (Table 15)**
Chi Square = 0.116 (*p* = 0.733)	Chi Square = 2.925 (*p* = 0.087)
Sensitivity = 4.0%	Sensitivity = 13.0%
Specificity = 94.3%	Specificity = 95.9%
False Positive Index = 7	False Positive Index = 1.67
Correctly Classified: 79.1%	Correctly Classified: 82.6%

In Tables 37 and 38, we show the comparisons between the VSA experts and the novices for the questions related to recent opiates use. Since so few respondents (n=2) tested positive for opiates, we do not dwell too long on these comparisons except to point out that the correlation between the CVSA experts (based on the 83 responses that were evaluated) was significant (r = 0.311) which is interesting since the specificity rate for the novices was considerably higher (94%) than it was for the experts (66%). The correlation between the novices and the LVA experts was not significant (r = 0.116). We note that neither set of experts were able to correctly classify any of the respondents who were deceptive about recent opiate use.

Table 37. Expert interpretation of deception about opiates by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	28 34.1%	28
2. No Deception Indicated	1 100%	54 65.9%	55
Total Sample	1 1.2%	82 98.8%	83

Expert	**Novice (Table 18)**
Chi Square = 0.515 (*p* = 0.473) Sensitivity = 0% Specificity = 65.9% False Positive Index = not calculated Correctly Classified: 65.1%	Chi Square = 0.129 (*p* = 0.720) Sensitivity = 0% Specificity = 93.9% False Positive Index = not calculated Correctly Classified: 92.8%

Table 38. Expert interpretation of deception about opiates by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	12 8.3%	12
2. No Deception Indicated	3 100%	133 91.7%	136
Total Sample	3 2.0%	145 98.0%	148

Expert	**Novice (Table 18)**
Chi Square = 0.270 (*p* = 0.603) Sensitivity = 0% Specificity = 91.7% False Positive Index = not calculated Correctly Classified: 89.9%	Chi Square = 8.484 (*p* = 0.004) Sensitivity = 33.3% Specificity = 97.3% False Positive Index = 4 Correctly Classified: 96.0%

In Tables 39 and 40, we present the comparisons between the VSA experts and the novices for questions related to recent PCP use. Compared with the CVSA and LVA experts, the novices have higher (though similar) specificity scores. The correlation between the novices and both sets of VSA experts was significant (r = 0.52 for CVSA and r = 0.21 for LVA), suggesting that there is a relatively higher level of inter-rater reliability for both instruments concerning recent PCP use than for the other drugs (although these data still do not approach the .80 threshold needed for inter-rater reliability).

Table 39. Expert interpretation of deception about PCP by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	20 24.2%	20
2. No Deception Indicated	1 100%	62 75.6%	63
Total Sample	1 1.2%	82 98.8%	83

Expert	**Novice (Table 22)**
Chi Square = .321 (*p* = 0.571) Sensitivity = 0% Specificity = 75.6% False Positive Index = Not calculated Correctly Classified: 74.7%	Chi Square = 0.233 (*p* = 0.629) Sensitivity = 0% Specificity = 94.5% False Positive Index = Not calculated Correctly Classified: 92.2%

Table 40. Expert interpretation of deception about PCP by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	1 20.0%	1 11.9%	18
2. No Deception Indicated	4 80.0%	126 88.1%	130
Total Sample	5 3.4%	143 96.6%	148

Expert	**Novice (Table 22)**
Chi Square = 0.298 (*p* = 0.585) Sensitivity = 20% Specificity = 88.1% False Positive Index = 1 Correctly Classified: 85.8%	Chi Square = 10.158 (*p* = 0.001) Sensitivity = 40% Specificity = 95.0% False Positive Index = 3.5 Correctly Classified: 93.1%

Finally, we present the comparisons of novices versus VSA experts for recent methamphetamine use in Table 41 and 42. Comparing the "expert" statistics in Table 40 to the "novice" data provided in Table 26 is problematic because so few respondents were coded by the expert. As a result, the significant correlation between the two sets of examiners (r = 0.32) does not seem to correspond to the data since the sensitivity of the CVSA experts (33%) is more than 5 times greater than the sensitivity of the novice data. Comparison of the novice data for just the cases evaluated by the CVSA expert (not shown here), however, shows much greater similarity.

Table 41. Expert interpretation of deception about methamphetamine by all respondents for CVSA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	3 33.3%	22 29.7%	25
2. No Deception Indicated	6 66.7%	52 70.3%	58
Total Sample	9 10.8%	74 89.2%	83

Expert	**Novice (Table 26)**
Chi Square = 0.050 (*p* = 0.824)	Chi Square = 1.284 (*p* = 0.257)
Sensitivity = 33.3%	Sensitivity = 6.3%
Specificity = 70.3%	Specificity = 82.8%
False Positive Index = 7.3	False Positive Index = 26.0
Correctly Classified: 66.3%	Correctly Classified: 75.5%

The correlation between the LVA and novice examiners was likewise significant (r = 0.18), but not sufficiently high to support a claim of inter-rater reliability. The LVA experts failed to correctly identify any of the respondents who were deceptive about recent methamphetamine use (compared to the very low 8.3% sensitivity rate of the novices). Both sets of examiners had relatively high specificity rates although the novices were higher (96% compared to 88%).

Table 42. Expert interpretation of deception about methamphetamine by all respondents for LVA.

VSA Classification	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	0 0%	16 11.8%	16
2. No Deception Indicated	12 100%	120 88.2%	13
Total Sample	12 8.1%	136 91.9%	145

Expert	**Novice (Table 27)**
Chi Square = 1.583 (*p* = 0.208) Sensitivity = 0% Specificity = 88.2% False Positive Index = Not calculated Correctly Classified: 82.8%	Chi Square = 0.368 (*p* = 0.544) Sensitivity = 8.3% Specificity = 95.6% False Positive Index = 6.0 Correctly Classified: 88.4%

In Table 43, we combined the individual assessments about deceptiveness for each drug for the CVSA sample. The novices made 844 such assessments while the CVSA expert made 415 deception assessments. The sensitivity of both sets of examiners is remarkably similar, with the novices correctly identifying 19.8% of the deceptive responses using the CVSA instrument compared to 19.4% for the CVSA expert. The novices have a higher specificity rate (90% vs. 73%) and a decidedly lower False Positive Index (4.9 vs. 14.6). The results suggest, therefore, little support for the inter-rater reliability between CVSA experts and novices. In general, the novices seemed to do a better job of distinguishing among the non-deceptive respondents.

Table 43. Summary table showing novice and expert interpretation of deception about all drug use by all respondents for CVSA.

VSA Classification		Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	Novice	16 19.8%	78 10.2%	94
	Expert	7 19.4%	102 26.9%	109
2. No Deception Indicated	Novice	65 80.2%	685 89.8%	750
	Expert	29 80.6%	277 73.1%	306
Total	Novice	81	763	844
	Expert	36	379	415

Expert	**Novice**
Sensitivity= 19.4%	Sensitivity= 19.8%
Specificity = 73.1%	Specificity = 89.8%
False Positive Index = 14.57	False Positive Index = 4.875
Correctly Classified: 68.4%	Correctly Classified: 83.1%

We present the same information for the LVA instrument in Table 44. The data suggest that the sensitivity score for the novices was more than three times higher (14.9%) than that for the LVA experts (4.4%). Both groups had about the same level of specificity but the LVA experts had a False Positive Index five times greater than the novices, owing mostly to the low sensitivity score recorded by the experts. This suggests relatively low inter-rater reliability for the LVA instrument on sensitivity but high reliability on specificity.

Table 44. Summary table showing novice and expert interpretation of deception about all drug use by all respondents for LVA.

VSA Classification		Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	Novice	10 14.9%	32 4.8%	42
	Expert	3 4.4%	45 6.7%	48
2. No Deception Indicated	Novice	57 85.1%	629 94.3%	686
	Expert	65 95.6%	623 93.3%	688
Total	Novice	67	667	734
	Expert	68	668	736

Expert	**Novice**
Sensitivity= 4.4%	Sensitivity= 14.9%
Specificity = 93.3%	Specificity = 94.3%
False Positive Index = 15	False Positive Index = 3.2
Correctly Classified: 85.1%	Correctly Classified: 87.1%

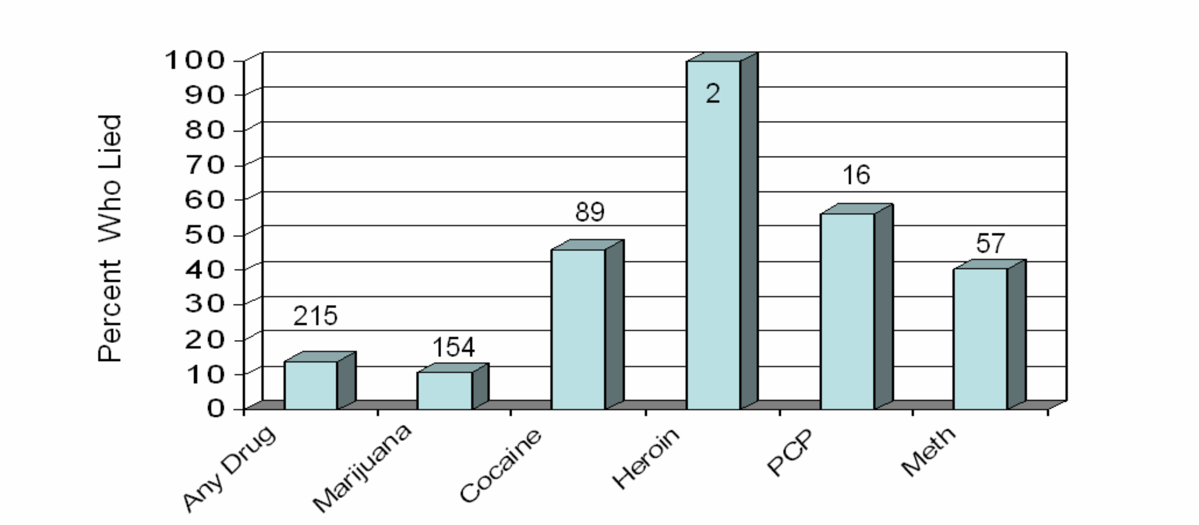
Summary. The data presented in these tables suggest that there is some similarity between expert VSA examiners and novice examiners. The deception assessments by the novices were more highly correlated with the CVSA experts (especially concerning sensitivity) but the novices also were very similar to the LVA experts in regards to specificity. The major difference seemed to be that LVA experts had much lower sensitivity than the novices either because of a reticence to make the call or an inability of the data to clearly indicate deception. When we compared the percent of responses that were correctly classified, the novices almost always had higher percentages. The only exception was the percent of correctly classified responses for marijuana for the LVA experts and novices when both were about 88%. The observed correlations between novices and experts, however, suggest a relatively low level of inter-rater reliability for both instruments. This suggests that the output is sufficiently open to interpretation that two raters looking at the same material regularly interpret the data differently.

c. Bogus Pipeline Effect.

During the course of the data collection, the interview team reported to the project director that they were concerned that the arrestees would be deceptive at a lower rate than the original ADAM project because they believed that the team was using a "lie detector." This effect is referred to as the bogus pipeline effect, where researchers increase truthfulness in self-report subjects by convincing participant that they have a "pipeline" into the truth (e.g., they will be able to tell if the subject is lying). The expectation is that subjects will answer more honestly if they believe that the truth can be tested for accuracy (Jones and Sigall, 1971). This phenomenon has been successfully tested in studies that rely on substance use (Aguinis, Pierce, and Quigley, 1993; Botvin, Botvin, Renick, Filazzola, & Allegrante, 1984; Sprangers and Hoogstraten, 1987) while other studies have suggested that the bogus pipeline is ineffective (Akers, Massey, Clarke, and Lauer, 1983; Campanelli, Dielman, and Shope, 1987).

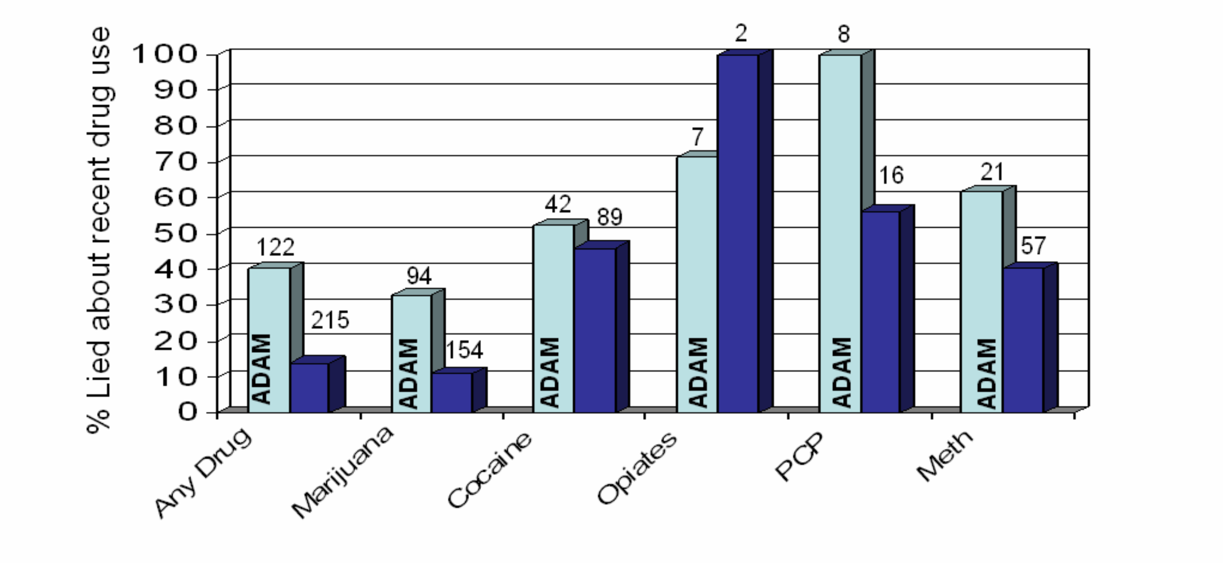
In our study, the arrestees were told before the survey that we were going to perform a urinalysis (just like in the ADAM study) but they were also aware that we were using a voice stress analysis program that might be able to detect any deceptive answers. We were interested in the impact of this knowledge on the arrestees' likelihood of being deceptive. We performed the analysis by calculating the percent of deceptive answers for "any drug" and then for each drug in turn. The data are presented in Figure 15. Each bar represents the percent of deceptive responses and the corresponding number of arrestees who had tested positive for the drugs. For example, 215 respondents tested positive for at least one drug but 30 arrestees (14.1%) were deceptive – they reported no recent drug use. Thus, only about 11% of the recent marijuana users were deceptive while approximately one half of the cocaine (46%), PCP (56%) and methamphetamine (40%) users were deceptive. There were only two people who tested positive for opiates, so the large percentage of "deceptive" responses is not reliable.

Figure 15. Percent of "users" who were deceptive about recent drug use (VSA 2006). Numerals reflect the number of respondents who tested positive for each drug.



In Figure 16, we present a chart that compares the VSA 2006 data (the current project) with ADAM data that were collected in the last quarter of 2003. The same procedures were used to create the chart in Figure 15. The results are relatively dramatic. More than 3 times as many users of at least one drug were deceptive in the ADAM study (40%) compared with the VSA study (14%). A similar pattern holds for marijuana users (33% in ADAM compared to 11% in VSA), cocaine users (52% in ADAM compared to 46% in VSA), PCP (100% in ADAM compared to 56% in VSA), and methamphetamine users (62% in ADAM compared to 40% in VSA). Since so few opiate users are included in the VSA data, no easy comparison is possible.

Figure 16. Comparison of "users" who were deceptive about recent drug use (ADAM 2003 and VSA 2006). Numerals reflect the number of respondents who tested positive for each drug.



In Table 45, we present a contingency table comparing the percent of deception among those who tested positive for any drug for the VSA 2006 and the ADAM 2003 data. The difference between the percent of respondents who were deceptive in the VSA project (14%) compared to those who were deceptive in the ADAM project (40.2%) is statistically significant (chi square = 30.0).

Table 45. Comparing "Used Any Drug Recently" for ADAM 2003 and VSA 2006 Data.

	Admitted Any Drug Use	Did Not Admit Any Drug Use
ADAM-2003 (N=122)	73 59.8%	49 40.2%
VSA-2006 (N=215)	185 85.9%	30 14.0%

Chi Square = 30.0 (1), *p*<0.001

It is important to point out, however, that there were some differences in the protocol for the two data collection efforts. First, the ADAM survey was much longer and a wider variety of questions were asked of the arrestees. One might say, though, that this difference would increase truthfulness because of added time for rapport to develop. Instead, truthfulness was lower with the longer survey instrument. A second difference is in the location of the interview. The ADAM survey was conducted in the hallway just outside of the booking facility while the VSA survey was conducted in an interview room next to the booking area. While the VSA room was quieter, it was not necessarily private. Indeed, other inmates were escorted past the room during data collection and the security staff stood in the doorway during the interviews. We feel that the ADAM interviews were probably more private in the hallway since the guards were positioned further away and all interviews were temporarily suspended when people walked past.

Clearly, though, the use of the VSA computer programs affected the likelihood of deceptive answers by arrestees. Arrestees who thought their interviewers were using “lie detectors” were much less likely to be deceptive when reporting recent drug use. Notice, however, that the “seriousness” pattern holds for both surveys. Arrestees are more likely to be deceptive for “serious” drugs than they are for marijuana. Either way, the findings suggest strong support for the bogus pipeline effect of voice stress analysis programs on self-reports of criminal behavior. This finding is very important for law enforcement since it suggests that just using the VSA programs may be more likely to encourage suspects to be more truthful. Indeed, most of the anecdotal support from investigators who successfully used VSA software suggests that this is the case. When police officers report that VSA programs “work,” they generally mean that they were able to obtain a confession from suspects by telling them that the computer “said they were lying.” The potential problem, of course, is with false confessions. Several high profile cases have emerged in the past decade that suggest impressionable suspects may confess to a crime that they did not commit because they believed the software. The rationalization is usually that they “must have forgotten” that they did it. Obviously, the bogus pipeline effect of VSA products has important positive and negative implications.

5. Conclusions and Implications for Policy and Practice.

The ability of law enforcement and treatment personnel to be able to know if someone is being deceptive is vital. In the aftermath of 9/11, for example, many have called for the use of VSA technologies at airports (e.g., Crites, 2005). If VSA technology can be shown to work effectively, then one more tool can be made available to law enforcement and treatment agencies. On the other hand, if the technology does not work as advertised, then agencies need to be made aware of that as well. Law enforcement is already investing heavily in VSA technology based on purely anecdotal evidence. According to CVSA literature, for example, over 1,400 law enforcement users have been trained to use the CVSA instrument.

Previous scientific evidence that supports the technology is scant but we argued that the field experiment described in this report would provide a more realistic test. The results of the project will allow law enforcement agencies to make a more informed decision when deciding whether to invest in VSA technology. In an age of limited police budgets, this money might be spent differently if the device is not accurate. Thus, the key policy implication of this study addresses the question of the validity of the VSA devices being studied.

The goal of this study was to evaluate the ability of two VSA instruments to detect deceptive answers about recent drug use among an arrestee population. Asking relevant questions about recent criminal behavior by arrestees in a jail setting was deemed to be a suitable method of introducing jeopardy and reality into the evaluation. In addition, our ability to create grounded truth (reality) through the use of a urinalysis was an added improvement over similar studies that have been conducted in the past.

Two VSA programs were selected for evaluation (CVSA and LVA). The research team attended week-long training sessions with each VSA vendor and worked closed with the VSA trainers to develop a research design that would adequately assess each instrument. The survey and urine data were collected at the Oklahoma County Jail in February and March, 2006. Throughout the summer of 2006, the research team cleaned and analyzed the VSA data so that decisions could be made about the deceptiveness of each respondent's answer according to the VSA software. Following the completion of the deception assessment for each respondent, the research team compared urinalysis results for marijuana, cocaine, opiates, PCP, and methamphetamine to the responses supplied by each respondent to determine who was being deceptive. Then, the research team compared "actual" deception to "indicated" deception to determine the validity of the VSA instruments for the detection of deception about past deviant behavior. The team then compared assessments about the deceptiveness of responses by VSA experts and by the research team (novices). Finally, the researchers examined the extent to which using the VSA instrument decreased deceptive responses by the respondents compared to similar data that was collected in a previous study.

The results are not promising. Although the LVA instrument tended to perform better than the CVSA instrument, both programs failed consistently to correctly identify respondents who were being deceptive. On average, only about 15% of the respondents who recently used drugs but reported that they had <u>not</u> used drugs were identified as being deceptive. For CVSA, this figure was 8% while it was greater than 20% for LVA. On the other hand, while over 90% of the non-deceptive respondents were correctly classified, that meant that almost 10% of non-deceptive respondents were <u>in</u>correctly classified. The False Positive Index (the ratio of false positive to true positives) for the VSA programs averaged 9.4 for all drugs. For the CVSA instrument, this number was much higher (averaging 14.1 for all drugs) while the FPI for the LVA instrument averaged about 5 for all drugs. The only significant relationship that was observed was the use of the LVA instrument to indicate deception for recent opiate and PCP use. Those findings deserve a caveat, however, since so few respondents used the drugs resulting in a small number of deceptive responses that were available to be detected. The important finding is that the VSA programs do not appear to provide any probability of detecting deception than chance.

These findings add to the growing literature on tests of voice stress analysis theory and devices. Even though early tests of the "theory" suggested that stress is related to measurable changes in voice patterns (Cestaro, 1996; Smith, 1977; Hansen, 1996; Hansen and Zhou, 1999; Haddad, Ratley, Walter, and Smith, 2002), it is not clear that VSA devices and software are able to distinguish stress from efforts to deceive (Haddad *et al*. 2002). We were unable to find any peer-reviewed and published studies that showed significant support for the effectiveness of VSA software to detect deception. All previously published research conducted in a lab setting has failed to find support for VSA theory or technology (Brenner, Branscomb, and Schwartz, 1979; Cestaro and Dollins, 1996; Hollien, Geison, and Hicks, 1987; Horvath, 1978, 1979; Janniro and Cestaro, 1996; Lynch & Henry, 1979; O'Hair, Cody, Wang, and Chao, 1990; Suzuki, Watanabe, Takeno, Kosugi, and Kasuya, 1973; Timm, 1983; Waln and Downey, 1987). Some researchers have also tried to test VSA products "in the field" but with limited success (Palmatier, n.d., 1999; 2000). Our research therefore complements previous research by failing to find support for the VSA products in a real world (jail) setting.

In addition, the programs do not seem to have very high inter-user reliability even though the programs were relatively easy to learn and implement. The research team was able to learn all that it needed to know in a one-week training session. The ability of the research team to detect deception was about as good as (if not better than) the VSA experts. Overall, the deception detection results for sensitivity were much more similar for the CVSA experts and the novices while the results for specificity were much more similar for the LVA experts and the novices. Unfortunately, though, the ambiguity of the output resulted in very little agreement in the interpretation of “deception” between the expert and novice raters. The highest correlation between the two sets of raters was less than 0.60, far short of the standard 0.80 threshold.

We experienced two methodological problems while conducting the research. First, the CVSA expert reported that several of the charts that the research team had created showed signs of over-modulation resulting in their not being useable. The research team knew this was going to be a problem because they had a lot of trouble getting the microphone to work correctly. Unless the microphone was held directly in front of the respondent’s mouth, the CVSA instrument would not record a chart. If the microphone was held too closely to the mouth, however, the captured chart was over-modulated. That said, the research team made as many calls about deception as possible, resulting in nearly twice as much data for the novices as from the CVSA expert. Importantly, this problem bodes poorly for the “ease of use” question in the real world. Unless a the microphone technology and the computer program do not become more compatible, then users face the real possibility of creating CVSA charts that are not interpretable.

Another problem was with the computer equipment. On the first day in the jail, we discovered that the LVA computer would not boot up. We were sent a used computer overnight but that problem factored into our decision to only collect CVSA data for the first 12 days (our original plan was to collect data on both machines simultaneously). Throughout the rest of data collection, however, the LVA computer proved to be very temperamental – often crashing in the middle of an interview or during the sequencing.

Finally, the LVA instrument is not technically designed to be used in the manner that we used it for this evaluation. When LVA representative demonstrate the product for potential customers, for example, they sometimes perform the deception detection test on a recording of a narrative monologue that is known to be false. A classic example that was used in the research team's training included former President Bill Clinton's statement "Now, I have to go back to work on my State of the Union speech. And I worked on it until pretty late last night. But I want to say one thing to the American people. I want you to listen to me. I'm going to say this again. I did not have sexual relations with that woman, Miss Lewinsky. I never told anybody to lie, not a single time; never. These allegations are false. And I need to go back to work for the American people." As a recording of that statement is played, the LVA instrument indicates when the speaker who is creating and presenting the narrative is being deceptive.

In our study, however, we were asking a series of yes or no questions. While this worked perfectly for the CVSA protocol, the LVA training staff required that we ask the respondent to reply in a conversational manner. Instead of responding "no" to the question about recent PCP use, the respondent was encouraged to answer in a full sentence as in "no, I have not used PCP in the past 72 hours." We recognize that this may have added some artificiality to the process but the LVA representatives said that was the only way to get this to work.

Even though the research team <u>specifically</u> requested that it be able to ask "yes/no" answers, the LVA trainers required the interviewers to ask the questions using a "conversational" manner. Subsequent discussions with representatives from LVA, however, suggested that they would have preferred (in retrospect) that LVA be tested in a different format, where "yes/no" answers where used instead of the "conversational" questioning that they originally required (see Letter from C. David Watson in Figure 17). After the data collection was completed, the LVA staff determined that the data collection protocol was problematic. Specifically, V-Worldwide worried that the respondents were being asked questions too quickly, creating a situation where they did not have enough time to process the question and to allow for "a feedback associated with embarrassment and deception." In addition, the LVA staff suggested that the answer pattern was artificially induced (not natural), resulting in a low likelihood that deception could be successfully indicated. As shown in the letter, the LVA staff recommended that the study be conducted again with a different protocol. To be clear, the research team did work with the developer to create what we were informed was a research design that would accomplish an acceptable test of the program. Indeed, before we began collecting data for the LVA project, we sent our protocol to our LVA contact/trainer. His reply was as follows:

> Your basic set up and questions structure should work fine. I do the same things with employee interviews. If they do come across with a yes or no question, just ask them to elaborate and they should then put it in narrative format. From your writing it appears that you doing the right thing. I do get concerned about noise, but if you are getting the noise out during the segmentation process that should not be a problem. The one issue you will have is that pausing during the interview may change the situation of the interviewee. In other words, while you are waiting their mental state may change. The best solution you can find is to not have the noise situation. But, you [sic] approach is about the best I have seen from anyone.
>
> We do use this type of narrative response in our employment interviews. I am pleased with the effort you are making in setting this up. I am confident that you will be able to do what you need to with this information. It looks like you have

> done better than anyone in setting up this information. It should work! (Personal correspondence, Thomas Winscher, March 13, 2006, emphasis added).

That said, we take seriously the recommendations of the LVA staff in our interpretation of our results and we suggest that further testing of LVA be conducted using the suggestions made by Mr. Watson. The problem, of course, is that his requirements almost make the LVA program untestable using the protocol outlined in this study.

Figure 17. Letter from V-Worldwide Regarding LVA Results

Thank you for providing us the opportunity to participate in this phase of your research. We welcome a rigorous evaluation of Layered Voice Analysis ("LVA") technology, and are committed to working with you and your team to develop a full and complete understanding of the capabilities of LVA.

As we have discussed with you, we have some concerns, which we believe you understand and acknowledge, with certain aspects of the data collection and how it impacts the ability to provide a meaningful analysis of the data. We understand that notwithstanding those concerns you would like for us to continue to participate with you to determine whether meaningful results can be obtained from the current data. We understand what one possible outcome of this phase of the research is to use what we have learned to establish better data collection protocols and redo the data collection.

The particular data collection concerns we have discussed with you are summarized below:

1. The questions are asked quickly and in a row. This creates a phenomenon associated with the lack of time to cognitively process what is being asked and creates a feedback associated with embarrassment and deception (false).

2. Speakers are being asked to answer in a "complete sentence." Because this requires the speaker to assemble a pre-structured sentence that is often inconsistent with the way the speaker normally thinks and speaks, it is not the "words" of a conversation. This generally creates some level of instability in the analysis as it makes what the speaker is saying come out in a way that is somewhat forced and unnatural. This often leads to a quick switch in the speaker's stress level that can create a false impression of deception based on stress and embarrassment. This leads to an analysis where the first segment could potentially be DI and the second segment part NDI.

3. The speaker was more confined in his or her answer and thus was not "speaking as freely" as LVA would prefer. LVA is conversational and thrives on free flow of thought and interactive responses. The format resulted in responses were sometimes unduly mechanical in nature.

4. Because the speaker knew he or she had no real accountability with respect to what was said, the speaker was not required to have the intent that is necessary to realize fully the impact of the statement.

5. There is a timing concern with the data collection. Is the time-line off in the speakers' accountability? Do the speakers have real time to process their responses and put a proper timeline in place?

Figure 17 Continued.

6. We also had some technical concerns with how the data was collected and assembled. We recommend that you never change the BG setting (rather keep it as what the LVA designated), do not use shift mix, do not unnaturally merge segments and keep the focus on segments that are clearly answering (“I have not”). In segmenting you should clean-up the inarticulate noise (“ummms,” “ahhhs” and bangs) but be careful not to delete too much. In addition, using the shift + mix button employs a higher stress level overall and can be too much data for the segment.

Because of the protocol and how the questions where asked, be careful as the responses are more indicative to the responses of a polygraph. Because the protocol did not employ control questions or questions to “bring the guy down” there are some unusual phenomenons (sic) that are occurring such as DI that is not really a DI. Rather it is stress and embarrassment, caused by the factors above, getting in the way accompanied with emotional reactions (increase in SPT) supported with a decrease in the cognitive levels (SPJ).

Although we will work with you to see what can be gleaned from the data already collected, our recommendation is that the data collection be repeated as follows:

1. Work with the developer to set-up a design that will harvest the information needed without all the side-effects that have been created.
2. Create a quieter work environment.
3. Do not have quick repetitive answers that are pre-structured.

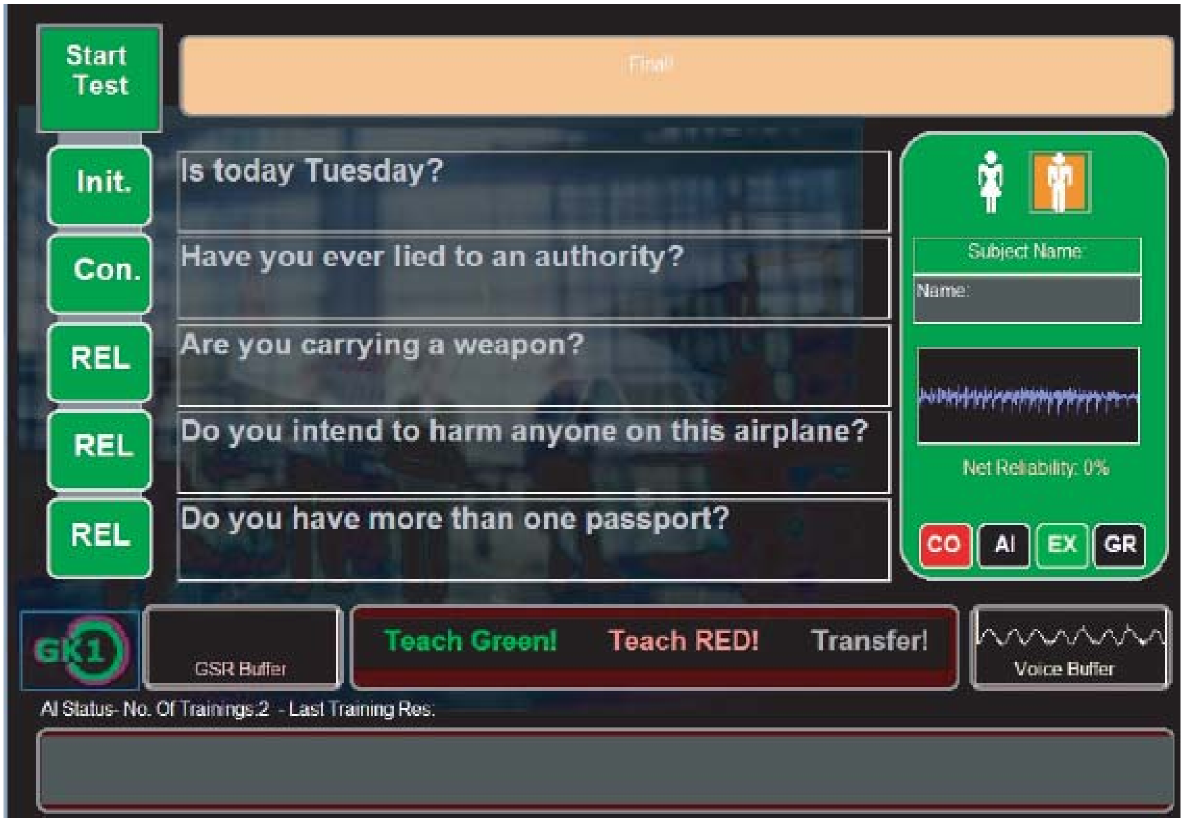
We look forward to continuing to work with you on this study.

Best regards,

C. David Watson
Chief Operating Officer

Interestingly, after the data collection was completed, the research team discovered that LVA had developed a different program that may have been more appropriate for this research. “Gate Keeper” is a program that is designed to assess deception to a short series of “yes” and “no” answers - just like the protocol used in this research project (see Figure 18). The program is designed to be used in airports. It is not clear why the Gate Keeper technology was not provided to the research team instead of LVA (if indeed the program would have been more appropriate). Future tests using this protocol should consider using Gate Keeper instead of LVA.

Figure 18. V-Worldwide Demonstration of How LVA Technology could be Adapted for Airport Security (Gate Keeper).



Is today Tuesday?

Have you ever lied to an authority?

Are you carrying a weapon?

Do you intend to harm anyone on this airplane?

Do you have more than one passport?

Subject Name

Name:

Net Reliability: 0%

AI Status- No. Of Trainings:2 - Last Training Res:

Source: http://www.v-lva.com/newsite/downloads/GK-1_Brochure.pdf (Accessed December 12, 2006)

Finally, our study showed that using VSA instruments in the jail setting resulted in respondents being less likely to be deceptive about recent deviant behavior (i.e., illicit drug use) compared to a situation where the results are simply being recorded by pen and paper. This finding has important implications for social scientists and for law enforcement. For social scientists, this “bogus pipeline” effect is a methodologically interesting sidebar that could be used to encourage more honest responses from subjects. From a law enforcement point of view, the results of the bogus pipeline effect show why law enforcement officers are likely to label VSA instruments as “successful” because they have been able to solve a case because the suspect confessed to the crime while undergoing a VSA interview. Leaving aside for the moment the potential problems associated with false confessions (Wylie, 2001), our results show that VSA instruments are likely to result in less deceptive responses under certain conditions.

Finally, we address the practicality of the use of VSA outside of the police interview setting. As the VSA programs that we tested stand now, it is not clear that these products would be useful in a setting such as an airport. We conducted the research in a relatively quiet room and we still had problems obtaining CVSA charts that were not “too loud” or over-modulated. The LVA requirement that the deception detection be conducted on conversational responses also makes the software impractical for “live” use in an airport setting. In addition, LVA requires a large amount of post-interview work to segment the responses. This also makes the instrument (as tested) impractical for use in a public setting such as an airport.

Products. The deception data that were collected for this study are available in SPSS format. Additional tests of the VSA data can be conducted using the data but the data will also supplement the sparse amount of data that are currently available for VSA studies. “The biggest challenge [with doing VSA research is] the unavailability of sufficient deceptive stress data with ground truth" (Haddad, Ratley, Walter, and Smith, 2002:19). Researchers who are interested in studying VSA are required to find audio data that are have ground truth (some criteria that allows the researcher to know if the person was telling the truth). Often, this is done by asking police departments to provide audio tapes of suspect interviews along with identification of those who were lying and those who were telling the truth. The ethical issues of using such data aside, these data are also faulty because of the variety of recording qualities and the variable amount of information that is available on the tape. In addition, the police may not know who is being deceptive. A respondent might be telling the truth but is later convicted. A respondent might be lying but is never convicted. Thus, using these data is extremely unreliable. Researchers have also used the SUSAS (Speech Under Simulated and Actual Stress) database to conduct VSA research, but there is no ground truth available for these data (Hansen, 1996). On the other hand, the data collected in this study are available to researchers who wish to test different devices (or other theories) "off-line." The data provide audio clips (in a *.wav file) with the ground truth (deceptive or not).

Caveats. In addition to the problems discussed above, we feel it is important to note other short-comings of our study. A key element of our validity study, for example, was the comparison of respondent answers to questions about recent drug use to a urinalysis test. The assumption, of course, is that a drug test is a 100% accurate way of testing drug use. While we know that is not the case, our drug lab reported that their accuracy rate was nearly 100%, so that adds to our confidence. That said, we are concerned about the number of people who said that they used marijuana in the past 30 days but tested negative for marijuana. It may be that asking about a relatively minor occurrence (for some people) over such a long period of time is just too much to expect. We dropped those individuals from many of our analyses but we wonder how many of those deceptive respondents who said they had not used but tested positive simply forgot. It might be that future tests of VSA instruments using this protocol should focus on drugs with a shorter metabolism rate (e.g., cocaine, opiates, PCP, and methamphetamine). Future studies should also consider NOT informing the respondents about the purpose of the experiment until after the data collection has been completed. IRB's regularly approved such deception as long as the subjects are debriefed after the data are collected. This would eliminate problems associated with allowing the subjects to know what questions are being asked in advance. It would also remove any potential testing effect in the study.

In addition, the two sources of data were not collected at the same time – CVSA data were collected first followed by LVA. There was a potential that the sample of arrestees who booked during the two different time periods were different or that the participation rate would change over time. In addition, the interviewers may have improved their technique over time so that they were more "comfortable" when the LVA project started compared to the CVSA project. Our attempt to guard against a possible "maturation" problem was to have one interviewer conduct all of the CVSA interviews while the second interviewer conducted all of the LVA interviews. Both interviewers were present at all times to provide assistance. Our hope was that the high experience level of the interview team would buffer any other problems with sample comparability but the protocol itself might have affected the similarity of the subjects. To check for any such differences, we conducted preliminary analysis (not shown here) that compared rates of deception, percent of subjects who tested positive for drugs, and demographics. The results showed that there were no significant differences between the CVSA and LVA samples.

We also acknowledge that there is potential for confusion between “opiate” use and “heroin” use since someone might not have used heroin but could still test positive for opiates. Future studies would do well to take this problem into account. We also recognize that VSA and polygraph technology are normally used to detect deception for single incident (e.g., one murder, a specific theft). Our test included a series of questions about recent use of five different drugs might not be the perfect test. Finally, we recognize that, even though many respondents seemed compelled to lie, our protocol did not (and could not) create the level of jeopardy that would exist in a typical interview between a police officer and a suspect. That said, we believe that a level of jeopardy (in the form of social desirability) did exist in the study since many respondents lied even though they knew we were going to perform a drug test, making this the first such study that was able to test these products in a live justice setting. We came as close to a “real world” situation as might be possible.

6. Bibliography/References.

ADAM. 2003. *Annualized Site Reports: 2002*. Washington, DC: National Institute of Justice.

Aguinis, H., C. Pierce, and B. Quigley. 1993. “Conditions under which a bogus pipeline procedure enhances the validity of self-reported cigarette smoking: a meta-analytic review.” *Journal of Applied Social Psychology*, 23(5), 352-373.

Akers, R., J. Massey, W. Clarke, and R. Lauer. 1983. “Are self-reports of adolescent deviance valid? Biochemical measures, randomized response, and the bogus pipeline in smoking behavior.” *Social Forces*, 62(1), 234-251.

Barland, G. 2002. “The Use of Voice Changes in the Detection of Deception.” *Polygraph*, 31(2), 145-153.

Botvin, E., G. Botvin, N. Renick, A. Filazzola, and J. Allegrante. 1984. “Adolescents’ self-reports of tobacco, alcohol, and marijuana use: Examining the comparability of video tape, cartoon and verbal bogus-pipeline procedures.” *Psychological Reports*, 55, 379-386.

Brenner, M., Branscomb, H., and Schwartz, G. 1979. “Psychological stress evaluator: Two tests of a vocal measure.” *Psychophysiology*, 16(4), 351-357.

Brown, T., S. Senter, and A. Ryan. 2003. *Ability of the Vericator to Detect Smugglers at a Mock Security Checkpoint*. Report No. DoDPI03-R-0002. Fort McClellan, AL: Department of Defense Polygraph Institute.

Bruck, S. No date. The Trusterpro Technology Reliability Test. Unpublished manuscripts accessed on November 17, 2006 at http://www.digilog.org/ppt_pages/The%20TrusterPro%20 Technology%20Reliability%20Test%20-%20Shlomo%20Bruck.pdf .

Campanelli, P., T. Dielman, and J. Shope. 1987. “Validity of adolescents' self-reports of alcohol use and misuse using a bogus pipeline procedure.” *Adolescence*, 22(85), 7-22.

Cestaro, V. 1996. “A comparison between decision accuracy rates obtained using the polygraph instrument and the Computer Voice Stress Analyzer (CVSA) in the absence of jeopardy.” *Polygraph*, 25(2), 117-127.

Cestaro, V. 1995. A Comparison Between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy. (DoDPI95-R-0002). Fort McClellan, AL: Department of Defense Polygraph Institute.

Cestaro, V. and Dollins, A. 1996. “An analysis of voice responses for the detection of deception.” *Polygraph*, 25(1), 15-342.

Cohen, J. 1960. “A coefficient of agreement for nominal scales.” *Educational and Psychological Measurement*, 20(1), 37-46.

Cohen, J. 1968. “Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit.” *Psychological Bulletin*, 70(4), 213-220.

Computerized Voice Stress Analysis Manual. 2006. West Palm Beach, FL: National Institute for Truth Verification.

Crites, J. 2005. *Aviation Security: Identifying Threats and Assessing Feasibility of Responses*. January 9, 2005 to the Aviation Group of the Transportation Research Board.

Fendrich, M., and Y. Xu. 1994. "Validity of drug use reports from juvenile arrestees." *International Journal of the Addictions*, 29(8), 971-985.

Fleiss, J. 1971. "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, 76(5), pp. 378--382

Fuller, B.F. 1984. "Reliability and validity of an interval measure of vocal stress." *Psychological Medicine*, 14(1), 159-166.

Haddad, D., Ratley, R., Walter, S., and Smith, M. 2002. *Investigation and Evaluation of Voice Stress Analysis Technology, Final Report*, NCJ Number: 193832. Washington, DC: National Institute of Justice.

Hansen, J. 1996. "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition." *Speech Communications, Special Issue on Speech Under Stress*, 20, 151-173.

Hansen, J. and Zhou, G. 1999. *Methods for Voice Stress Analysis and Classification, Final Technical Report*. Rome, NY: US Air Force Research Laboratory.

Harrison, L. and Hughes, A. 1997. *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. NIDA Research Monograph, Number 167. Washington, DC: NIDA.

Harrison, L. 1995. "The validity of self-reported data on drug use." *Journal of Drug Issues,* 25, 91-111.

Hollien, H., Geison, L., Hicks, J. 1987. "Voice stress analysis and lie detection." *Journal of Forensic Sciences*, 32(2), 405-418.

Hopkins, C., Ratley, R., Benincasa, D., and Grieco, J. 2005. "Evaluation of Voice Stress Analysis Technology." Proceedings of the 38th Hawaii International Conference on System Sciences.

Horvath, F. 1978. "An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception." *Journal of Applied Psychology*, 63(3), 338-344.

Horvath, F. 1979. "Effect of different motivational instructions on detection of deception with the psychological stress evaluator and the galvanic skin response." *Journal of Applied Psychology*, 64(3), 323-330.

Hser, Y. 1997. "Self-reported drug use: Results of selected empirical investigations of validity." *NIDA Research Monograph*, 167, 320-343.

Janniro, M. and Cestaro, V. 1996. "Effectiveness of detection of deception examinations using the Computer Voice Stress Analyzer." *Polygraph* 27(1), 28-34.

Johnson, R. 2004. "Lie-detector glasses offer peek at future of security." *EE Times*, January 16, accessed on December 1, 2006 at http://www.eetimes.com/story/OEG20040116S0050.

Jones, E. and H. Sigall. 1971. "The bogus pipeline: a new paradigm for measuring affect and attitude." *Psychological Bulletin*, 76, 349-364.

Kubis, J. F. 1973. *Comparison of Voice Analysis and Polygraph as Lie Detection Procedures*. (Technical Report No. LWL-CR-03B70, Contract DAAD05-72-C-0217). Aberdeen Proving Ground, MD: U.S. Army Land Warfare Laboratory.

Lippold, O. 1971. "Physiological Tremor." *Scientific American*, 224(3), 27-34.

Lu, N., Taylor, B. and Riley, K. 2000. "The validity of adult arrestee self-reports of crack cocaine. *American Journal of Drug and Alcohol Abuse,* 27(3), 399-407.

Lynch, B. and Henry, D. 1979. "A validity study of the psychological stress evaluator." *Canadian Journal of Behavioral Science*, 11(1), 89-94.

Meyerhoff, J., Saviolakis, G., Koenig, M., Yourick, D. In Press. *Physiological and biochemical measures of stress compared to voice stress analysis using the Computer Voice Stress Analyzer (CVSA)*. Report No. DoDPI98-R-0004. Fort Jackson, SC.

Mieczkowski, T., Barzelay, D., Gropper, B., and Wish, E. 1991. "Concordance of three measures of cocaine use in an arrestee population: Hair, urine, and self-report." *Journal of Psychoactive Drugs*, 23, 241-246.

National Research Council. 2003. *The Polygraph and Lie Detection*. Washington, DC: National Academy of Sciences.

O'Hair, D., Cody, M. J., Wang, S., & Chao, E. Y. (1990). Vocal stress and deception detection among Chinese. *Communication Quarterly*, 38(2, Spring), 158ff.

Palmatier, J. 1999. The Computerized Voice Stress Analyzer: Modern Technological Innovation or "The Emperor's New Clothes"? *General Practice, Solo & Small Firm Division Magazine (16)*:224-232.

Palmatier, J. 2000. *The Validity and Comparative Accuracy of Voice Stress Analysis as Measured by the CVSA: A Field Study Conducted in a Psychophysiological Context*. Unpublished manuscript, Michigan Department of State Police.

Palmatier, J. No date. *The CVSA, Polygraph, and Trustech / Vericator Voice Analysis Technologies: Preliminary Results from a Comparative Analysis Conducted in a Criminal Justice Field Setting.* Unpublished manuscript access on December 1, 2006 at http://www.nemesysco.com/vericator.pdf .

Shrout, P. & Fleiss, J. 1979. "Intraclass correlation: uses in assessing rater reliability." *Psychological Bulletin*, 86, 420—428.

Smith, G. 1977. "Voice analysis for the measurement of anxiety." *British Journal of Medical Psychology*, 50, 367-373.

Sprangers, M. and J. Hoogstraten. 1987. "Response-style effects, response-shift bias and a bogus-pipeline." *Psychological Reports*, *61*, 579-585.

Suzuki, A., Watanabe, S., Takeno, Y., Kosugi, T., & Kasuya, T. 1973. "Possibility of detecting deception by voice analysis." *Reports of the National Research Institute of Police Science*, 26(1, February), 62-66.

Timm, H. 1983. "The efficacy of the psychological stress evaluator in detecting deception." *Journal of Police Science and Administration*, 11(1), 62-68.

Waln, R. and Downey, R. 1987. "Voice stress analysis: Use of telephone recordings." *Journal of Business and Psychology*, 1(4), 379-389.

Wylie, M. 2002. "Police Use of Voice Stress Analysis Generates Controversy." Newhouse News Service (accessed at http://polygraph.com.au/pdf/police_use_of_voice_stress_analysis_generates_controversy.pdf on December 1, 2006.

Yacoubian, G. 2000. "Reassessing the need for urinalysis as a validation technique: Correlation estimates from the Arrestee Drug Abuse Monitoring (ADAM) Program." *Journal of Drug Issues*, 30(2), 323-334.

Yule, G. 1912. "On the methods of measuring association between two attributes." *Journal of the Royal Statistical Society* 75, 576–642.

7. Appendices

a. Jail Officer/Research Associate Checklist for VSA
b. VSA Facesheet
c. CVSA Survey Form
d. LVA Survey Form
e. Verbal Consent Script
f. Summary of Drug Testing
g. Tests Comparing Validity of VSA Programs for Deceptiveness Related to All Drugs Including Those Who Reported Marijuana Use but Tested Negative
h. Table Comparing CVSA and LVA data.

APPENDIX A.
JAIL OFFICER/RESEARCH ASSOCIATE CHECKLIST FOR VSA

Please initial that you have read and understand the following guidelines.

___l Both the interview and the urine sample are voluntary. The arrestee can decline to participate altogether, or at any point during the interview process up to and including providing a urine specimen.

___2 Each shift will be staffed by two jail officers and two research associates (RA's). RA's have been trained in professional interviewing techniques and in jail protocol.

___3 RA's arrive on site approximately 30 minutes before each shift begins and obtain a list of potential participants from the booking officer/staff member on duty. During this time, one jail officer will bring a water jug from the kitchen while the other officer assists the RA with arrestee selection.

___4 The booking officer will provide the supervisor with a printout of demographic and charge information for each selected arrestee. RA's may request assistance from jail officers in resolving questions about charges at arrest and other arrestee information.

___5 The jail officer will provide the RA with a list of sampled arrestees. The jail officer should record cell locations and notify the RA if a sampled arrestee cannot be located.

___6 When the RA's are ready to begin, the arrestee should be summoned from the cell by the jail officer and brought to the interview area with minimal explanation or interaction. *The nature of the study should not be discussed*. If the arrestee asks where s/he is being taken, the jail officer should say that there is someone conducting a survey who wants to speak with him/her. The study <u>should not</u> he referred to as a "drug study" and the urine sample <u>should not</u> be mentioned. The jail officer should not say the name of the arrestee in the presence of the interviewer. Rather, the arrestee should be matched to the correct RA using the VSA identification number. The officer should announce the arrestee by saying, for example, "Here's case M304."

___7 Occasionally, a selected arrestee may tell the jail officer that s/he does not want to participate in the study. The jail officer should escort the arrestee to the interview area and explain that s/he must decline directly to the RA. That said, this is a voluntary study and no inmate should be forced to meet with the RA. If a jail officer finds that an arrestee is too ill to leave the cell or is otherwise indisposed, the officer should tell the RA's why the arrestee could not be brought to an interviewer.

___8 Jail officers should screen out arrestees who present a security risk (e.g., obviously violent or too sick to move). Interviewers are trained to terminate the interview if the arrestee is too high on drugs or alcohol or is too aggressive or menacing to continue the interview.

___9 Jail officers should maintain sufficient distance from the interviewing area so that they cannot overhear any conversation between the interviewer and the arrestee, but remain in view of the interviewer and arrestee. We want to avoid a situation in which an arrestee feels uncomfortable answering a question honestly because s/he feels s/he can be overheard. Officers should not leave RA's alone with arrestees and should not leave the interviewing area without informing an RA.

___10 The interviews will take about 10-15 minutes plus about 5 minutes for the urine sample. Jail officers should be prepared to escort arrestees to appropriate location to provide urine samples and then return them to their cells unless RA's request further clarification on the questionnaires.

___11 Candy bars or a non-sugar substitute are provided as incentives to those who complete the interview and provide a urine specimen.

___12 If an arrestee is unable to provide a urine sample, s/he will be asked to wait until they are able to do so. Under no circumstances should the collection bottle be left with the arrestee.

___l3 Because the interview is confidential, the identity of specific arrestees and the contents of their interviews may not be discussed.

After completing this checklist, jail officer and RA should sign below.

___________________________________ ______________________________
Jail Officer Research Associate

Date:_____________

APPENDIX B.
VSA FACESHEET

INTERVIEW IDENTIFICATION

ID1.	Respondent ID #	(LABEL)
ID2.	Data Collection Date	___________/__________/20 __ __

ARREST RECORD

AR1.	Arrest Date	___________/__________/20 __ __
AR2.	Arrest Time	________________________ a.m. l p.m **[CIRCLE ONE]**
AR3.	a) Arrest location and b) Zip	a)_________________________b)_____________________
AR4.	Respondents birthdate	________________/______________/19__ __
AR5.	Gender [1=Male, 2=Female, 3=Other]	______________
AR6.	Race/Ethnicity 1= White 2= Black or African American 3= Hispanic/Latino 4= American Indian or Alaskan Native 8= Native Hawaiian or Other Pacific Islander 9= Asian 6= Other [Specify} ____	______________
AR9.	Respondent address Zip	______________________
AR10.	3 Most serious arrest charges: Severity:	1=Felony, 2-Misdemeanor, 3=Traffic Violation, 4=Municipal/Summary, 5=Probable Cause, 6=OTHER [Specify} **WARRANT**; 0=No, 1=Yes

	A) PENAL CODE	B) OFFENSE	C) CODE	D) SEVERITY	E) WARRANT
1					
2					
3					

AR11.	Booking Date	___________/__________/20 __ __
AR12.	Booking Time	________________________ a.m. l p.m. **[CIRCLE ONE]**
AR13.	Sample source [1=Stock, 2=Flow, 3=Other]	______________
AR14	Interviewer ID# (of person filling out this side of facesheet)	______________

APPENDIX C.

CVSA SURVEY FORM

Date________________________ ID #_______________________

Sample Stock Flow

1. Y N Are we in the state of Oklahoma?

2. Y N Is the color of the wall_____?

3. Y N Am I sitting down?

4. Y N OTWWD: Have you used Marijuana in the past 30 days?

5. Y N Am I wearing glasses?

6. Y N OTWWD: Have you used cocaine in the past 72 hours?

7. Y N Are you sitting in a chair?

8. Y N Have you ever driven over the posted speed limit?

9. Y N Am I wearing a watch?

10. Y N OTWWD: Have you used Heroin in the past 72 hours?

11. Y N Are we in the city of Oklahoma City?

12. Y N OTWWD: Have you used Methamphetamine in the past 72 hours?

13. Y N Is this the month of Feb/March?

14. Y N OTWWD: Have you used PCP in the past 72 hours?

15. Y N Are the lights on in this room?

APPENDIX D.

LVA SURVEY FORM

Date________________________ ID #_______________________

Sample Stock Flow

1. Y N Have you used Marijuana in the past 30 days?

2. Y N Have you used cocaine in the past 72 hours?

3. Y N Have you used Heroin in the past 72 hours?

4. Y N Have you used Methamphetamine in the past 72 hours?

5. Y N Have you used PCP in the past 72 hours?

APPENDIX E.
VERBAL CONSENT SCRIPT

Hi, my name is ______________________. (first name only)

This is an Oklahoma Department of Mental Health and Substance Abuse Services directed project designed to collect information on drug use among individuals who have been arrested. Your participation is voluntary. There is no benefit or risk associated with your participation. I work for KayTen Research and the information you provide is confidential, unavailable to anyone outside the research project, so it will not help or hurt your case. I will ask you some questions that will take 10-20 minutes. Your responses will be analyzed by the technology but your name will not be written on any of our paperwork or on the data. Your responses will be recorded to test the how well voice stress analysis determines stress. You may find some questions embarrassing or distressing, but you can refuse to answer any question. At the end of the interview, I will ask you to provide a urine sample. There will be no way to identify you from this urine sample. If you listen to all my questions and provide the urine sample, then you will be given a candy bar or sugar-free substitute. Can we begin now?

Interview Status

IN1. a) Interviewer initials and b) ID# a)________________________ b) ___________________

IN2. a) Interview status a) __________________________

1= AGREED TO INTERVIEW **[SKIP TO IN3]**
2= DECLINED TO INTERVIEW
3= NOT AVAILABLE

b) Reason b) __________________________

1=DID NOT WANT TO
2=TAKEN TO COURT
3= RELEASED
4= TRANSFERRED
5= MEDICAL UNIT
6= VIOLENT OR UNCONTROLLED BEHAVIOR
7= PHYSICALLY ILL
8= LANGUAGE [SPECIFY]
9= SHIFT ENDED ____
10=OTHER [SPECIFY ____

[IF DID NOT AGREE, DISCONTINUE INTERVIEW]

IN3. How many hours ago were you arrested? **[RECORD NUMBER OF HOURS]** ___________________

[IF GREATER THAN 48 HOURS, DICONTINUE INTERVIEW]

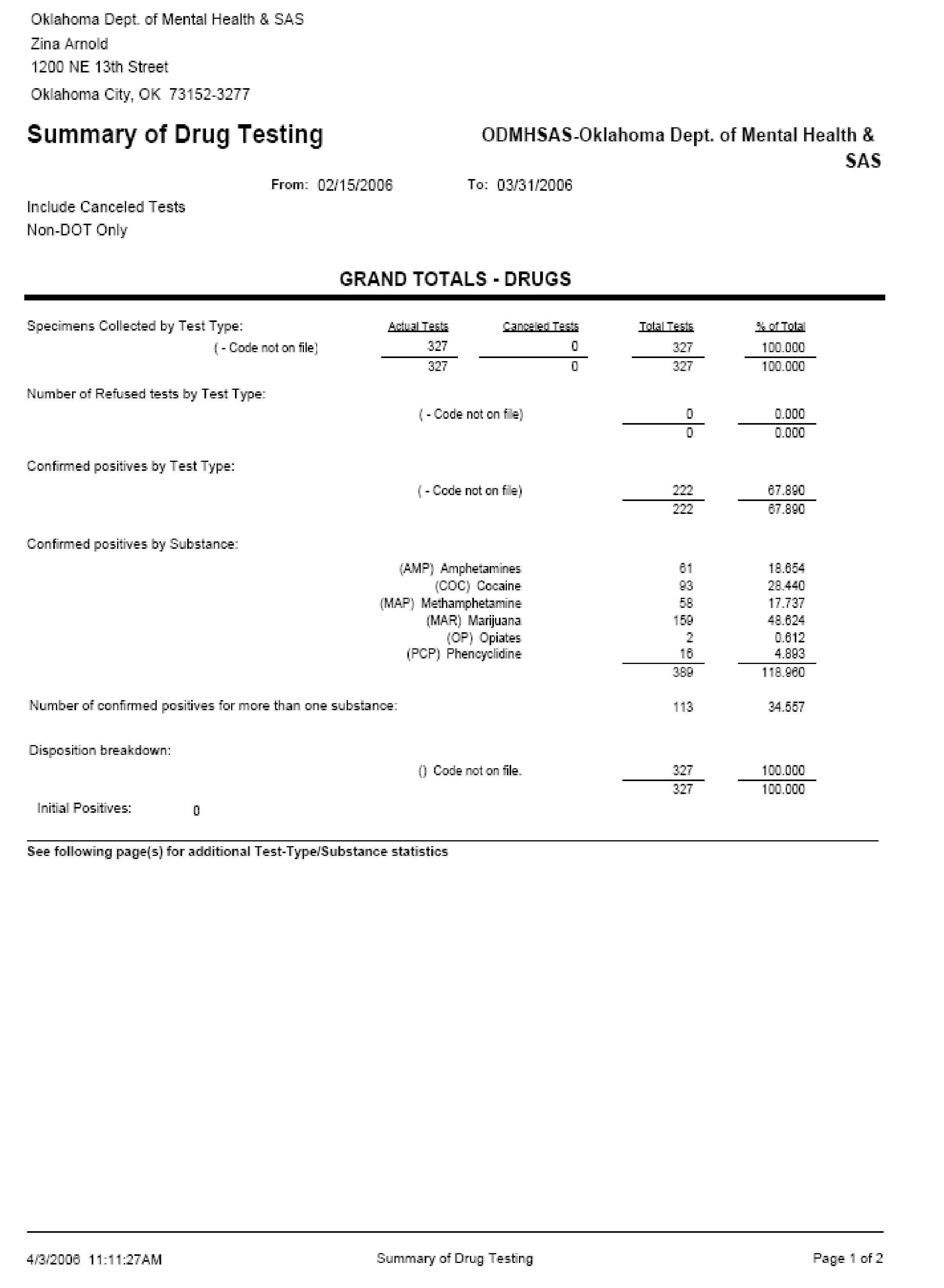
IN4. Interview a)start time, b) end time

a) __________________________ p.m.

b) __________________________ p.m.

A.

Appendix E. Summary of Drug Testing



Oklahoma Dept. of Mental Health & SAS
Zina Arnold
1200 NE 13th Street
Oklahoma City, OK 73152-3277

Summary of Drug Testing

ODMHSAS-Oklahoma Dept. of Mental Health & SAS

From: 02/15/2006 **To:** 03/31/2006

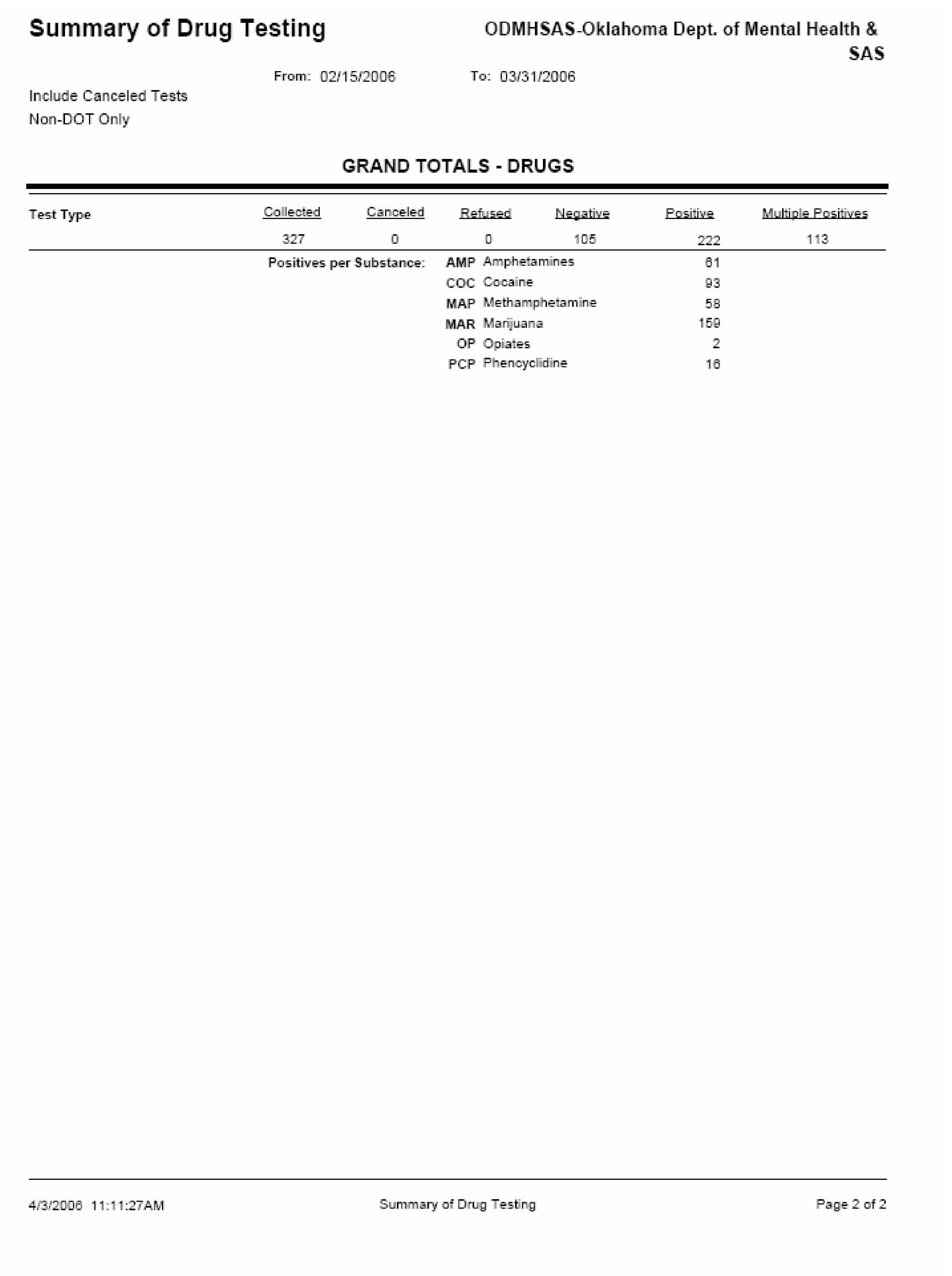
Include Canceled Tests
Non-DOT Only

GRAND TOTALS - DRUGS

		Actual Tests	Canceled Tests	Total Tests	% of Total
Specimens Collected by Test Type:					
	(- Code not on file)	327	0	327	100.000
		327	0	327	100.000
Number of Refused tests by Test Type:					
	(- Code not on file)			0	0.000
				0	0.000
Confirmed positives by Test Type:					
	(- Code not on file)			222	67.890
				222	67.890
Confirmed positives by Substance:					
	(AMP) Amphetamines			61	18.654
	(COC) Cocaine			93	28.440
	(MAP) Methamphetamine			58	17.737
	(MAR) Marijuana			159	48.624
	(OP) Opiates			2	0.612
	(PCP) Phencyclidine			16	4.893
				389	118.960
Number of confirmed positives for more than one substance:				113	34.557
Disposition breakdown:					
	() Code not on file.			327	100.000
				327	100.000
Initial Positives:	0				

See following page(s) for additional Test-Type/Substance statistics

4/3/2006 11:11:27AM Summary of Drug Testing Page 1 of 2



Summary of Drug Testing

ODMHSAS-Oklahoma Dept. of Mental Health & SAS

From: 02/15/2006 To: 03/31/2006

Include Canceled Tests

Non-DOT Only

GRAND TOTALS - DRUGS

Test Type	Collected	Canceled	Refused	Negative	Positive	Multiple Positives
	327	0	0	105	222	113

Positives per Substance:

AMP	Amphetamines	61
COC	Cocaine	93
MAP	Methamphetamine	58
MAR	Marijuana	159
OP	Opiates	2
PCP	Phencyclidine	16

4/3/2006 11:11:27AM Summary of Drug Testing Page 2 of 2

Appendix G. Tests Comparing Validity of VSA Programs for Deceptiveness Related to All Drugs Including Those Who Reported Marijuana Use but Tested Negative

Table G1. Interpretation of deception about any drug by all respondents.

Condition	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	36 31.6%	52 25.6%	88
2. No Deception Indicated	78 68.4%	151 74.4%	229
Total Sample	114 36.0%	203 64.0%	317

Chi Square = 1.294 (*p* = 0.255)
Sensitivity = 31.6%
Specificity = 74.4%
False Positive Index = 1.44
Correctly Classified: 59.0%

Table G2. Interpretation of deception about any drug by all respondents for CVSA.

Condition	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	19 32.2%	38 35.2%	57
2. No Deception Indicated	40 67.8%	70 64.8%	110
Total Sample	59 35.3%	108 64.7%	167

Chi Square = 0.151 (*p* = 0.698)
Sensitivity = 32.2%
Specificity = 64.8%
False Positive Index = 2.0
Correctly Classified: 53.3%

Table G3. Interpretation of deception about any drug by all respondents for LVA.

Condition	Deceptive Response	Non Deceptive Response	Total
1. Deception Indicated	17 30.9%	14 14.7%	31
2. No Deception Indicated	38 69.1%	81 85.3%	119
Total Sample	55 36.7%	95 63.3%	150

Chi Square = 5.557 (*p* = 0.018)
Sensitivity = 30.9%
Specificity = 85.3%
False Positive Index = 0.82
Correctly Classified: 65.3%

Appendix H. Table Comparing CVSA and LVA data.

		Group Statistics	
	Expert-Type	N	Mean
Respondent's age	CVSA	159	33.95
	LVA	148	33.72
Lied about mar - said no but tested positive	CVSA	86	0.10
	LVA	68	0.12
Lied about coc - said no but tested positive	CVSA	45	0.42
	LVA	44	0.50
Lied about her - said no but tested positive	CVSA	0	0.00
	LVA	2	1.00
Lied about pcp - said no but tested positive	CVSA	7	0.57
	LVA	9	0.56
Lied about meth - said no but tested positive	CVSA	33	0.33
	LVA	24	0.46
Test Positive For Any Drug	CVSA	168	0.68
	LVA	151	0.67
Marijuana Positive Test	CVSA	168	0.51
	LVA	151	0.45
Cocaine Positive Test	CVSA	168	0.27
	LVA	151	0.29
Opiates Positive Test	CVSA	168	0.00
	LVA	151	0.01
Amphetamine Positive Test	CVSA	168	0.21
	LVA	151	0.17
Methamphetamine Positive Test	CVSA	168	0.20
	LVA	151	0.16
PCP Positive Test	CVSA	168	0.04
	LVA	151	0.06