2

Credibility Assessment of Verbatim Statements (CAVS)

Final Report of Office of Naval Research Grant No. N00014-92-J-4006

Charles R. Honts, Ph. D.
Principal Investigator

and

Mary K. Devitt
Research Assistant

15 October 1993

University of North Dakota
Psychology Department
P.O. Box 8380
Grand Forks, ND 58202-8380

Credibility Assessment of Verbatim Statements (CAVS)

Final Report of Office of Naval Research Grant No. N00014-92-J-4006

Charles R. Honts, Ph. D.
Principal Investigator

and

Mary K. Devitt
Research Assistant

15 October 1993

University of North Dakota
Psychology Department
P.O. Box 8380
Grand Forks, ND 58202-8380

The University of North Dakota

EXECUTIVE SUMMARY

We examined the possibility of detecting deception with a statement analysis technique that we called Credibility Analysis of Verbatim Statements (CAVS). Twenty-six subjects were interviewed about their work history by an experienced interviewer. Subjects were previously instructed to tell one true story about a job that they had held and to tell one false story about a fictitious job. Subjects were motivated to tell convincing lies by the offer of a monetary reward if they fooled the interviewer with their false statement. The interviewer was no better than chance at discriminating the true and false statements and was correct with only 13 of the 26 subjects. The subjects' interviews were videotaped and were then transcribed. Fifteen Content and eleven Structural criteria were then scored in the statements by trained evaluators who had reference only to the transcripts.

The data were analyzed with discriminant analysis and logistic regression. Logistic regression proved to be the more powerful approach. A logistic regression solution using 9 of the criteria correctly classified 78.85% (40 of 52) of the statements. This was considered good performance and is comparable to the results of polygraph examinations conducted in similar situations. However, when the results of the logistic regression were applied on a within-subjects basis, the statements of 24 of the 26 subjects were correctly categorized as either true or false. One subject's statements were not classifiable, and one subject's statements were misclassified.

The results of this study suggest that CAVS can be a very powerful tool for assessing the credibility of statements made by adults. The within-subjects discriminations were particularly powerful. Given the need for flexible and effective credibility assessment in the national security system, we believe that CAVS is a good candidate for application and deserving of additional research and development.

BACKGROUND

The information obtained through the statements of applicants, their references, and other human intelligence (HUMINT) sources are a central part of the security clearance process. However, the veracity of such statements must always be suspect. An applicant may lie to hide derogatory or embarrassing information in her/his background or he/she might lie to hide the true purpose of his/her application for clearance. Similarly, references and other information sources may give false statements for any of a wide variety of reasons. Other HUMINT sources may lie for reasons that are all too obvious. To the extent that any of these falsehoods are accepted as veridical, the security clearance process may be compromised, perhaps with disastrous consequences.

A number of approaches have been taken to safeguard the clearance process from the effects of false information given by applicants and other HUMINT sources. Objective measures of such things as financial behavior can be useful in some cases, but their application is limited. What is needed is a method to directly assess the credibility of the information given by human sources. For many years polygraph testing has been used to assess the credibility of applicants for security clearance, to periodically assess the veracity of persons holding security clearances, to verify HUMINT information, and to investigate suspected security breaches. However, the utility of polygraph tests is severely limited by several factors. First, because of their nature, polygraph tests are limited in the number and types of issues that can be addressed. False negative outcomes could result because the subject of the polygraph test may simply have not been asked appropriate questions. Second, polygraph testing is expensive in terms of resources and is cumbersome in its implementation. An examiner and equipment are necessary. The subject must agree to the testing, and must cooperate with that testing. A polygraph test covering only a few issues will take at least two hours, and may often take longer. Third, recent research conducted at the Department of Defense Polygraph Institute indicates that polygraph tests are not very good at detecting deception in security screening situations. Barland, Honts, & Barger (1989) tested the validity of the security screening polygraph tests of four federal agencies and found that the best of them barely detected half of the deceptive subjects. The worst agency detected less than 10% of the deceptive subjects. On blind evaluation, only one of the agencies in the Barland et al. study produced better than chance discrimination of truthful and deceptive subjects. Those results were extended in another laboratory study reported by Honts (1992). Honts found that the Department of Defense's Counterintelligence Scope Polygraph test was a poor discriminator of subjects innocent and guilty of committing an act of mock espionage. Moreover, Honts (1991) conducted an analysis of field data from the Counterintelligence Scope Polygraph program. His results suggest that the CSP program detects only about 2% of the deceptive subjects. Finally, polygraph tests are readily susceptible to countermeasures. Research by Honts and his colleagues has clearly shown that polygraph tests can be defeated by relatively simple physical and mental countermeasures after brief periods of training (Honts, Hodes, & Raskin, 1985; Honts, Raskin, & Kircher, 1987, 1993; Winbush, 1993). Considering these limitations, polygraph tests are, at best, a small part of the solution to the credibility assessment of human information sources.

Several psychological techniques, new to English speaking countries, suggest themselves as possibly useful in assessing the credibility of HUMINT source's statements. Statement Validity Assessment (SVA), originally called Statement Reality Analysis, is a technique developed in Germany for assessing the credibility of child witnesses. For over 30 years, Statement Analysis of the credibility of child witness statements has not only been admissible in the German courts, it is often mandated by those courts in cases where there is a disputed allegation of child sexual abuse (Undeutsch, 1989). SVA has just recently been introduced to the English speaking world by Raskin and his colleagues (see the review by Raskin & Esplin, 1991). SVA assesses the credibility of a free narrative statement by formally examining the motivational setting of the accusation and by scoring the specific contents of the statements using a scoring system know as Criteria-Based Content Analysis ([CBCA] Steller & Koehnken, 1989). The notion underlying CBCA is that statements recalled from true memory will differ qualitatively from statements produced as a deliberate falsehood, or from fantasy. This notion is known as the Undeutsch Hypothesis, after the originator of the technique (Steller, 1989). However, despite its long history of successful application in Germany, CBCA has been the subject of very little formal research. What little research exists has been recently reviewed by Horowitz (1992), and the reliability and validity of CBCA appears very promising.

Unfortunately, to date, there has been little study of the extension of CBCA to adult statements. Landry and Brigham (1992) conducted a study where college students were given brief training in CBCA. They were then shown videotapes or given transcripts of people making true or false statements about a personal traumatic event. Subjects who received the brief training and saw the videotape were significantly more accurate in their classifications of statements then were naive subjects or subjects who received only the written statements. Although these results are somewhat supportive of CBCA, and they have implications for the criminal justice settings, they have little to say about the effectiveness of trained evaluators and credibility assessment with adults. To our knowledge there are no other studies of the validity of CBCA applied to statements given by adults.

However, there are other statement-based schemes for assessing the credibility of adult statements. Sapir (1988) has developed a technique for assessing the credibility of adult statements that is based on structural and content criterion analysis of the source's statement. This technique has become very popular with law enforcement in the U.S. and in Canada. The Royal Canadian Mounted Police .([RCMP] Kaster, 1991) have extended and developed this structural statement analysis and they report a considerable amount of anecdotal success with the technique (personal communication, Kaster, August 1993). However, to our knowledge, no research has ever been published on the reliability or validity of structural statement analysis applied to adults.

The present study examined the reliability and validity of statement analysis as applied to adult sources in a laboratory experiment. In the present study, college students gave two narrative statements about jobs that they might have held. They were instructed to give one of the statements as a true statement. They were also instructed to give another statement as a false statement about a fictitious job that they never held. The obtained statements were then transcribed and we extracted both content criteria and structural criteria from those statements for statistical analysis.

METHOD

Subjects

The subjects were 13 male and 13 female college student volunteers recruited from the Introductory Psychology classes at the University of North Dakota. They received extra credit toward their final grades for participating in the study. In addition, subjects were paid a cash bonus of $25.00 if the interviewer was not able to tell which of the two statements they made was the false statement.

Interviewer

All of the interviews were conducted by the same interviewer. The interviewer was a male psychologist. He held the Ph. D. degree in Experimental Psychology. The interviewer was also an experienced polygraph examiner and he had a great deal of experience in conducting both forensic and employment related polygraph examinations. In addition, the interviewer had received training in the Statement Validity Assessment interview technique at a number of workshops over the past 8 years. The interviewer also had received professional training in psychological interviewing.

Procedure

Subjects initially volunteered via sign-up sheets posted in the common area of the Psychology Department. The sign-up sheet specified that potential subjects must have had at least one full time job. Potential subjects were then contacted by phone by one of the experimenters. During the phone interview, potential subjects were read the following:

> This is a study of the ability of an experienced interviewer to determine if the interviewee is telling the truth about his/her previous work experience. If you agree to participate in this study you will be interviewed about your work history by an expert interviewer. During the interview the interviewer will ask you to tell, in as much detail as you can, everything you did during your

> worst day at work on each of two jobs. You will tell the interviewer one true story about a job that you actually have held. However, you will also tell the interviewer a story about a job that you never had. You will have to make that story up completely. If the interviewer cannot tell which of the stories is false you will be given a $25.00 bonus. Do you want to participate in this study?

If the subject agreed to participate, the experimenter obtained information about the job on which the subject was to base the true story. The experimenter then told the subject to work on a convincing but completely false story about a day on a job they never held. An appointment was then made for the subject to be interviewed. The interview took place no sooner than 48 hours after, but within one week of the telephone interview.

When subjects arrived for the interview they were initially met by an assistant who obtained written informed consent from each subject. The assistant then confirmed the information about the job on which the true story was to be based. This assistant also obtained information about the false story. Subjects were instructed to tell either the true or the false story first, so that half of the time the true story was told first. Subjects were then interviewed. The interviews were video-taped in their entirety. The interviewer used open ended questions to obtain as much free narrative as possible from the subjects about their worst day at work at two jobs. Following the conclusion of the interview, the interviewer made his best guess as to which of the stories was the false one.

The assistant informed the subject of the interviewer's decision. If the subject fooled the interviewer, arrangements were made to pay the subject. The assistant then debriefed the subject and interviewed him or her regarding the strategies she or he may have taken in his or her effort to beat the interviewer. The assistant stressed the need for subjects to be completely forthright at this time and reminded them that information obtained in the debriefing would have no impact on their receiving the monetary bonus.

Scoring

Following the interview, the tapes were transcribed in the their entirety. The transcripts were coded so that the original interviewer could evaluate the statements with CBCA. Three additional copies of the transcripts were prepared and mailed to three volunteer evaluators. One of those evaluators was trained in, and conducted scientific research on, CBCA. Both of the CBCA evaluators used a modified version of scoring procedures developed for Criteria Based Content Analysis system (Raskin & Esplin, 1991). In that system content criteria are evaluated on a three-point scale where: 0 = absent, 1 = ⁿresent, and 2 = strongly present. In CBCA, the presence of content criteria ,s considered as indicating a truthful statement. Therefore, the larger the score, the more confidence in the veracity of the statement. The CBCA scoring system has been reported to be acceptably reliable in a study conducted at the National Institutes of Health (Horowitz et al., 1993).

The two other independent evaluators were trained in and experienced with the scoring of structural criteria and they were asked to evaluate the transcripts for the presence of structural criteria. The Structural Criteria adapted from the Adult Statement Analysis approach of the RCMP represent inconsistencies in the structure of the statement. The presence of structural criteria is indicative of a false statement. The Structural Criteria were also evaluated on a three-point scale, but to reflect the nature of the Structural Criteria they were scored: 0 = absent, -1 = present, and -2 = strongly present.

Criteria

The following is a list the criteria used in this study. The Content Criteria were adapted from CBCA (Raskin & Esplin, 1991, 1992; Steller & Koehnken, 1989). The Structural Criteria were adapted from the RCMP version of Statement Analysis (Kaster, 1991). Additional information on these criteria is available in the original sources.

Content Criteria (After Raskin & Esplin, 1992, p. 279)

C1 **Logical Structure**. Is the statement coherent and the content logical?

C2 **Unstructured Production**. Are the descriptions unconstrained and the report somewhat unorganized? Are there digressions and/or spontaneous shifts of focus? This criterion requires that the account be logical.

C3 **Quantity of Details**. Are there specific descriptions of places, times, persons objects and/or events?

C4 **Contextual Embedding**. Are events placed in a spatial and temporal context? Is the action connected to other incidental events?

C5 **Descriptions of Interactions**. Are there reports of actions and reactions or conversations composed of a minimum of three elements involving at least the witness and one other person?

C6 **Reproduction of Conversation**. Is speech or conversation reported in its original form?

C7 **Unexpected Complications During Specific Incidents**. Was there an unplanned interruption or an unexpected complication during the main incident being reported?

C8 **Unusual Details**. Are there details that are unusual, but meaningful in this context?

C9 **Superfluous Details**. Are peripheral details described in connection with the alleged events that are not essential and do not contribute directly to the specific incident?

C10 **Accounts of Subjective Mental State**. Does the witness describe feelings or thoughts experienced at the time of the incident?

C11 **Spontaneous Corrections**. Were corrections offered, or information added, to material previously provided in the statement? Answers to direct questions do not qualify.

C12 **Admitting Lack of Memory**. Did the witness indicate a lack of memory or knowledge about the incident? In response to a direct question, the response must go beyond a simple "I don't know" to qualify.

C13 **Raising Doubts About One's Own Testimony**. Did the witness express concern that some part of the statement might not be believable?

C14 **Self-Deprecation**. Did the witness describe some aspect of his/her behavior related to the specific incident as wrong or inappropriate?

C15 **Attributions About The Mental States of Others**. Is there reference to the mental or emotional states of other participants?

Structural Criteria (After Kaster, 1991)

S1 **Verb Choice**. Does the choice of verbs indicate that the action took place or merely that it was indicated? Does the choice of verbs suggest hidden meaning?

S2 **Use of Connectors**. Is there a high use of connectors? Connectors often indicate editing of detail, or a need for time to make up detail.

S3 **Failing to Answer Questions**. Does the witness fail to give direct answers to questions?

S4 **Inconsistent Speech Style**. Does the style of speech change during the statement?

S5 **Changes in verb tense**. Are there changes in verb tense during the telling of the story?

S6 **Improper Use of Pronouns**. Is the use of pronouns appropriate?

S7 **Use of Generic Terms**. Are generic terms used in situations where people would normally use a specific detail?

S8 **Definite and Indefinite Articles**. Does the witness use definite articles when normal conversation would call for the use of indefinite articles, or vice versa?

S9 **Changing of Terms Used to Describe the Same Object or Person**. Do terms used to describe the same person or object change in the course of the statement?

S10 **Using Others to Attest to Your Honesty**. Does the witness make reference to others to support her/his credibility. Does the witness use others to make statements that he/she will not make oirectly?

S11 **Inconsistent Subjective Chronometry**. Is the chronometry of the statement inconsistent? That is, are certain portions of the statement longer or shorter than they should be in reference to the amount of time and the complexity of events covered?

Data Analysis Strategy

Twenty-six criteria represented too many criteria for a meaningful multivariate analysis with either discriminant analysis or logistic regression. Therefore we decided to reduce the number of predictors by the following procedure:

1. Initially, univariate statistical tests and univariate correlations with the True or False criterion were calculated.
2. Variables with correlations with the True or False criterion of less than $r = 0.15$, were eliminated from further consideration
3. The remaining variables were then subjected to a reliability analysis. Any variables that clearly hurt the reliability of the scale could then have been eliminated.
4. The remaining variables were then subjected to Logistic Regression and Discriminant Analysis.

RESULTS AND DISCUSSION

Evaluations by the Original Interviewer

At the end of each interview, the original interviewer made a decision about which of the subject's statements he believed to be the true statement. Those decisions were based on the interviewer's initial subjective impression of the quality of the statements, his impressions based on the subject's body language and demeanor, and on his overall subjective impression of the interview. The original interviewer made a correct decision on 50% (13 of 26) of the subjects. This performance represents the exact chance expectation.

Inter-rater Reliability

Individual content scores and total scores from the two independent content criteria evaluators were correlated in order to estimate the reliability of the content criteria scoring. Reliability *r*s for the scoring of the 16 individual content criteria ranged from -0.03 to 0.82, M = 0.26. Reliability of the total content scores was calculated to be, r = 0.35, p = .005. The second independent structural evaluator failed to provide a scoring. Therefore, no data are available to assess reliability of scoring the structural criteria.

The low values for the reliability of the two content criteria evaluators is surprising. Recent studies conducted at the National Institute of Child Health and Human Development (Horowitz, Lamb, Esplin, Boychuk, Krispin, & Reiter-Lavery,1993) and at the University of Utah (Anson, Golding, & Gully, 1993) have demonstrated that the scoring of these content criteria can be reliable. In the Horowitz et al. study, reliabilities for total CBCA scores ranged from 0.78 to 0.89, with a mean reliability of r = 0.84. Those results indicate that the scoring of content criteria can be quite reliable. Why performance in this study did not reach the level of performance attained in the Horowitz et al. study is not clear. However, it is clear that the evaluators in the Horowitz et al. study spent a great deal of time calibrating their scoring methods before that study began. This calibration process involved the group of evaluators working together on practice statements until a criterion level of agreement on the scorings was reached. Although the two content criteria evaluators in this study had the same initial training, they did not go through a calibration process before this study. The lack of reliability between these two evaluators suggests that such calibration procedures may be necessary to achieve acceptable reliability, since the common training is apparently insufficient. Future research using content criteria should include procedures for assuring calibration between evaluators.

Given the low reliability of the two evaluators, a question arose as to what data should be subjected to additional analyses. Since one of the content criteria evaluators (Horowitz) was a member of the calibrated team of evaluators in the Horowitz et al. (1993) study, it seemed likely that his evaluations would be more representative of evaluations obtained from a calibrated system. In addition, of the two evaluators, Horowitz was by far the more experienced in scoring statements. Moreover, Horowitz had no contact with the subjects, other than the transcripts. The other content evaluator was the original interviewer and his scoring may have been contaminated by his contact with the subjects during the interviews. Therefore, we decided to subject Horowitz's data to the additional analyses.

Univariate Analyses

Content Criteria Scores

Mean content criteria scores of the second evaluator (Horowitz) are shown in Figure 1. The means, univariate *t*-test values, and univariate

Figure 1. Mean Content Criteria Scores

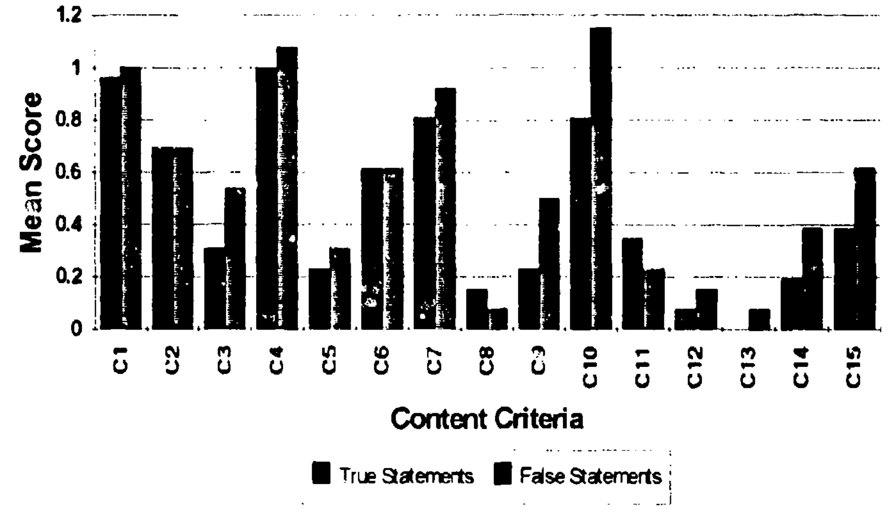


Table 1. Results of the Univariate Analyses of the Content Criteria

Criterion	Mean True	Mean False	*t* (25) *p* value 2-tailed	*r*
C1	0.96	1.00	1.00, *p* = 0.32	-0.14
C2	0.69	0.69	0.00, *p* = 1.00	0.00
C3	0.31	0.54	2.29, *p* = 0.03	-0.24
C4	1.00	1.08	0.70, *p* = 0.49	-0.10
C5	0.23	0.31	0.57, *p* = 0.57	-0.07
C6	0.62	0.62	0.00, *p* = 1.00	0.00
C7	0.81	0 92	1.14, *p* = 0.26	-0.15
C8	0.15	0.08	0.81, *p* = 0.42	0.12
C9	0.23	0.50	2.27, *p* = 0.03	-0.24
C10	0.81	1.15	2.37, *p* = 0.03	-0.26
C11	0.35	0.23	0.90, *p* = 0.38	0.13
C12	0.76	0.15	0.81, *p* = 0.42	-0.12
C13	0.00	0.08	1.44, *p* = 0.16	-0.20
C14	0.19	0.38	2.00, *p* = 0.06	-0.21
C15	0.38	0.62	1.66, *p* = 0.11	-0.21
Total	6.80	8.35	2.74, *p* = 0.01	-0.29

correlations with the true or false criterion are shown for the same content criteria in Table 1. The univariate analyses revealed significant differences at traditional 0.05 α level with three of the criteria, C3 (Quantity of Details), C9 (Superfluous Details), and C10 (Accounts of Subjective Mental State). Four additional variables, C7 (Unexpected Complications), C13 (Raising Doubts), C14 (Self-Deprecation), and C15 (Attributions About Mental States of Others), produced correlations with the criterion of an absolute value of 0.15 or greater and were retained for additional analyses. The total of all content criteria scores was also significantly different for True and False statements, *t*(25) = 2.74, *p* = 0.01, *r* = -0.29.

Inspection of Table 1 reveals an interesting finding; all of the criteria that were significant, or approached significance, produced negative correlations with the true/false criterion. The expectation based on the Undeutsch Hypothesis, and on research with children, was that these predictors should produce positive correlations with the true/false criterion. It may be that adults, having more experience in deception, have adopted strategies of hiding their deception by overemphasizing some of the characteristics that are associated with truth telling. In other words, adults have learned how to tell qualitatively good lies. Nevertheless, if these criteria provide discriminating information, they can be useful regardless of their sign. The multivariate analyses reported in a later section address the discriminative power of the collection of criteria.

Structural Criteria Scores

Mean structural criteria scores are show in Figure 2. The means, univariate *t*-test values, and univariate correlations with the true/false criterion are shown for the same structural criteria in Table 2. At the traditional 0.05 α level, the univariate analyses revealed significant differences between true and false statements with only one of the structural criteria, S2 (Use of Connectors). However, two additional structural criteria, S7 (Use of Generic Terms) and S9 (Changing Terms), did produce correlations with the criterion of an absolute value of 0.15 or greater, and they were also retained for additional analyses. The total of all structural criteria did not differ significantly across true and false statements, $t(25) = 0.18$, $p = 0.86$, $r = 0.03$, ns.

As with the Content Criteria, an inspection of Table 2 reveals an interesting outcome. The one Structural Criterion that reached statistical significance (S2, Use of Connectors) produced a negative correlation with the criterion, despite an expectation of a positive relationship. We have no ready

Figure 2. Mean Structural Criteria Scores

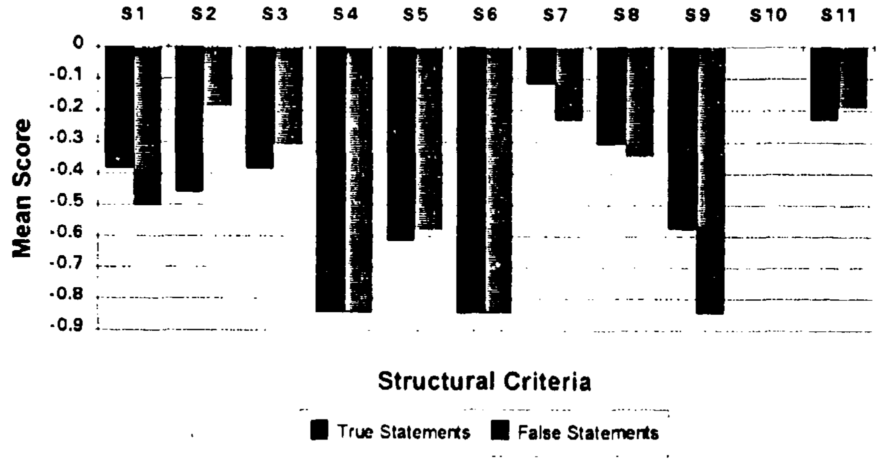


Table 2. Results of the Univariate Analyses of the Structural Criteria

Criterion	Mean True	Mean False	*t* (25) *p* value 2-tailed	*r*
S1	-0.38	-0.50	0.72, *p* = 0.48	0.10
S2	-0.46	-0.19	2.06, *p* = 0.05	-0.26
S3	-0.38	-0.31	0.46, *p* = 0.65	-0.07
S4	-0.85	-0.85	0.00, *p* = 1.00	0.00
S5	-0.62	-0.58	0.17, *p* = 0.87	-0.02
S6	-0.85	-0.85	0.00, *p* = 1.00	0.00
S7	-0.12	-0.23	1.14, *p* = 0.26	0.15
S8	-0.31	-0.35	0.30, *p* = 0.77	0.04
S9	-0.58	-0.85	1.16, *p* = 0.26	0.17
S10	0.00	0.00	-	-
S11	-0.23	-0.19	0.27, *p* = 0.79	-0.04
Total	-4.77	-4.88	0.18, p = 0.86	-0.03

explanation for this finding. It may be that when attempting deception subjects deliberately try to limit the use of connectors because they may hold the notion that when persons are being deceptive they use many connectors. Alternately, it may be that the rehearsed false story was easy to recall and did not generate a need for the use of connectors during production. The issue of what naive persons believe about the characteristics of false statements is an interesting question on its own. Research addressing this question might shed some light on the results of this study and is deserving of support.

Reliability Analysis

On the basis of the univariate analyses, 10 variables were retained for further consideration. Those 10 variables, C3 (Quantity of Details), C7 (Unexpected Complications), C9 (Superfluous Details), C10 (Accounts of Subjective Mental State), C13 (Raising Doubts), C14 (Self-Deprecation), C15 (Attributions About Mental States of Others), S2 (Use of Connectors), S7 (Use of Generic Terms) and S9 (Changing Terms) were next subjected to a reliability analysis. The results of the reliability analysis are shown in Table 3. Chronbach's Alpha for the scale was a modest 0.35. However, considering the limited number of items and their restricted range, low values for Alpha are not unexpected. What is more important, an examination of the column titled, *Alpha if Item Deleted*, suggests that all of the variables were reasonably consistent. There were no one or two variables whose elimination would have resulted in a major improvement in the overall Alpha. Therefore, all 10 of the variables were retained for the multivariate analyses.

Table 3. Reliability Analysis of the Retained Variables

Criterion	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha If Item Deleted
C3	3.59	3.34	0.26	0.35	0.28
C7	3.15	3.98	-0.05	0.19	0.39
C9	3.65	3.79	-0.03	0.19	0.40
C10	3.04	2.74	0.39	0.37	0.18
C13	3.98	3.82	0.26	0.24	0.32
C14	3.73	3.41	0.26	0.40	0.28
C15	3.52	3.51	0.13	0.15	0.33
S2	4.35	3.76	0.02	0.15	0.38
S7	3.84	3.66	0.17	0.11	0.32
S9	3.31	3.12	0.08	0.14	0.38

Multivariate Analyses

Discriminant Analysis

Two approaches were taken to the discriminant analysis. Initially, forced variable entry was used to develop a model that contained all 10 of the retained variables. That analysis failed to produce a significant discriminant function at $\alpha = 0.05$, although it was very close, canonical correlation = 0.57, Wilks' Lambda = 0.67, χ^2 (10) = 17.96, $p = 0.0556$. Overall this non-significant function correctly classified 71.15% (37 of 52) of the statements. The function correctly classified 79.6% (20 of 26) of the truthful and 65.4% (17 of 26) of the false statements.

Next the discriminant analysis was run in a forward stepwise fashion with the analysis set to maximize Rao's statistic. In effect, this causes the discriminant analysis to develop a function that maximizes the distance between the group centroids in discriminant space. The stepwise analysis did produce a significant discriminant function containing five variables, canonical correlation = 0.54, Wilks' Lambda = 0.71, χ^2 (5) = 16.16, $p = 0.0064$. Overall this discriminant function correctly classified 69.23% (36 of 52) of the statements. This function correctly classified 69.20% (18 of 26) of the truthful and 69.20% (18 of 26) of the false statements. The variables retained in this discriminant function and their standardized discriminant function coefficients are shown in Table 4.

Table 4. Variables and Discriminant Function Coefficients for the Significant Discriminant Function.

Variable	Coefficient
C9	0.68
C10	0.50
C13	0.43
C15	0.33
S2	0.63

Logistic Regression Analyses

Two approaches were also taken to the Logistic Regression analyses. Initially, forced variable entry was used to develop a model that contained all 10 of the retained predictor variables. That analysis produced a significant solution after 7 iterations, Model χ^2 (10) = 22.94, $p = 0.011$, that correctly classified 76.92% (40 of 52) of the subjects. With the truthful statements 80.77% (21 of 26) were correctly classified, while 73.08% (19 of 26) of the false statements were correctly classified. The variables and their Logistic Regression *B* values are shown in Table 5.

Table 5. Variables and Logistic Regression *B* Coefficients for the Significant Logistic Regression Function.

Variable	*B* Coefficient
C3	0.73
C7	1.64
C9	1.88
C10	0.94
C13	7.35
C14	1.04
C15	0.42
S2	2.43
S7	-0.55
S9	-0.14

A second Logistic Regression analysis was conducted on the same 10 variables using a backward stepwise procedure that was set up to maximize the likelihood ratio criterion. Accuracy with that analysis peaked with a solution containing 9 variables, Model χ^2 (9) = 22.87, p = 0.006. That analysis correctly classified 78.85% (41 of 52) of the subjects. The 9 variable solution correctly classified 84.6% (22 of 26) of the truthful statements and 73.08% (19 of 26) of the false statements. The variables in the 9 variable solution and their Logistic Regression *B* values are shown in Table 6.

Table 6. Variables and Logistic Regression *B* Coefficients for the Significant 9 Variable Logistic Regression Function.

Variable	*B* Coefficient
C3	0.70
C7	1.65
C9	1.92
C10	0.98
C13	7.52
C14	1.03
C15	0.41
S2	2.54
S9	-0.57

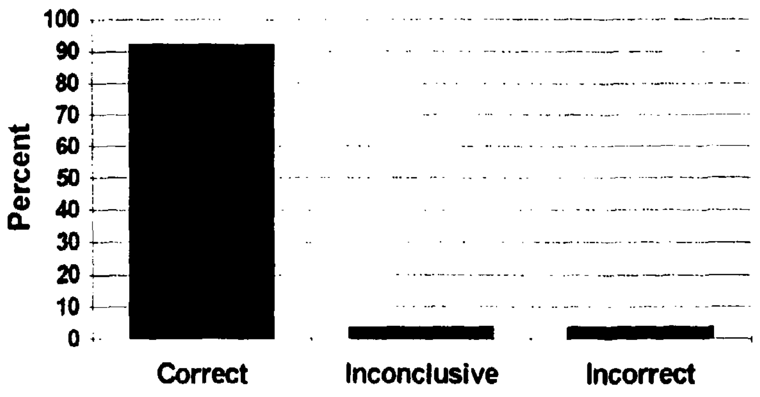
The performance of the logistic regression analysis was promising. The accuracy levels obtained with the 9 variable Logistic Regression solution are comparable to the levels of correct decisions obtained with polygraph examinations of laboratory mock crime subjects. For example, Honts and Devitt (1992) were able to correctly classify 82.05% of their sample of 200 subjects by using stepwise Discriminant Analysis. In the present study, we were able to correctly classify 78.85% of our subjects with a stepwise Logistic Regression.

Within-Subjects Classifications

Although the results of the Logistic Regression were very promising, it is important to remember that the present analyses fail to take full advantage of the information in the data. The analyses described above assume that the truthful and false statements come from different subjects. We felt that classification could be improved if the analysis took advantage of the fact that the data were obtained within-subjects. To estimate discrimination within-subjects, we had the Logistic Regression produce a *p|false* for each subject's statement based on the 9 variable solution. The value of *p|false* for each subject's false statement was then subtracted from the *p|false* value for each subject's true statement. The difference scores were then evaluated. A negative difference score was considered a correct decision since this indicated that the *p|false* for false statements was larger than the *p|false* for true statements.

The results of the within-subjects classifications are illustrated in Figure 3. When the analysis was conducted this way, 24 subjects (92%) were correctly classified, 1 subject (4%) was incorrectly classified and 1 subject (4%) was inconclusive. The mean difference in *p|false* values was -0.36, $s = 0.27$. **This performance is particularly impressive when one considers that the experienced interviewer was no better than chance (13 of 26 correct) in his within-subjects analysis of the subjects at the end of the interview.** The impact of this finding is amplified by the fact that the original interviewer had the advantages of actual contact with the subject and had a chance to observe the subject during the prod iction of the statements. The outcomes of the CAVS analysis illustrated in Figure 3 were based solely on an analysis of the transcribed statements.

Figure 3. Within-Subjects Outcomes



CONCLUSIONS

The present study examined the ability of a Content Analysis of Verbatim Statements to discriminate true and false statements given by laboratory subjects regarding their work experience. On the basis of the data analyses we believe that the following tentative conclusions are warranted:

- Our experienced interviewer was no better than chance at discriminating true from false statements.
- The results of the between-subjects analysis with Logistic Regression suggest that CAVS can provide good discrimination between true and false statements. An accuracy rate of 78.85% was obtained with a 9 variable Logistic Regression model. Such performance is comparable to polygraph examinations conducted on subjects under similar conditions. Moreover, it exceeds DoD's own estimates of the validity of polygraph tests used in employment screening.
- When the output of the Logistic Regression was recast as a within-subjects analysis, 24 of the 26 subject's statements were correctly classified. The within-subjects analysis misclassified only one subject's statements. This seems to be very promising performance indeed, and suggests that CAVS is worth additional research.
- Our results suggest that reliability of CAVS scoring may be an issue of concern. The two CBCA evaluators in this study failed to achieve an acceptable level of reliability. Despite the finding in other research that acceptable reliabilities with CBCA were possible, we believe that the reliability question deserves considerable scrutiny.
- Both of the evaluators who applied the structural criteria reported that their job was more difficult because the structural criteria had originally been developed for use on written statements given by the subject. The relatively poor performance for the structural criteria in this study may have been due to their application to spoken rather than written statements. Additional research is needed to address this question.

RECOMMENDATIONS

On the basis of our experience in this study, we make the following recommendations:

- Additional research should be conducted on the reliability of CAVS scoring.
- Toward the goal of examining the reliability problem, we would recommend that several individuals from the same laboratory receive professional training in CBCA and in the use of the Structural Criteria. It would then be possible to have control of the rater calibration process and the progress of research could be expedited. The conclusion of the present study was delayed because of dependence on outside evaluators. In the end one evaluator failed to provide data and was thus not included in the study.
- Additional research should be conducted as soon as possible to attempt replication of the initial findings of this study and to increase the data base for multivariate model building. Considerably more confidence could be placed on multivariate models based on larger samples of subjects.
- In order to increase the data base, we would recommend collecting between-subjects as well as within-subjects data. Between-subjects validation will have to be done before the technique can be applied in the field. Our recommendation is for a follow-up study including at least 100 truthful and 100 false statements. At least half of the statements should be obtained between-subjects. More subjects than this would be desirable if resources are available.
- CAVS should be applied to archival field data at the first opportunity. We have a database of statements obtained from confirmed field polygraph examinations. While these statements were not obtained in employment situations they might be useful for a first attempt at field validation.
- The structural criteria should be examined with written statements. This could be done as part of the study described above.

FINAL COMMENTS

In conclusion, we believe that the present study suggests that CAVS has great potential for credibility assessment in the national security clearance process. The results of the within-subjects analysis were particularly strong. Such within-subjects discriminations could be very valuable in the field. For example, it is often possible to independently verify some items of information in a subject's employment background but not others. The present results suggest that we could compare a subject's statements about a known truth to statements of unknown veracity and make valid credibility assessments about the unknowns with a high degree of accuracy. CAVS is particularly attractive for this application because it requires no instrumentation and no highly trained examiner to obtain the data. Moreover, CAVS can be applied covertly and in a retrospective manner, as long as the verbatim nature of the subject's statements are retained. The actual CAVS analysis could then be performed minutes, days or even years after the statement was obtained. Given the potential value of such a technique to the national security system, we strongly urge PERSEREC and ONR to pursue this line of inquiry and to fund additional research.

REFERENCES

Anson, D. A., Golding, S. L., & Gully, K. J. (1993). Child sexual abuse allegations: Reliability of criteria-based content analysis. Law and Human Behavior, 17, 331-342.

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). The validity of detection of deception for multiple issues. Psychophysiology, 26, S13. (Abstract)

Honts, C. R. (1991). The emperor's new clothes: Application of polygraph tests in the American workplace. Forensic Reports, 4, 91-116.

Honts, C. R. (1992). Counterintelligence scope polygraph (CSP) test found to be a poor discriminator. Forensic Reports, 5, 215-218.

Honts, C. R., & Devitt, M. K. (1992). Bootstrap decision making for polygraph examinations: Final report of DOD/PERSEREC Grant No. N00014-92-J-1794. University of North Dakota, Grand Forks.

Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. Journal of Applied Psychology, 70, 177-187.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1987). Physical countermeasures may reduce the physiological detection of deception. Journal of Psychophysiology, 1, 241-247.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1993). Mental and physical countermeasures reduce the accuracy of polygraph tests. Manuscript accepted for publication in the Journal of Applied Psychology.

Horowitz, S. W. (1992). Empirical support for statement validity assessment. Behavioral Assessment, 13, 293-313.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1993). Reliability of criteria-based content analysis of child witness statements. Manuscript submitted for publication.

Kaster, J. (1991). Interviewing witnesses and statement analysis. Ottawa: Canadian Police College.

Landry, K. L., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. Law and Human Behavior, 16, 663-676.

Raskin, D. C., & Esplin, P. W. (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), The suggestibility of children's recollections: Implications for eyewitness testimony (pp. 153-164). Washington, D.C.: American Psychological Association.

Raskin, D. C., & Esplin, P. W. (1992). Statement validity assessment: Interview procedures and content analysis of children's statements of sexual abuse. Behavioral Assessment, 13, 265-291.

Sapir, A. (1988). SCAN: Scientific content analysis. Unpublished manuscript.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille, (Ed.), Credibility assessment (pp. 135-154). Dordrecht, The Netherlands: Kulwer.

Steller, M., & Koehnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed.), Psychological methods in criminal investigation and evidence (pp. 217-245). New York: Springer.

Undeutsch, U. (1989). The development of statement reality analysis. In J. C. Yuille (Ed.), Credibility assessment (pp. 101-120). Dordrecht, The Netherlands: Kulwer.

Winbush, M. (1993). Countermeasures in the concealed knowledge test: Unpublished Honors Thesis submitted to the Psychology Department of the University of North Dakota, Grand Forks.

Acknowledgments

The authors would like to acknowledge and thank Jason Sickler and Tonia Timm for their help in running the subjects of this study. We also want to recognize and thank Steve Horowitz and Larry Driscoll for providing the independent evaluations used in this study. We would also like to thank John Kaster, John Kircher, and David Raskin for their helpful comments during the design, conduct, and analysis of this study.

The research depicted in this report was sponsored by the Dɘpartment of the Navy, Office of the Chief of Naval Research. The contents of this report do not necessarily reflect the position or the policy of the United States government, and no official endorsement should be inferred.