Law and Human Behavior, Vol. 29, No. 6, December 2005 (© 2005)
DOI: 10.1007/s10979-005-7370-z

Behavioral Cues to Deception vs. Topic Incriminating Potential in Criminal Confessions

Martha Davis,[1,4] Keith A. Markus,[1] Stan B. Walters,[2] Neal Vorus,[1] and Brenda Connors[3]

Coding statements of criminal suspects facilitated tests of four hypotheses about differences between behavioral cues to deception and the incriminating potential (IP) of the topic. Information from criminal investigations corroborated the veracity of 337 brief utterances from 28 videotaped confessions. A four-point rating of topic IP measured the degree of potential threat per utterance. Cues discriminating true vs. false comprised word/phrase repeats, speech disfluency spikes, nonverbal overdone, and protracted headshaking. Non-lexical sounds discriminated true vs. false in the reverse direction. Cues that distinguished IP only comprised speech speed, gesticulation amount, nonverbal animation level, soft weak vocal and "I (or we) just" qualifier. Adding "I don't know" to an answer discriminated both IP and true vs. false. The results supported hypothesis about differentiating deception cues from incriminating potential cues in high-stakes interviews, and suggested that extensive research on distinctions between stress-related cues and cues to deception would improve deception detection.

KEY WORDS: deception; nonverbal communication; criminal confessions; stress.

Researchers have identified verbal and nonverbal behaviors that significantly discriminated false from truthful statements (deTurck & Miller, 1985; Ekman, Friesen, & O'Sullivan, 1988; Zuckerman, DePaulo, & Rosenthal, 1981). However, the list of discriminators has varied from study to study, and differences in measurement have made results difficult to compare (Knapp & Comadena, 1979). DePaulo et al. (2003) identified a limited set of deception cues from a long list of possibilities in their meta-analysis of 120 deception studies. They found that deception cues were more prominent when the lies were about transgressions and the speaker was especially motivated to succeed at the deception (DePaulo et al., 2003).

[1] John Jay College of Criminal Justice, New York, New York.
[2] Stan B. Walters Associates, Inc., Versailles, Kentucky.
[3] Naval War College, Newport, Rhode Island.
[4] To whom correspondence should be addressed at Psychology Department, John Jay College of Criminal Justice, 445 W. 59th Street, New York, New York 10019; e-mail: marthadavisr@mac.com.

Because deception cue research has primarily involved student volunteers and experimentally-evoked lies, the results may not generalize to naturalistic contexts in which the stakes are very high (Miller & Stiff, 1993). Studies of deception cues based on careful examination of videotapes and transcripts of actual high stakes encounters are rare (Horvath, Jayne, & Buckley, 1994; Mann, Vrij, & Bull, 2002; Shuy, 1998; Vrij & Mann, 2001). Increasingly, interrogations and confessions are being videotaped in the United States, Great Britain and other countries (Baldwin, 1992), and research on criminal suspect interviews is needed in particular. While methodologically there are advantages to studying lies evoked in the laboratory, Horvath et al. (1994) and Mann et al. (2002) have demonstrated the efficacy of using information from criminal investigations to corroborate which suspect statements were true and which false in preparation for an analysis of behavioral cues to deception.

The present study is the first to combine an extensive analysis of deception cues in criminal suspect statements with an assessment of how potentially incriminating the subject matter or topic was to the speaker. How the suspect first met the victim of a crime poses a less threatening question than where he or she was standing in the room when the victim was shot. Incriminating potential (IP) was treated in this study as irrespective of whether the suspect in fact incriminated him or herself. Although some suspects might be more threatened by the same question than others, we regarded IP as a situational stressor especially important with respect to brief changes in psychological stress within suspect interviews.

Mann et al. (2002, p. 368) appeared to look for differences in topic incriminating potential by choosing truthful utterances that were "as comparable as possible in nature to the lies" and excluding name and address information as too easy. However, some examples they cite appear potentially more incriminating than others (e.g., talk about the victim's alcohol problem vs. denial of when the speaker entered the crime scene). The present study examined cues associated with differences in IP, such as the contrast between crime-relevant background information vs. crime scene questions. If, as seems likely, deception increases as the subject matter becomes potentially more incriminating, then cues found related to deception may be related to IP as well.

Cues and Deception Processes

Several theories have been proposed as to how demeanor cues relate to deception processes (Anolli & Ciceri, 1997; Ekman, 1992; DePaulo et al., 2003). Cues associated with lying have been interpreted as (a) manifestations of tension, anxiety, or heightened arousal, (b) efforts to control output to minimize cues and mistakes, (c) communication disruptions due to cognitive overload, and (d) contradictions between the expression of a truly experienced emotion and the words or expressions that belie or conceal it. Conceivably, some of these possibilities may co-occur. For example, vocal pitch increases associated with heightened fear (cf. Ekman, O'Sullivan, Friesen, & Scherer, 1991) may be displayed with efforts to control one's performance, such as inhibiting hand motions during deception (Vrij, Akehurst, & Morris, 1997).

Explicitly or implicitly, parallels have been drawn between cues to deception and signs of conflict or arousal that can occur during truthful communications.

> ... people sometimes act differently when they are lying and telling the truth. But these differences are not communicants of deception per se, but instead reflect internal states like heightened cognitive processing, fear, guilt, excitement, or arousal, which may be associated with deception under some conditions (Kraut, 1980, p. 213).

The question of whether deception cues differ in discernible ways from manifestations of internal states accompanying stressful truth-telling has theoretical and practical implications. Lists of behaviors found associated with deception (cf DePaulo et al., 2003) can seem indistinguishable from signs of nervousness and heightened arousal during truthful statements. In the present study we tested implications of the model shown in Fig. 1 in which some cues to deception differ from behaviors occurring in a heightened-arousal-but-honest condition. The model predicts three types of cues: those related to both IP and deception processes (mixed cues), those related only to IP, and those related only to deception. The underlying theory views deception cues as distinctive in type, form and how they fit into the verbal exchange. They constitute special variants of common behaviors that marked the particular internal processes involved in deception and would not be displayed in precisely the same way during high stakes, truthful statements.

Identifying these special variants depends in part on the precision of the observation. In the classic experiment by Ekman et al. (1988), some nurses who described a burn victim video as a pleasant scene displayed tense smiles. Close inspection revealed that the smile contained a trace of an expression of disgust that people openly displayed without masking smiles when they viewed the gory tape and did not lie

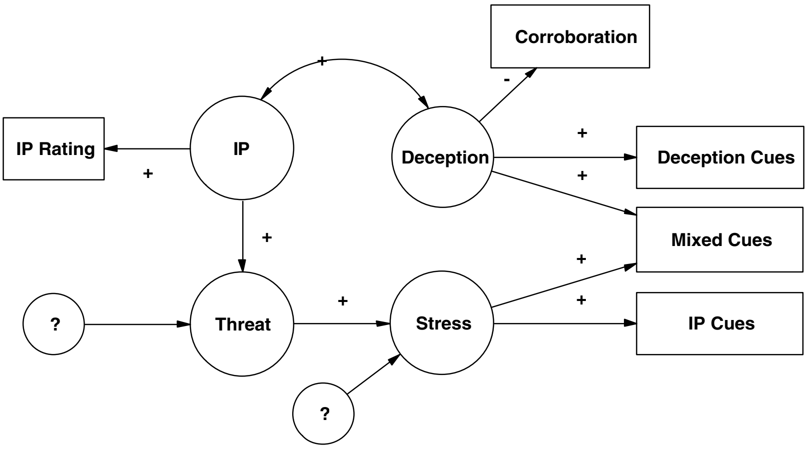

Fig. 1. Theoretical model of cue production.

about what they saw (Ekman, 1988). The tense facial expression – which in a less vigorous analysis might be interpreted as a sign of emotional stress, performance anxiety, displeasure with the task, and so on – appeared on careful examination to contain a facial expression which directly contradicted the verbal content. This is a prime example of how precision in identifying tiny but critical differences can decrease the ambiguity of cues, narrow the interpretive possibilities, and help differentiate deception from heightened arousal cues.

Selecting Cues

The behaviors subjected to coding were chosen because they appeared promising as cues to heightened arousal and/or deception according to the research literature and the experience of two of the authors with analyzing forensic interviews (Davis & Hadiks, 1995; Walters, 1996). We considered verbal content cues, vocal or paralinguistic details, and nonverbal behavior. Some categories such as consistency, logic, and quantity of detail included in methods for assessing the credibility of verbal content (Sapir, 1987; Shuy, 1998; Steller & Koehnken, 1989; Vrij, Edward, Roberts, & Bull, 2000) were not applicable because, as will be discussed, our study had to be limited to brief, isolated utterances. However, we could consider short disclaimers and qualifiers such as “I don’t know” or “I just . . .” added to an answer.

Behaviors found related to anxiety such as speech disturbances (Kasl & Mahl, 1965) have frequently been included in deception cue studies (DePaulo et al., 2003). We examined several of these (filled pause sounds, stammering, silent pauses, phrase interruptions or breaks in thought line). Several cues commonly associated with nervousness or deception, such as fidgeting, postural shifting, and face rubbing (Miller & Stiff, 1993), were dropped because they were either too rare, unreliable, or difficult to observe in these videotapes. Indications of output, such as number of words per utterance and amount of speech-related gesticulating, were included in the present study, as were signs of pressure such as increases in speech speed (DePaulo et al., 2003). Preference was given to recording incidence of specific behaviors, but we also included ratings involving more judgment, e.g., degree of animation. Practical considerations affected what was coded as well. Behaviors such as blink rate (Mann et al., 2002), pupil size (Hess & Polt, 1963), subtle differences in facial expression (Ekman et al., 1988), and foot and leg movements could not be included because of the video quality and medium camera shots from the waist up characteristic of our tape collection.

Although we examined cues commonly found in deception cue research, how we operationalized many of them was distinctive and influenced by the experience of two of the authors with suspect interviews (Davis & Hadiks, 1995; Walters, 1996). In our experience, lies in criminal confessions were often brief utterances within otherwise truthful passages. Also, what could be corroborated with independent information was often limited to brief passages. This meant we had to concentrate on relatively short true and false speech units, and the cues that might distinguish false from true would be relatively infrequent, discrete behaviors or marked changes. Therefore, we predicted that a reliable deception cue in this context was a *special variant* of a common behavior made distinctive in part by the way that it (a) clustered or spiked in frequency, e.g., three phrase and/or consecutive word repeats within a few seconds; (b) was exaggerated, e.g., protracted shaking "no" of the head; or (c) deviated well above or below the speaker's baseline, e.g., one word very low in volume.

Deception and IP Cue Differences

Bavelas, Black, Chovil, and Mullet (1990) have criticized deception research that treats veracity as a simple true vs. false dichotomy. They contend that many answers given under pressure are equivocal and fall somewhere along a truthfulness continuum. In the present study, we selected assertions and denials that were unequivocal and unambiguous, so our true and false selections would fall at the extremes of the continuum Bavelas et al. examined, and treating true vs. false as a simple dichotomy was appropriate. We considered brief lies that occurred in the midst of otherwise truthful responses in a high stakes interview as discrete events that involved a special complex of cognitive and affective processes qualitatively different from the truthful portions of the statements. We therefore expected that cues to deception in the context under study would involve discrete, relatively infrequent spikes, clusters, or exaggerations of common behaviors, especially those that possessed specific lexical meaning.

We considered IP as one prominent source of threat to the speaker, and the IP rating (IPR) of a question or topic as an operational measure of variations in IP. IP is therefore one of several contributors to psychological stress within the interview (Fig. 1). The circles with question marks in Fig. 1 indicate that IP is only one source of threat to the speaker and threat is only one source of psychological stress in this model. Because psychological stress exhibits relatively continuous variation in degree, we expected cues reflecting continuous increase or decrease along a single dimension of behavior to be related to IP, not to true vs. false. Thus, for example, we expected speech rate to be related solely to IP because it was varied gradually along a single dimension as would characterize changes in arousal level related to psychological stress, but we predicted that an exaggerated, overlong headshaking "no" that was an infrequent event with a lexical meaning would be a cue to deception.

Goals and Hypotheses

The first goal of this study was to identify vocal and nonverbal cues that reliably discriminated false from true answers in a set of real, high stakes interviews. The second goal was to identify behavioral cues to IP level and compare these with the deception cues. In the process, we tested the following hypotheses:

1. The proportion of false answers will increase as IP increases.
2. Certain cues will discriminate true vs. false, but will not discriminate IP.
3. A different set of cues will discriminate IP, but not true vs. false.
4. A third set will discriminate both IP and true vs. false.

METHOD

Videotaped Subjects

The 28 subjects were criminal suspects videotaped giving statements to assistant district attorneys (ADAs) after they had been interrogated by police officers or detectives. While sometimes referred to as a confession, the ADA interview is formally called a statement. Suspects consented to the videotaping and each interview began with a repeat of the Miranda warning. Given the age of the tapes and the geographic dispersal of subjects, it was not possible to obtain consent for research use. Subject confidentiality was protected by strict controls on who could see the tapes, disguise of identities in reports, and destruction of the tapes when the study was completed.

The 28 suspects (4 women and 24 men) were ethnically diverse (10 Caucasian, 12 African American and 6 Hispanic). They ranged from 18 to 58 years of age with 14 falling between 18 and 25, 10 between 26 and 34, and 4 subjects over 35. Socioeconomically, 13 could be characterized as in a low SES group, 9 lower middle, 5 middle and 1 upper middle. From information that was available, just over 50% were first offenders. The list of convictions was as follows: 2 premeditated murder, 17 unpremeditated murder, 2 negligent homicide, 4 assault with a deadly weapon, 1 sexual assault, and 2 illegal weapons possession.

What suspects were convicted of, however, did not perfectly correspond to what they admitted to on tape. Ten confessed guilt but claimed mitigating circumstances such as being under the influence of drugs. Eight admitted accompanying a perpetrator, but maintained that they were innocent of the crime. Seven argued self-defense or efforts to stop the victim who was described as the aggressor. In other words, almost all minimized their role to some extent and admitted to less than the crime of which they were convicted.

Transcript Preparation and Utterance Selection

To isolate the segments that would be coded in detail, the interviews were transcribed verbatim and examined for utterances that could be corroborated as true or false from information from the investigations.

Transcription

Research assistants not involved in behavioral coding typed transcripts of the interviews from the audio alone. Tapes were reviewed many times to insure accurate transcription including partial words, filled pause sounds (umms, uhhs) and where silent pauses occurred. Two assistants completed each transcript, with one checking the typescript of the other and both reviewing any corrections for final agreement.

Corroboration

The next step was to identify brief utterances within each transcript that could be confirmed as true or false. This was done in two ways. For the first batch of 21 interviews, a detective active on the case was available to meet with a psychologist and go over the transcript line by line in order to find answers that could be confirmed by information from the investigation. The psychologist recorded which utterances were corroborated true or false along with the source of the corroborating information. An additional seven transcripts (the second batch) were corroborated with information from criminal records and the source of the information was noted. In this case a highly experienced detective examined the record to confirm utterances that were true or false in the transcript.

In both batches every effort was made to confirm the veracity of an utterance with strong evidence. 20.2% of the corroborations were based on laboratory evidence, 39.8% on crime scene analysis, and 40.1% on witness or suspect accounts. Witness accounts were based on close-up and fairly protracted exposure to the criminal activity, not obscured or fleeting observations, and most witnesses were not implicated in the crime. Among the witness or suspect accounts, 11.3% was based on two or more witnesses, at least one of whom was impartial; 24.0% were based on one impartial or two or more partial witnesses. Eight utterances or 2.4% were corroborated by speaker recant or correction, and another 2.4% were based on one implicated witness but the information did not appear to help his or her case.

For the utterance to be designated true, every detail of it had to be independently corroborated as true. If it was designated false, at least one part of it had to be confirmed false. The number of corroborated utterances ranged from 6 to 20 and averaged 12 per subject. The number of false utterances per subject varied from 1 to 11. We were able to confirm only one or two false utterances for 12 of the 28 subjects. Four subjects had a few more false than true utterances. An utterance had to contain unequivocally expressed information about events or actions that could be confirmed. However, a few dealt with information about intent or what the subject knew because there was a great deal of compelling independent evidence to confirm this. The total number of utterances selected from 28 interviews in this way was 337 (229 true and 108 false).

Unit Refinement

Some answers contained phrases that could not be corroborated or changes in topic such that the speech unit subjected to coding had to be refined to exclude such content. The exact delineation of each utterance was time-consuming and had to be done after the corroboration procedure. The precise start of an utterance was defined either by the end of a question or the point within a longer answer at which the information could be corroborated. The end of the utterance was defined as (a) the start of another question, (b) a change of topic mid-answer or (c) up to three pieces of information about the topic. After marking the unit on the transcript, a research assistant listened to the audio and stopped the computer event recorder at the precise onset and end of each unit, yielding computer recording of durations accurate to .03 s. Given the unit criteria, utterances subjected to coding were quite brief (from 0.15 to 41.6 s; $M = 5.28$ s, $SD = 4.67$). Number of words per utterance varied from 1 to 66 ($M = 12.55$, $SD = 11.37$). Utterances were short not only because it was difficult to corroborate every detail of a longer segment, but because it would be hard to identify which contents of a long answer were associated with which behavioral details.

Rating Incriminating Potential

Four forensic psychology graduate students were taught IP rating of utterances. IPR had already been completed by a psychologist and detective on 21 interviews (authors reference Davis, Walters, Vorus, Meiland, & Markus, 2000), but we had this redone by students because they were naïve to the design, behavioral coding and hypotheses of the study. The batch of seven interviews added later was coded by one of these students and two additional students. All student coders were instructed in the criteria for each level of the four-point rating (Table 1) and practiced coding until they displayed adequate agreement. In order to have sufficient context for making their judgments, coders read the entire interview transcript and

Table 1. Incriminating Potential Rating: Coding Criteria, Frequencies and Percentage False

IPR[a]	Definition	T[b]	F	*N*
1 Negligible	Identifying information such as nicknames or car ownership. No weapon information	37 (100.0%)	0 (00.0%)	37 (11.0%)
2 Medium low	Background information on weapon, characteristics of car used in crime, previous crime history, how suspect knew victim or accomplice in past. Non-criminal activity of self and/or accomplices preceding crime activity. EXS: "I used to deal crack cocaine." "I went to school with [accomplice]. "I had a lot to drink and fell asleep." "I got it last month at a gun show."	55 (76.4%)	17 (23.6%)	72 (21.4%)
3 Medium high	Description of victim or bystander appearance, location or activity related to crime events; description of objects, locations, layouts, time factors of crime event; information on whether speaker was truthful or not. EXS: Q: "What did [victim] do then?" A: "She ran down the street." Q: "You told the officer he was there?" A: No, I told him he wasn't there."	34 (54.0%)	29 (46.0%)	63 (18.7%)
4 Highest	Activity or reactions of speaker and/or accomplice directly related to crime in question; e.g., what immediately leads up to or follows crime; who participated, their locations, who had weapon; crime planning; what done with evidence and weapon. EXS: "I grabbed the gun from him." "I threw the gun in the trash." "Then she told them where the safe was."	103 (62.4%)	62 (37.6%)	165 (49.0%)

[a] If more than one type of information in an utterance, the one judged primary was rated.

[b] T and F refer to corroborated true and corroborated false. T and F percents are percent of utterances within that IPR level.

were given information as to who was convicted of what and who was a victim or bystander. However, the student coders did not know which utterances were confirmed true and which false.

Reliability for the first pair of student coders was $r = .78$, $N = 341$. After separately coding the 267 utterances of the first batch of interviews, the coders reviewed disagreements and settled them by consensus. Therefore, the IPR data from the first 21 interviews subjected to analyses was based on 2 of 2 agreement or review of disagreements and consensus. The 83 utterances of the second batch were coded by three students. Consistency in coding between the first and second teams was secured because one coder was on both teams, and because during training the two new coders practiced on utterances from the first batch until levels of agreement between them and the first pair were adequate. Reliability for each pair on the second team was as follows: $r = .80$, $r = .77$, and $r = .73$, $N = 83$. IPR second batch data used in subsequent analyses was based on 2 or 3 out of 3 agreement plus coder review and consensus for the utterances initially lacking agreement.

Behavioral Coding

Lists of potential verbal, paralinguistic and nonverbal cues to deception or anxiety were devised from study of the literature, and the authors' experience with videotaped forensic interviews (Davis & Hadiks, 1995; Davis et al., 2000; Walters, 1996), and consideration of the limits of the videotapes (video and audio quality, camera angle, etc.). Table 2 lists the behaviors coded in the present study.

Following the categorical vs. continuous change distinction discussed earlier, the cues in Table 2 predicted to discriminate true vs. false were discrete moments of protracted headshaking, speech disfluency spikes, soft/weak vocal quality, nonverbal overdone, long silent pauses, word/phrase repeats, the qualifier "I [or we] just," and addition of "I don't know" to an answer. Also, a brief burst of non-lexical sounds (umms, uhhs, sighs, gutturals) was predicted to be a true vs. false cue because it was a relatively infrequent, discrete event. Continuous measures predicted to discriminate IP level were word number, nonverbal animation, gesture amount, and speech speed.

Coder Training

The training of the behavioral coders involved extensive instruction by the authors who developed the coding (Davis & Hadiks, 1995; Walters, 1996), practice with non-research items, and periodic checks of agreement until they were ready to code the research items independently. The first 21 interviews were coded by different teams for each of three modalities. Experts in movement analysis from the faculty of the Laban/Bartenieff Institute of Movement Studies (NYC) were trained in the selected nonverbal codes and limited to video presentation without audio or transcript. One team of forensic psychology graduate students was trained in vocal coding (from audio with transcript but without video) and another in verbal content categories (from transcript alone).

Only one coder from these teams was available two years later to work on an additional seven interviews. All behavioral coding of these interviews was

Table 2. Behavior Categories and Coder Reliability

Category	Definition per utterance	Reliability[a,b]
Verbal phrases and output		
Word/phrase repeats	Exact phrase and/or consecutive word repeats (threshold[c]: 3 or more)	word $\kappa(241) = .87$, $SE = .06$ phrase κ $(240) = .49$, $SE = .07$
"I/we just"	"Just" qualifier, e.g., "I just shot him once"	$\kappa(241) = .58$, $SE = .12$
"I don't know"	If additional to answer	$\kappa(241) = .93$, $SE = .05$
Number of words	Count of words per utterance	r $(240) = .99$
Vocal coding		
Disfluency spike	Interrupt of phrase or line of thought or speech stammering or mumbling (threshold: 2 or more)	$\kappa(353) = .41$, $SE = .06$ alpha (83) = .71
Soft/weak volume	Almost inaudible, indefinite intonation	$\kappa(76) = .64$, $SE = .12$
Non-lexical sounds	Filled pause sounds (FPS: ums and uhs) and/or sighs, gutteral sounds (threshold: 2 FPS and/or 1 or more sounds)	$\kappa(173) = .86$, $SE = .05$ alpha(83) = .88
Speech speed	Number of subject words divided by duration measured to 3/100ths of second	Duration taken from computer event recording of onset/end
Long silent pause	Before or within answer pause a second >2s	$\kappa(173) = .80$, $SE = .08$
Nonverbal coding		
Long head shaking	Continuous back/forth "no" motion 5 or more	$\kappa(223) = .82$, $SE = .07$
Gesture amount	Proportion of utterance (rated 1–4) with hand gestures that accompany speech	r $(307) = .87$
Nonverbal overdone	Facial expression, gesticulation or action exaggerated, excessive or "put on" plus 2+ shoulder shrugs per utterance	alpha(160) = .65 shrug r $(109) = .84$
Nonverbal animation	4-point rating of motor activity/expressiveness	r $(358) = .74$

[a] Cohen's κ based on two coders; alpha reliability coefficients based on three coders.
[b] Number in parentheses is the n used for the reliability check.
[c] The threshold criterion is number per utterance predicted to discriminate false.

completed by one team of forensic psychology graduate students. First, they learned and completed the coding of nonverbal behavior without audio or transcript. After that, the team learned to code verbal content cues from transcript alone and completed this before instruction in vocal coding (with transcript but without video). One coder from the old teams was able to join them for the last stage, vocal coding.

This raised questions as to whether coding in one procedure was different from coding in the other. For several reasons, we were assured that it was not. First, as part of their training, the new coders had to achieve adequate levels of agreement with the old coders on a selection of items from the first batch of 21 interviews before they could code the second batch of utterances. Secondly, we did not find differences in the distribution of behaviors. Given only seven interviews in the second batch and the relatively low incidence of categorical variables in general, it seemed unreasonable to compare first and second batch coding of each cue separately, so we clustered three of the cues for one comparison and two of the cues for another. There was no difference between first and second batch coding as regards the percentage of utterances containing word repetition, I don't know phrases, and protracted head shaking cues (old coding = 12%; new coding = 13%), and first vs. second batch coding was not related to presence/absence of these cues ($\chi^2(1, N = 337) = .04$, $p = .84$). There was also little difference between first and second batch coding as regards the percentage of utterances containing nonverbal overdone and speech disfluency spikes (old coding = 13%; new coding = 8%: $\chi^2(1, N = 337) = 1.49$, $p = .22$).

In both the one-team-per-modality coding of the first batch and the one-team-each-modality-in-sequence coding of the second, observation of the nonverbal behaviors was separate from coding of paralinguistic and verbal content cues. Every coder coded alone before review of disagreements and was free to replay or examine utterances as long as needed. None of the coders knew the hypotheses of the study or whether the utterances were true or false.

Coder Reliability

Table 2 presents coder reliability data with the definitions of the coded variables. Most of the behavioral coding involved present/absent, categorical determinations rather than ordinal data, and for this Cohen's kappa (κ) was preferable to percent agreement because it corrects for chance. According to Landis and Koch (1977), our four lowest κ coefficients fell within a fair to good range of .40–.75 and the other five were excellent (i.e., above .75). Monitoring agreement levels during training was only the first stage in the effort to insure observation accuracy. For those variables originally coded by at least two coders, disagreements were either reviewed by the team that coded the behavior and settled by consensus, or a third and, if necessary, fourth coder reviewed the observation until there was either 2 of 3 or 3 of 4 agreement. For the first batch nonverbal coding originally done by only one observer, at least two trained coders repeated the observations and 2 of 3 or 3 of 4 agreement was required for the observation to be included in the final dataset.

RESULTS

This section reports analyses testing each hypothesis starting with the association between IP and true vs. false (T/F). Secondly, assessment of the association between the behavioral cues and T/F is discussed, behavioral cues and IP is third, and cues related to both T/F and IP is fourth. The fourth section is a report of individual differences in cue patterns, and the fifth section addresses the predictive accuracy of the cues.

True vs. False and Incriminating Potential

Hypothesis 1 stated that the proportion of false responses increase as IP increases. Correlation between T/F and IPR partialing out subject dummy codes was $r_{\mathrm{ti.s}} = .28$, $df = 308$, $p < .001$. Inclusion of the subject dummy codes addressed the clustered data structure with a fixed-effects model (Cohen, Cohen, West, & Aiken, 2003), limiting statistical generalization to utterances from the participants in the sample. The limited number of subjects and the nonrandom sample of subjects precluded reasonable use of a random-effects model. As Table 1 shows, IPR 1 (identifying information) contained no false utterances, and the percentage of false utterances increased markedly between IPR 2 (background information potentially incriminating but tangential to the crime) and IPR 3 (information related to the crime itself but not to crime actions of the suspects). However, against prediction, the percentage of false utterances decreases somewhat between IPR 3 and 4 (potential criminal actions of suspects), so that while the number of false responses increases linearly from lowest to highest, the proportion of false responses decreases between 3 and 4. This suggests that measures of linear association may not provide the best test of the hypothesis. A hierarchical binary logistic regression predicting T/F from subject (categorical) and IP (numeric) improved statistically significantly in fit ($\chi^2(1, N = 337) = 12.1, p < .001$) with the addition of a polynomial term for IP squared ($b_{\mathrm{IP}} = 3.78$, $SE = 1.08$, $b_{\mathrm{IP}^2} = -0.80$, $SE = -0.27$). The linear partial correlation remains strong over the 172 cases with the lowest three IPR ($r_{\mathrm{ti.s}} = .42$, $df = 143$, $p < .001$) whereas the correlation disappears for the 228 cases with the highest two IPR ($r_{\mathrm{ti.s}} = .04$, $df = 199$, $p = .58$). These results suggest a ceiling effect but provide partial confirmation for Hypothesis 1.

True vs. False Cues

We examined the bivariate correlations between all the cues, T/F and IPR. A stem and leaf plot of the correlations revealed a roughly normal distribution ($M = .12$, $SD = .13$) with approximately half of the cases falling between .04 and .19 (median = .10). The extreme low correlations remained weak and all involved speech speed with the negative correlation making theoretical sense: −.23 (long pauses), −.21 (ums, uhs, non-lexical sounds), and −.16 (vocal weak). The extreme high correlations, on the other hand, appeared to reflect two genuine outliers each of which makes theoretical sense (.54 between animation level and gesture percent, .53 between speech speed and word number). The remaining correlations all fell below .40. None of the correlations raised concern about colinearity in the multivariate analysis. The same analysis including the subject dummy variables produced similar conclusions. The resulting leptokurtic distribution ($M = .00$, $SD = .09$) had the same extreme values with a narrower inter-quartile range ($Q_1 = -.04$, median $= -.03$, $Q_3 = .02$).

A hierarchical binary logistic regression analysis predicting T/F provided a test of Hypothesis 2. The first model included the four-point IPR, subject (categorical), and three types of corroboration source (categorical: lab evidence, crime scene analysis, and verbal account). The second model added to these covariates

Table 3. Binary Logistic Regression: True vs. False Cues

Block	Variables	χ^2	*df*	*p*
1	Subject (Cat.), incriminating potential level (Cat.),	Block 109.70	32	.000
	corroboration source (Cat.)	Model 109.70	32	.000
2	Predicted deception cues: Word/phrase repeats, speech	Block 54.35	9	.000
	Disfluency spikes, protracted head shake, nonverbal overdone, "I don't know," "I [we] just" qualifier, long pauses, soft/weak vocal, non-lexical sounds	Model 164.06	41	.000
3	Predicted incriminating potential cues: word number, nonverbal animation,	Block 1.21	4	.876
	gesture amount, speech speed	Model 165.27	45	.000

the cues hypothesized to predict T/F (word/phrase repeats, speech disfluency spikes, protracted head shake, nonverbal overdone, long pauses, soft/weak vocal, "I don't know" phrases, "I [or we] just" qualifier, and non-lexical sounds). The third model added the cues hypothesized to predict IPR, but not predict T/F: word number, nonverbal animation, gesture amount, and speech sound. Table 3 presents the fit statistics for the three models. The fact that the second model improves prediction over and above the first supports Hypothesis 2.

Table 4 presents the regression weights for the third model. To control alpha inflation, we only evaluated univariate tests of statistical significance if the omnibus tests for the block of variables produced statistical significance. Table 4 lists the number of utterances in 337 that contained a given cue, and the percentage of these cues that were confirmed false. With the exception of non-lexical sounds, all of the predicted T/F discriminators were relatively rare.

Table 4. False vs. True Discriminators

			Binary logistic regression		
Category^a	n^b	(%) False	*B*	*SE*	*p*
Predicted deception cues					
Word/phrase repeats	12	83.3	2.32	1.04	.026
Sp. disfluency spikes	21	71.4	2.15	.78	.006
Protracted headshake	18	77.8	2.41	.83	.004
Nonverbal overdone	22	72.7	1.38	.71	.052
"I don't know"	18	88.9	3.11	1.23	.012
"I/we just"	18	66.7	.61	.80	.445
Long pauses	22	40.9	1.08	.76	.157
Soft/weak vocal	13	69.2	.63	.93	.502
Non-lexical sounds	73	23.3	−1.00	.51	.050
Predicted incriminating potential cues					
Word number			.12	.21	.558
Nonverbal animation			.14	.24	.570
Gesture amount			−.10	.19	.605
Speech speed			.06	.21	.772

[a] Subject, corroboration source, incriminating potential included in analysis but not in Table.
[b] *n* = number of utterances in 337 that contains cue with next column listing percentage of *n* confirmed false. Variables with no entries in first two columns were continuous measures.

Word/phrase repeats, protracted head shaking, nonverbal overdone, speech disfluency spikes and addition of the phrase “I don’t know” discriminated T/F in the predicted direction (presence of cue with corroborated false). Three cues failed to predict T/F (long silent pauses, “I [we] just” and soft/weak vocal quality). Non-lexical sounds proved to be a statistically significant discriminator of T/F but in the opposite direction. Of the 73 utterances containing non-lexical sounds (ums, uhs, sighs), 76.6% were corroborated true. As expected, the behaviors predicted to discriminate IPR (word number, nonverbal animation, gesture amount, and speech speed) did not discriminate T/F, providing additional support for Hypothesis 2.

Incriminating Potential Cues

Hierarchical binary logistic regression analyses tested Hypothesis 3 regarding cues that discriminated IPR. An apparent nonlinear relationship between some predictors and IPR precluded the use of a linear model even as an approximation. However, the pattern of cases across the underlying multidimensional cross-tabulation of the variables precluded multinomial logistic regression. As a means of balancing statistical power against appropriate statistical assumptions, we ran three binary logistic regressions of dichotomous contrasts. Rather than dummy coding the values of IPR, we analyzed 1, vs. 2, 3, 4; 1, 2 vs. 3, 4; and 1, 2, 3 vs. 4 to retain the ordinal properties of the variable.

Parallel to the previous analysis, the first model included T/F, subject (categorical), and the three-category corroboration source (categorical). The second model added the cues hypothesized to discriminate IPR (word number, nonverbal animation, gesture amount, and speech speed). The third model added the cues hypothesized to discriminate T/F (word/phrase repeats, speech disfluency spikes, protracted head shake, nonverbal overdone, long pauses, soft/weak vocal, “I [or we] just” qualifer, addition of “I don’t know” phrase, and non-lexical sounds).

Table 5 summarizes the model fit statistics for the three models for the dichotomies (1, vs. 2, 3, 4; 1, 2 vs. 3, 4; 1, 2, 3 vs. 4). The second block of variables

Table 5. Binary Logistic Regression: Incriminating Potential Discriminators

	IPR 1 vs. 2, 3, 4			IPR 1, 2 vs. 3, 4			IPR 1, 2, 3 vs. 4		
Block variables	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1 Subject (Cat.), true vs. false, corroboration source (Cat.)									
Block test	114.64	30	.000	92.20	30	.000	75.52	30	.000
Model test	114.64	30	.000	92.20	30	.000	75.52	30	.000
2 Predicted IP cues[a]									
Block test	20.46	4	.000	29.16	4	.000	26.38	4	.000
Model test	135.94	34	.000	121.36	34	.000	101.91	34	.000
3 Predicted deception cues[b]									
Block test	8.86	9	.450	6.71	9	.668	20.41	9	.016
Model test	143.95	43	.000	128.07	43	.000	122.32	43	.000

[a] Animation level, gesture amount, word number, speech speed.

[b] Word repeats, sp disfluency spikes, protracted headshake, nonverbal overdone, “I don’t know” phrase, “I (or we) just” qualifier, long pause, vocal weak, and non-lexical sounds.

Table 6. Incriminating Potential Discriminators

	IPR 1 vs. 2, 3, 4			IPR 1, 2 vs. 3, 4			IPR 1, 2, 3 vs. 4		
Category[a]	*B*	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
True vs. false	24.70	113.41	.828	1.13	.42	.007	.56	.36	.120
Predicted stress cues									
Word number	−.25	.50	.612	−.22	.20	.267	−.25	.19	.173
Nonverbal animation	−.50	.60	.401	.53	.24	.027	.57	.21	.008
Gesture amount	2.00	.98	.042	.41	.22	.059	.25	.17	.146
Speech speed	1.34	.42	.002	.45	.19	.017	.29	.17	.097
Prediction deception cues									
Word/phrase repeats	10.70	228.76	.963	.93	1.38	.502	−.60	.90	.503
Speech disfluency spikes	4.83	60.80	.937	.51	.77	.509	.35	.61	.560
Protracted head shake	10.96	189.03	.954	.53	.80	.511	.17	.66	.798
Nonverbal overdone	10.45	249.54	.967	.13	.91	.888	−.25	.68	.710
"I don't know"	−14.22	91.04	.876	.45	.96	.638	1.48	.75	.047
"I (or we) just"	7.99	304.76	.979	1.98	1.21	.101	3.61	1.26	.004
Long pauses	2.53	1.66	.128	1.04	.79	.189	.66	.67	.327
Soft/weak vocal	.47	368.58	.999	−.03	.96	.977	−2.20	1.07	.039
Non-lexical sounds	−1.28	.89	.152	−.02	.43	.957	.32	.40	.422

[a] Subject, corroboration source in analysis but not in Table.

improved predictions statistically significantly over T/F, subject dummy codes and corroboration source for each of these dichotomies. This provides general support for Hypothesis 3. Further, recall from the previous analysis that these cues failed to improve prediction of T/F when added into the third model in that analysis. This result provides further support for the hypothesis because cues that collectively discriminated IPR failed to discriminate T/F. Likewise, the T/F discriminating cues failed to add to the prediction of the IPR dichotomies 1 vs. 2, 3, 4 and 1, 2 vs. 3, 4, further support for Hypothesis 2.

Table 6 reports the regression weights for the three models predicting the three pairs of IPR. Again, to control alpha inflation we only inspected univariate tests for specific regression weights for variables included in blocks of variables with statistically significant omnibus tests. Nonverbal animation distinguished IPR for dichotomies 1, 2 vs. 3, 4 and 1, 2, 3 vs. 4. Gesture amount distinguished IPR for the dichotomy 1 vs. 2, 3, 4 and tended to distinguish 1, 2, vs. 3,4 ($p = .059$). Speech speed distinguished IPR for 1 vs. 2, 3, 4 and 1, 2 vs. 3, 4. Word number per utterance did not discriminate IPR for any of the dichotomies. Two cues that we predicted would discriminate T/F proved to be IP cues instead. Both "I[or we]just" and soft/weak vocal discriminated IPR for 1, 2, 3 vs. 4. As expected, no deception cues involved continuous degrees of change. However, against expectation, IPR was related to both continuous change cues and isolated event cues, including one with lexical content ("I [or we] just").

Overall, the results confirmed Hypothesis 2 that some behaviors would discriminate deception but not IP, and Hypothesis 3 that a different set would discriminate IP but not deception. The basic argument that cues to deception should not be confounded with cues to incriminating potential remains supported by the results.

Mixed Cues

We also hypothesized that some cues would discriminate both IP and deception but did not predict what they would be. The phrase "I don't know" added to an answer discriminated both deception and IPR 1, 2, 3 vs. 4 which gives tentative support for Hypothesis 4.

Individual Differences

In our data, utterances displayed at most three cues, and 70% of the 69 utterances with deception cues displayed only one type. Individuals varied in which deception cues they displayed, but the cues were observed in from 9 to 13 subjects, and 27 of the 28 subjects displayed at least one deception cue.

Degree of Accuracy

How well do the results of the microcoding discriminate true and false utterances compared with detection accuracy rates of judgment studies? To examine this we combined cues that discriminated T/F in one direction (word/phrase repeats, protracted head shake, speech disfluency spikes, nonverbal overdone, and "I don't know" phrases) into a dichotomous (none/some) variable called false cues. True accuracy (212 corroborated true utterances with no cues over the 229 corroborated true) was 92.6% and false accuracy (52 corroborated false with cues over 108 corroborated false) was 48.1%. An overall accuracy of 78.3% (true plus false hits over 337) was considerably higher than the detection accuracy of judges that in research studies hovers around 55% (Vrij, 2000).

The above indices of accuracy rest on a fixed cut-score, in this case corresponding to 1 or more cues. Different cut scores produce different classification tables and different percentages. The area under an ROC curve (AUC) provides a useful cut-score free index of predictive accuracy. We plotted the outcome variables against the predicted probability values from the logistic regressions to assess accuracy through AUC. Predicting T/F, this yielded AUC = .895, $SE = .02$ (95% CI from .86 to .93). Predicting IP 1 vs. 2, 3, 4, the same procedure yielded AUC = .977, $SE < .01$ (95% CI from .96 to .99). Predicting IP 1, 2 vs. 3, 4, AUC = .760, $SE = .03$ (95% CI from .71 to .81) and predicting IP 1, 2, 3 vs. 4, AUC = .626, $SE = .03$ (95% CI from .56 to .69). The last two IP results were lower than those for T/F but still well above the .50 baseline.

DISCUSSION

Intuitively, it follows that the hotter the topic, the more likely one will lie about it, if one is going to lie. There was partial support for Hypothesis 1 that the proportion of false answers increases as the topic becomes potentially more incriminating. The false utterances did not appear to occur randomly as if they were unwitting errors or differences of opinion, but in a self-serving pattern. Executives from the District Attorneys Office determined which videotapes could be studied and, not surprisingly, they selected cases supported by a great deal of evidence and information. Also, the interviews were conducted soon after the crimes, so memory errors would seem unlikely. We therefore have treated the false answers as intentional lies and called true vs. false discriminators deception cues. Nevertheless, our method of corroborating true and false utterances had limits. Although witnesses had good exposure to the individuals they described, the reliability of witness accounts and self-reports remains a serious question. Refined methods for corroborating veracity are critical for future research on high stakes interviews.

We did not anticipate a higher proportion of false utterances at IPR 3 than at IPR 4, but this may be an artifact of the data collection procedures. In the first batch of 21 interviews, the proportion of false responses in IPR 3 was 35.7 and 39.2% in IPR 4, a slight increase. However, in the 7 interviews of the second batch, the proportion of false in level 3 was 66.7% dropping to 32.5% in level 4. It is possible that this difference was due to differences in how the true and false utterances were corroborated. It was difficult to corroborate IPR 3 true utterances from the records (the method used for the second batch). In other words, the drop from IPR 3 to IPR 4 in percentage false may be due to low numbers of true utterances at IPR 3, and indicate how strongly the distribution of true and false utterances at each level can be affected by the method of corroborating segments selected from naturalistic recordings.

Incriminating Potential vs. Deception Cues

Overall, our results support a model in which deception and psychological stress are separate, albeit related, processes. As hypothesized, some behaviors discriminated IP only, while others distinguished T/F alone. One cue discriminated both T/F and IP.

As predicted, the deception cues tended to be discrete, relatively rare, and lexical (e.g., a nonverbal cue with a word-equivalent such as head shaking “no”). However, against expectation, IP cues could be either continuous measures of rate or intensity (e.g., speech rate and degree of animation) or discrete, relatively rare behaviors with or without lexical content (e.g., the “I (or we) just” qualifier). Some of the deception cues appeared to be micro signs of protesting too much in Shakespeare’s sense (e.g., protracted head shaking, repeating phrases). In past research, deception cues have typically been forms of control, affective contradiction, speech disruption or nervous behaviors. This study indicated that various forms of protesting-too-much – verbal and nonverbal – are important cues in high stakes interviews and warrant further investigation.

Non-lexical sounds (ums, uhs, sighs, gutturals) occurred primarily with true responses, not false ones in this study. In research on speech disturbances and anxiety, Kasl and Mahl (1965) have contended that non-lexical sounds such as ums and uhs function differently from other forms of speech disturbance and we would concur. Non-lexical sounds occurred in a pattern different from all the other behaviors in (Markus, Davis, & Walters, in preparation). In this study three or more repetitions of a phrase and/or consecutive word within a brief utterance was associated with deceptive responses. Non-lexical sounds, most of which involved two or more ums and uhs, were associated with true responses. If speech disturbances such as these serve cognitive functions, then our data suggest that different disturbances serve different functions. Future research could test, for example, whether non-lexical sounds such as umms and uhhs aid retrieval of truthful information or, at least, taking time to find a safer description or euphemism, while exact word/phrase repetitions serve taking time to construct a deceptive answer. At the very least, our results indicate why, when non-lexical sounds are combined with other forms of speech disturbance as in Mann et al. (2002), the more inclusive variable does not discriminate deception.

Study Limitations

While no voices were raised and the style of interviewing was quiet and polite, there is little question that the stakes were high and the suspects were highly motivated to cast themselves in a better light. On the whole the confirmed false answers tended to minimize the speaker's role in the crime, while other statements admitted some involvement. That is, the lies appeared to be attempts to lessen the seriousness of the accusations, with only a few suspects making the case for complete innocence. Nevertheless, questions about intention and motive qualify our results – including why any suspect would truthfully admit incriminating information, and how IP and deception cues relate to compliance behavior. Also, we cannot assess the extent to which our results were skewed because we could only obtain tapes of subjects judged guilty. We would expect the behavior of a guilty suspect who confessed for personal gain, psychological need, or compliance-with-authority motives to be very different from the behavior of an innocent suspect who was coerced or psychologically motivated to make a false confession. The same general admission should be communicated in quite different ways, with different demeanor cues. Also, for legal reasons, we could not compare the false denials and fabrications of criminal suspects found guilty with the accurate denials and narratives of suspects who were proven innocent and whose records were therefore sealed. It is critical to compare statements made by convicted suspects, exonerated suspects and innocent suspects who make false confessions, but there are major legal and protection of human subject issues to be addressed before this can be done. Exonerated prisoners who consent to research use of their interrogation or confession tapes might make such comparisons possible in the future. Until then, studies on convicted suspects, such as this one, must be interpreted and applied to law enforcement very cautiously. 7.4% of the corroborated true utterances contained false cues and up to half the lies were missed in this dataset. For investigators conducting such interviews, the deception cues we identified are, at most, possible leads requiring investigation.

Comparisons with Past Research

Whether our results generalize to other groups remains an empirical question that our research design necessarily leaves to plausibility arguments for external validity rather than statistical generalization. The present study involved behavioral cues accompanying brief utterances about events and actions during very high stakes interviews. Our focus on brief utterances is important because many interview situations contain brief lies within otherwise truthful narratives. There is reason to predict that longer passages such as an elaborate fabrication during an interrogation would have additional cues to incriminating potential and deception as well as greater frequency of the cues found in the brief utterances (Walters, 1996). The list of cues in Table 2 is by no means exhaustive. How cue displays in brief utterances differ from displays in longer passages is an empirical question. We would also expect distinctions between deception and IP cues to occur in longer speech segments and with different types of high stakes interviews. However, the present study cannot confirm or disconfirm this conjecture.

The ADA interviews followed interrogations by detectives or police officers and this raises another issue affecting external validity. Would the cue displays differ on second telling and with the shift from investigator to prosecutor? This is a complex question. While the element of practice might decrease deception cue displays, the pressure to recall what one said and be consistent could increase them. Again, these are questions for future research.

Like DePaulo et al. (2003), we identified word repetition as a deception cue, but beyond this, our results appear to diverge from past research. Unlike Mann et al. (2002), we did not find a relationship between silent pauses and deception, not for short (.5–1.5 s) pauses in the original group of 21 interviews (Davis et al., 2000) nor for long (2 or more s) pauses in the total group of 28 interviews. However, we studied only confessions for which pausing before or during a serious admission was as common as pausing before lying. Mann et al. (2002) studied both confessions and interrogations in which suspects appeared to be denying involvement, and this may be why the number of pauses with lies was greater than pauses with serious admissions.

Increased cognitive overload and high motivation to lie have been associated with decrease in animation and greater behavioral control (DePaulo, Kirkendol, Tang, & O'Brien, 1989; Ekman & Friesen, 1972), and control-type deception cues have often been noted in the research literature (cf Buller & Aune, 1987). We found little evidence of control-type deception cues in what was certainly a context of high cognitive overload and motivation to lie. We did find some behaviors that might be related to control in the sense of buying time to fabricate an answer (exact word or phrase repeats), but in our subjects, restricted motion was not related to deception. For example, we did not find that speech gesticulating decreased with deception as reported in Vrij, Akehurst, and Morris (1997). Our subjects showed a tendency to gesticulate more when the topics were more incriminating. This is one example where a cue found related to deception in the experimental literature turned out to be a cue related to IP, not deception in our study. To cite another example, Buller, Burgoon, White, and Ebesu (1994) found that speakers displayed higher ratings of "kinesic expressiveness" (expressive, animated, impassive) when lying than when telling the truth. In the confession tapes, nonverbal animation level was related to IP, not deception.

Future Research

In many experimental cue studies, those who volunteer to lie or tell the truth are asked to address the same topic, so there is no way to compare cues to topic incriminating level or other psychological stress potentials with cues to deception. We found small changes in behavior that were related to subtle shifts in the IPR of brief utterances. We would argue for designing experimental deception cue research in such a way that both cues to deception and topic incriminating potential can be assessed and compared, e.g., by asking participants to construct their own alibis (cf Porter & Yuille, 1996).

There are many reasons why a cue may not generalize across contexts and from one study to another. Our results suggest that conflating T/F and IP cues is one of the reasons. Comparing evoked lies with relatively non-threatening truthful statements may be another. For example, speech speed and animation level were IP cues in the confession tapes. The personal stakes for participants in deception experiments are necessarily lower than the stakes for suspects in the confession tapes, but it is quite possible that a cue which is actually related to psychological stress will distinguish lies from true statements in an experimental condition because telling a lie is all that is personally threatening and following experimenter instructions to tell the truth entails relatively minor performance or compliance challenges for the participant. In the confession tapes telling the truth was often enormously threatening, hence we could investigate whether there were different cues for two different types of threatening answers, true vs. false. Experimental research on deception has been particularly attentive to the importance of motivation and whether the consequences of telling a lie are sufficient to elicit cues (DePaulo et al., 2003). But equally important is the need to compare lying with threatening rather than non-threatening truth-telling.

There has been relatively little attention to – or at least little explicit discussion in the literature about – how the experimental task and verbal content may determine the types of cues that occur. For example, the Ekman et al. (1988) study of nurses was a situation in which it made sense to search for behavioral cues to negative affect that would contradict narratives of pleasant events. It is possible that direct and unambiguous contradictions are more likely to occur with descriptions of affect or attitude (e.g., "I am sure of it" with hesitant speech or "this is a pleasant scene" with facial expression of disgust) than with the accounts of events and actions that were the focus of the present study. Suspects in the confession tapes sometimes demonstrated holding the gun as they admitted using it, but we did not see clear contradictions between speech and action, (e.g., someone pulling the trigger finger two times while saying "I just shot him once.") Most of the deception cues that we identified were related to explicit or implicit denials or subtle ways of protesting too much.

Detection accuracy based on combining cues was relatively high considering the brevity of the utterances, the individual differences as to which cues were displayed, and the variation in subject matter between interviews. For several reasons (tape quality, reliability issues, coding resources), our coding was limited and did not include many cues that could prove valid discriminators of stress or deception in better quality tapes. Adding reliable deception cues to the combination should increase detection accuracy (Ekman et al., 1991).

The study showed that fine-grained coding of multiple behavioral cues in relation to measures of both veracity and incriminating potential was a viable and productive approach to deception detection in real, high-stakes interviews, despite the enormous labor required. Markus, Davis, and Walters (in preparation) demonstrate behavioral profiles for true responses that differ in content and another set of behavioral profiles for false answers that differ in content. Advances in understanding the nature of deception processes and how demeanor reflects them in real-world contexts requires multivariate studies of deception cues in relation to both incriminating potential and specific contents.

ACKNOWLEDGMENT

This study began as part of a US Government-sponsored project on the Analytic Potentials of Nonverbal Communication directed by Brenda Connors. Continuing support has been provided by John Jay College of Criminal Justice Presidential Research Grants 2002–2003. We thank Jack Zlotnick, Dan Juda and Gabrielle Salfati of John Jay for their help. We are indebted to NYC District Attorney executives for permission to study the videotapes, to the detectives who provided background information, and to our consultant, Det. (ret.) Raymond Pierce, formerly of the New York City Police Department. We are grateful to faculty from the Laban/Bartenieff Institute of Movement Studies who contributed expertise in nonverbal observation and to the John Jay graduate students who were behavioral coders and research assistants.

REFERENCES

Anolli, L., & Ciceri, R. (1997). The voice of deception: Vocal strategies of naïve and able liars. *Journal of Nonverbal Behavior*, *21,* 259–284.

Baldwin, J. (1992). Video taping police interviews with suspects: An evaluation. London: Police Research Series: Home Office Police Department.

Bavelas, J. B., Black, A., Chovil, N., & Mullett, J. (1990). *Equivocal Communication*. Newbury Park: Sage Publications.

Buller, D. B., Burgoon, J. K., White, C. H., & Ebesu, A. S. (1994). Interpersonal deception VII. Behavioral profiles of falsification, equivocation, and concealment. *Journal of Language and Social Psychology*, *13,* 366–395.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

Davis, M., & Hadiks, D. (1995). Demeanor and credibility. *Semiotica*, *106,* 5–54.

Davis, M., Walters, S. B., Vorus, N., Meiland, P. A., & Markus, K. A. (2000). *Verbal and nonverbal cues to false testimony in criminal investigations*. Paper presented at the American Psychological Association Convention, Washington, DC.

DePaulo, B. M., Kirkendol, S. E., Tang, J., & O'Brien, T. P. (1989). The motivational impairment effect in the communication of deception: Replications and extensions. *Journal of Nonverbal Behavior*, *12,* 177–201.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). *Cues to deception. Psychological Bulletin*, *129,* 74–118.

DeTurck, M. A., & Miller, G. R. (1985). Deception and arousal: Isolating the behavioral correlates of deception. *Human Communication Research*, *12,* 181–201.

Ekman, P. (1988). Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior*, *12,* 163–175.

Ekman, P. (1992). *Telling Lies*. New York: W. W. Norton.

Ekman, P., & Friesen, W. V. (1972). Hand movements. *Journal of Communication*, *22,* 353–374.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54,* 414–420.

Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. R. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, *15,* 125–135.

Hess, E. H., & Polt, J. M. (1963). Pupil size in relation to mental activity during simple problem-solving. *Science*, *140,* 1190–1192.

Horvath, F. S., Jayne, B., & Buckley, J. (1994). Differentiation of truthful and deceptive criminal suspects in behavior analysis interviews. *Journal of Forensic Sciences*, *39,* 793–807.

Kasl, S. V., & Mahl, G. F. (1965). The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, *1,* 425–433.

Knapp, M. L., & Comadena, M. E. (1979). Telling it like it isn't: A review of theory and research on deceptive communications. *Human Communication Research*, *5,* 270–285.

Kraut, R. E. (1980). Humans as lie detectors: Some second thoughts. *Journal of Communication*, *30,* 209–216.

Landis, J. R., & Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics*, *33,* 671–679.

Mann, S., Vrij, A., & Bull, R. (2002). Suspects, lies, and videotapes: An analysis of authentic high stakes liars. *Law and Human Behavior*, *26,* 365–376.

Markus, K. A., Davis, M., & Walters, S. B. (2004). A behavioral typology of deception in criminal confessions. Manuscript in preparation.

Miller, G. R., & Stiff, J. B. (1993). *Deceptive communication.* Newbury Park: Sage Publications.

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior*, *20,* 443–458.

Sapir, A. (1987). *The LIS course on scientific content analysis (SCAN).* Phoenix, AZ: Laboratory for Scientific Interrogation.

Shuy, R. W. (1998). *The language of confession, interrogation, and deception*. Thousand Oaks, CA: Sage Publications.

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.

Vrij, A. (2000). *Detecting Lies and Deceit*. Chichester: Wiley.

Vrij, A., Akehurst, L., & Morris, P. (1997). Individual differences in hand movements during deception. *Journal of Nonverbal Behavior*, *21,* 87–102.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, *24,* 239–264.

Vrij, A., & Mann, S. (2001). Telling and detecting lies in a high-stake situation: the case of a convicted murderer. *Applied Cognitive Psychology*, *15,* 187–203.

Walters, S. B. (1996). *Principles of kinesic interview and interrogation*, Boca Raton: CRC Press.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology.* (Vol. 14, pp. 1–59). San Diego, CA: Academic Press.