Journal of Police and Criminal Psychology (2018) 33:302–315
https://doi.org/10.1007/s11896-017-9250-0

Clusters of Nonverbal Behaviors Differ According to Type of Question and Veracity in Investigative Interviews in a Mock Crime Context

David Matsumoto[1,2] · Hyisung C. Hwang[2]

Published online: 22 December 2017

Abstract
Evaluating truthfulness and detecting deception is a capstone skill of criminal justice professionals, and researchers have long examined nonverbal cues to aid in such determinations. This paper examines the notion that testing clusters of nonverbal behaviors is a more fruitful way of making such determinations than single, specific behaviors. Participants from four ethnic groups participated in a mock crime and either told the truth or lied in an investigative interview. Fourteen nonverbal behaviors of the interviewees were coded from the interviews; differences in the behaviors were tested according to type of question and veracity condition. Different types of questions produced different nonverbal reactions. Clusters of nonverbal behaviors differentiated truth tellers from liars, and the specific clusters were moderated by question. Accuracy rates ranged from 62.6 to 72.5% and were above deception detection accuracy rates for humans and random data. These findings have implications for practitioners as well as future research and theory.

Keywords Deception · Nonverbal behavior · Facial expressions · Voice · Gestures · Truthfulness

Introduction

Conducting investigative interviews is an important part of the criminal justice process, and evaluating truthfulness, detecting deception, and assessing credibility are important determinations made during these interviews. For decades, researchers have examined nonverbal behavior (NVB) for clues to such determinations because they are dynamic actions of the face, voice, and body that communicate messages. Early studies produced preliminary evidence for facial expressions, vocal characteristics,[1] body movements, and gestures to differentiate truths from lies (e.g., see Ekman et al. 1988, 1991; Mehrabian 1971; Streeter et al. 1977), but later studies produced conflicting or null results (e.g, Hocking and Leathers 1980; Klaver et al. 2007; Vrij et al. 2000). Subsequent meta-analyses have corroborated this mixed picture, suggesting that the ability of NVB to differentiate truths from lies is equivocal (DePaulo et al. 2003; Sporer and Schwandt 2006, 2007).

Most of the research to date has examined specific, single NVB. A handful of studies, however, have suggested that clusters of NVB (sometimes with words), instead of single behaviors, can reliably differentiate truths from lies (see review by Vrij 2008). Repeated words and phrases, speech dysfluencies, and head shaking, for instance, discriminated between true and false statements in 28 videotaped confessions by criminal suspects (Davis et al. 2005). Speech disturbances, hand and finger movements, and response latencies differentiated between true and false statements by nurses who witnessed a videotape of a theft (Vrij et al. 2000). And a combination of facial expressions and vocal pitch produced high and significant deception detection accuracy rates (Ekman et al. 1991).

The notion that clusters are better than single behaviors in differentiating truths from lies is rooted in a consideration of the complexity of and degrees of consciousness about one's cognitions and emotions, and that NVBs are signals of these cognitions and emotions. At any one time, the mind is replete with multiple thoughts and feelings that can and often exist simultaneously, particularly when interacting with others.

[1] Consistent with many other writers in this area (e.g., Hirschberg 2002; Scott and McGettigan 2016), we consider vocal characteristics, including vocal pitch (tone), rate, intensity, response latencies, response durations, and the like, as a subset of NVB. Contrastingly, verbal behavior focuses on the messages associated with verbal content (words).

✉ David Matsumoto
dm@sfsu.edu

[1] Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, USA

[2] Humintell, El Cerrito, CA, USA

Some thoughts and feelings are complementary to each other, some not, and they may be directly, indirectly, or unrelated to each other. People are consciously aware of only a portion of these thoughts and feelings and direct their limited cognitive resources to verbalizing certain, specific, selected thoughts and to managing their appearance or the impressions of others about themselves (for extended discussions, see Baumeister and Masicampo 2010; Lambie and Marcel 2002; Murphy and Zajonc 1993).

Thus, spoken words do not reflect the entire contents of one's mind, only part of it. There are many other cognitive and emotional contents of the mind that are not spoken; some may be consciously accessible but much is not. The existence of multiple cognitions and emotions in the mind suggests that multiple and different NVBs may be associated with these cognitions and emotions, because NVBs are signals of cognitions and emotions. That is, unspoken, diverse, complex, and at least partially unconscious mental contents may be expressed nonverbally through multiple channels such as the face, voice, gesture, body posture, gait, or interpersonal space (Matsumoto et al. 2013). These NVBs may embody cognitions or signal emotions, and they may be complementary to each other or not and related to each other or not. NVB may or may not be directly related to the content of words expressed verbally and may occur outside of conscious awareness. Thus, NVB can be (and should be) quite complex because the cognitive and emotional contents of the mind are complex and varied.

Truth telling and lying occur within this process described above. Lying facilitates certain types of cognitions and emotions, thereby introducing cognitive and emotional loads that differ from truth telling (Ekman 1985; Frank 2009; Vrij 2008). These loads, in turn, influence the NVB that signals those cognitions and emotions. Lying requires not only knowledge of the truth, but also knowledge about the contents of one's lies, whether fabricating false information or omitting knowledge; thus, liars must lie about their cognitions. Liars also have thoughts about the fact that they are lying, requiring the liar to remember where and how the lies are being (or were or will be) perpetrated, as well as the consequences of being caught (or not); thus, liars have additional cognitions about their lies. Additional cognitions about lies result in additional emotions about the fact that one is lying; thus, liars have emotions about their lies. And to the extent that emotions are associated with the truth and about the knowledge that one is lying, lying requires falsification of those feelings; thus, liars must lie about their emotions.

These characteristics of the mental states of liars—cognitions about lying, lying about cognitions, emotions about lying, and lying about emotions—suggest a very complex picture of how NVB may function when people lie. Signs of any of these mental states may occur across multiple nonverbal channels, are transient and fleeting, and wax and wane across time in an interview. They may be different for each individual, and even within each individual, different across contexts and questions. Moreover, different types of signals occur in different parts of the body; faces signal specific emotions (Hwang and Matsumoto 2016), the voice and body signal general affective orientations (Matsumoto et al. 2016; Scott and McGettigan 2016), the head and hands signal specific verbal phrases (Cartmill and Goldin-Meadow 2016), and the face and body signal cognitive processes (Hwang and Matsumoto 2016; Matsumoto et al. 2016). Thus, examination of any one channel alone may not differentiate truth tellers from liars as well as clusters would because clusters cast a broader net of signal sources with which to capture possible leakage (i.e., the expression of unconscious or suppressed mental states) of the various cognitive and emotional states that occur dynamically.[2]

This framework better explains what has been found in the literature. Examinations of a single NVB may sometimes lead to positive findings and sometimes not, which is exactly what meta-analyses have reported (DePaulo et al. 2003). Examinations of clusters of NVB should more reliably differentiate truths from lies or provide higher deception detection accuracy rates, which is what the few available studies have reported (Davis et al. 2005; Ekman et al. 1991; Vrij et al. 2000). Thus, testing the ability of multiple nonverbal channels simultaneously may be a more fruitful and realistic methodology than the testing of single channels in isolation, as there is no guarantee that a single channel is consistently available in reality or accessed when people tell the truth or lie.

Another factor that should influence the differential production of NVB between truth tellers and liars in investigative interviews is the types of questions asked. Different types of questions should elicit different cognitions and emotions for truth tellers and liars, which should produce different NVB. For example, open-ended questions (e.g., "Tell me what happened") are different than direct questions about lying ("Did you take the money?"). In open-ended requests, cognitions and emotions are tied to a memory and should vary according to the contents of the story. For example, someone who has stolen money and lies is likely to experience the cognitions and emotions that occurred at the time of the theft (e.g., fear, exhilaration, nervousness, or joy), because these cognitions and emotions would be encoded in memory just as facts are (this perspective is consistent with the reality monitoring framework; see Johnson 1988; Johnson and Raye 1981). At the same time, that individual will likely engage in recall and

[2] In fact, the same argument could be made for verbal cues to lying. Research has demonstrated that a number of verbal cues—both related to content and to grammatical and linguistic features of speech—can differentiate truth tellers from liars. But this literature also shows that no one single, specific verbal cue can differentiate truth tellers and liars reliably; instead, this literature has shown that multiple, different types of verbal cues can differentiate truths and lies (Deeb et al. 2017; Hwang et al. 2016; Matsumoto et al. 2015a, b; Vrij et al. 2011), akin to clusters of NVB.

reporting strategies to regulate embodiments of those cognitions and emotions as if to appear truthful. Thus, that same individual will have the additional burdens of having emotions about the fact that they are lying and having to lie about their emotions and the additional cognitive loads associated with knowing they are lying and lying about what they know.

Closed-ended, direct questions that require a yes/no response (e.g., "did you take the money?" or "are you lying to me now?") are different. These may not be as varied or complex as open-ended questions because direct questions require only a yes or no answer. Responding to such questions does not require accessing as many multiple, complex, and varied cognitions and emotions, and as a response strategy individuals need focus only on a simple yes or no response. Cognitive and emotional leakage in NVB may occur, but the influence of these direct questions is likely different than those of open-ended ones.

Individual differences ensure that different people have different cognitive and emotional reactions to these complexities; thus, the leakage that may occur when responding to open-ended questions may be complex and varied. For instance, liars may leak their fears if they actually felt fear either during the incident or were afraid of being caught, or they may leak their joy if they were exhilarated during the incident or if they were joyful about the fact they were lying. They may not be able to hide their momentary feelings of disgust toward the interviewer for being in the interview and having to respond to questions.

The framework introduced above suggests that different types of cognitions and emotions that are recruited by different types of questions should be associated with different clusters of NVB. Examining NVB across questions compounds the difficulty for NVB to differentiate truths from lies because doing so likely washes out any differentiability due to the uniqueness of questions. For example, one type of question may cause liars to become angrier than truth tellers, while another type of question may cause truth tellers to become angrier than liars. If data are averaged across questions, a typical data analytic strategy, anger would likely not differentiate truths from lies. Testing the possibility of NVB to discriminate truth tellers and liars, therefore, requires researchers to examine reactions separately for each question, or for different types of questions.

Another factor to consider in this line of research is that interview quality is easily contaminated by nuisance factors. Among these are times when interviewees do not understand the questions asked, or when the interviewer impedes or negatively influences the interview process. For example, interviewers may misstate or rearrange words of a question so that the meaning of the original question is altered, may interrupt an interviewee, or may interject words ("keep going," "go on") during a response. Interviewers sometimes even volunteer words to help interviewees complete a response. Because NVB are transient reactions, these nuisance factors can blur the ability of NVB to differentiate truths from lies.

Thus, the nature of the question asked of interviewees needs to be considered when examining NVB associated with veracity and deception, and the quality of the interview process itself requires integrity. With few exceptions (Anolli and Ciceri 1997; Reynolds and Rendle-Short 2011; Vrij et al. 2007), however, most studies have not examined how differences in NVB as a function of truth telling and lying may be moderated by the type of question asked, nor has any study systematically characterized different types of questions or checked for interview quality. Here, we categorize three types of questions and examine if differences in clusters of NVB between truth tellers and liars are moderated by these question types. We also control for the quality of the interviews by coding for interview contamination.

Participants from four ethnic groups—European Americans, Chinese, Hispanics, and Middle Easterners—either stole a check and lied about it or did not and told the truth. After being assigned to either the steal-lie or do not steal-truth condition, each participant engaged in three interviews, all in English, two prior to committing the crime and one afterwards (the investigative interview). We examined three categories of NVB participants produced during the third investigative interview to test whether or not they differentiated truths from lies, and if question type moderated those differences. The three categories of NVB examined were facial expressions of emotions (six types), gestures (three types), and vocal characteristics (five types). These NVBs have been commonly tested in previous studies on deception (DePaulo et al. 2003).

The three immigrant samples were included in order to test for ethnic group differences in NVB produced during the interviews. Ethnicity is often a marker of cultural differences (Matsumoto and Juang 2016), and the ethnic groups sampled represented some of the same ethnic groups in which differences in expressivity and cultural norms for emotional expression have been documented within the USA (Matsumoto 1993; Tsai and Levenson 1997; Tsai et al. 2000a). The ethnic groups in this study were also representative of the cultural and ethnic diversity that law enforcement officers in the USA (and other multicultural societies) face.

We tested the following hypotheses:

> Hypothesis 1: That NVB would differ as a function of type of question. More specifically, we predicted that closed-ended, direct questions would produce the least amount of NVB than other question types.
> Hypothesis 2: That different clusters of NVB would reliably differentiate truths from lies (2a), and that type of question asked would moderate this effect (2b). That is, the specific NVB that differentiated truths from lies would be different for different types of questions.

Method[3]

Participants

Participants from the four ethnic/cultural groups (European Americans and Chinese, Hispanic and Middle Eastern immigrants) were recruited from student and non-student communities in the San Francisco Bay Area and Buffalo, NY, through ads (online and print) seeking "European American," "Chinese," "Hispanic," or "Middle Eastern" participants. The European Americans were all US born-and-raised Caucasians. Participants in the other three ethnic groups were immigrants born and raised in their home country or born in the USA but whose first language was that of the home country and whose parents were both born and raised in the home country. Home country was defined for Chinese as the People's Republic of China, Hong Kong, or Taiwan, and the first language was Mandarin or Cantonese; for Hispanics, home country was any country in Central or South America, and the first language was Spanish; and for Middle East, home country was any country in Northern Africa or Western Asia, and the first language was Arabic.

Two hundred twenty-six individuals participated for cash payment (standard participation fee was $20, with the possibility of receiving more depending on outcomes described below). The European Americans included $n = 40$ and 38 in the lie and truth conditions, respectively; the Chinese, Hispanics, and Middle Eastern samples included $n = 46$ and 36, $n = 28$ and 18, and $n = 8$ and 12, respectively.[4] Seventy percent of the sample was comprised of students, none of whom participated in a similar study. They were roughly evenly distributed between males (47.4%) and females (52.6%) with an average age of 27.32 (range 19–47) across the four ethnic groups and within conditions. All experimental procedures occurred in English.

Measures

Participants completed a basic demographics questionnaire (with questions reconfirming inclusion criteria), the General Ethnicity Questionnaire (GEQ; Tsai et al. 2000b), an emotion checklist, the Machiavellianism Scale (Christie 1970), and the Self-Monitoring Scale (Snyder 1974). Participants also completed the emotion checklist at the end of the experiment. This checklist included 12 emotion words (guilt, fear, anger, embarrassment, worry, contempt, excitement, disgust, amusement, nervousness, surprise, and interest) rated on nine-point scales labeled 0, none; 4, moderate amount; and 8, extremely strong.

The GEQ is a commonly used scale to measure acculturation and ethnic identity and was included as a manipulation check for ethnic/cultural differences. This questionnaire contained 38 statements, 25 rated on a five-point Likert scale, and 13 rated on a five-point scale from very much to not at all. The target group mentioned in the GEQ was modified to be applicable to each ethnic group. Analyses of the GEQ total score, which was the mean of all items after reverse coding those negatively loaded, indicated that the Chinese sample had significantly higher scores than American born Chinese and Chinese who immigrated to the USA before the age of 12 reported by Tsai et al. (2000b), $t(74) = 8.07$, $p < .001$, $d = .93$; $t(74) = 1.71$, $p < .05$, $d = .20$, respectively. Thus, our Chinese sample identified themselves as Chinese and with Chinese culture more so than American born Chinese. Although norms for Hispanics and Middle Easterners do not exist, their scores were comparable to the Chinese in our sample.

Procedures

Participants were introduced to the study and told that they would be randomly assigned to either take a $100 check or to look at but not take the check. They were also told that they had to convince all interviewers of their honesty and innocence. The stakes associated with the experiment were explained (below). Participants then completed the pre-session measures. When done, the experimenter conducted a random assignment to veracity condition procedure in view of the participants. Participants were reminded of their condition and the stakes associated.

Participants were then escorted out of the instruction area and, after a short wait, were greeted by an interviewer, who conducted an initial screening interview. After this was completed, participants were informed that they were selected for a second interview. After this second interview, participants were left alone to go to a room and steal or not steal the check. After returning, the participants were informed they had been selected for a third interview and were escorted to a separate interview room. An interviewer entered and conducted this investigative interview, which was the focus of this paper and the analyses below. Upon completion of this interview, the participant was escorted back to the instruction area, completed post-session measures, debriefed, paid, and excused. A different interviewer conducted each interview.

Interviewers, Questions, and Question Types

Ten male actors, all above the age of 30, served as interviewers. All received training to deliver the interviews in a neutral and objective manner and to stick with the

[3] Portions of the methodology have been previously reported in Matsumoto and Hwang (2015) and Matsumoto et al. (2015b).

[4] Sample sizes for specific analyses reported below differed because of differing missing cases occurring because of technical issues in the various methods of data extractions.

predetermined interview questions. The first author also served as an interviewer.

The questions used were modeled after questions used in real-life security and investigative interviews, and were developed after consultation with subject-matter experts (SMEs) from various law enforcement entities with interests in the practical application of the findings. For the post-event investigative interview, we incorporated questions typically used by law enforcement officers as well as questions based on the Strategic Use of Evidence (SUE) (Hartwig et al. 2005, 2006). The investigative interview included 11 questions and lasted an average of 9 min 46 s.

Six of the questions were selected for analysis because they were designed to discriminate truth tellers from liars. We categorized the questions into three question types, based on the rationale that they were different types of questions that should have elicited different cognitive and emotional reactions: (1) Open-ended questions were requests to the interviewee to tell their side of the story (e.g., "Describe everything you did in the room"). (2) Direct questions were closed-ended questions that directly asked whether or not the interviewee was lying or had stolen something (e.g., "Did you steal the check?"). (3) Indicator questions were questions often used by law enforcement professionals that were designed to increase cognitive and emotional load on the interviewees and to which truth tellers and liars should respond differentially (e.g., "What should happen to someone who steals money and is caught?").

Stakes and Manipulation Checks

Participants were told they will earn a minimum of $20 and bonuses of $0 to $80 depending upon their assigned condition and the determinations of the interviewers. That is, the participants were informed that the interviewers would be making a determination of their role in stealing the check, and if judged as honest, the participants would receive additional money and would be allowed to leave early; but if they were judged as dishonest, they would receive no additional money and would have to stay an additional hour completing a long questionnaire.

As a manipulation check, participants rated the severity of the consequences if they were judged to be lying in the experiment, using a scale from 1, no consequence, even slightly pleasurable, to 10, maximum consequence, even slightly painful. The rating was obtained after the instructions and stakes were explained to the participants and after the participants were assigned their condition and reminded of their tasks and stakes. The mean rating was above the mid-point, mean = 5.68, SD = 2.24, suggesting that the stakes were perceived to be of moderate severity. There were no ethnicity or condition differences.

We also computed a pre-post (2) × emotion (12) × veracity condition (2) mixed ANOVA on the self-reported emotions. The three-way interaction was significant, $F(11, 2046) = 4.96$, $p = .000$, $\eta_p^2 = .03$, as was the pre-post × veracity condition interaction, $F(1, 186) = 18.49$, $p = .000$, $\eta_p^2 = .09$. To decompose the three-way, we computed single-*df*, pre-post × veracity condition simple interaction contrasts, separately for each of the emotions. Significant effects were found for seven emotions $.02 < \eta_p^2 < .19$. Participants who stole the check and lied about it reported significant increases in guilt, fear, embarrassment, worry, and nervousness compared to participants who told the truth. Participants who did not steal the check and told the truth reported significantly less excitement and interest than participants who lied. Thus, the conditions elicited emotions differentially in the participants.

Analysis of Facial Expressive Behavior

Interviews were video-recorded with a head and shoulder shot of the participants. Facial expressive behavior was coded using an automated facial expression analysis program known as FaceReader (Noldus Information Technology 2013). Initial validation tests were conducted by comparing FaceReader output against the intended emotions portrayed in the Radboud Faces Database (Langner et al. 2010). Overall accuracy of the FaceReader software to the emotions portrayed in the Radboud Faces Database was 90% (Bijlstra and Dotsch 2011). A number of subsequent studies have successfully used FaceReader to analyze facial behavior and classify the behavior into emotion categories (Chentsova-Dutton and Tsai 2010; Harley et al. 2012; Terzis et al. 2012). Facial behavior is classified into one of eight categories for each video frame analyzed. The eight categories are anger, disgust, fear, happiness, sadness, surprise, neutral, and unknown; the unknown and neutral categories were not used in the analyses below. For each category, a probability score for the existence of that facial emotion is calculated based on the full-face prototypical expression on which FaceReader was trained. The frequency of each expression category that occurred within each interview question period was then calculated; the interview question period was defined as the time period from the start of the interviewer's question to the beginning of the next question.

Gesture Coding Procedures

Three types of gestures were selected for coding—head nods, headshakes, and shrugs. These gestures were chosen because coders and investigators could easily see them, and they were likely not to be culture-specific (as opposed to culture-specific emblematic gestures). Shrugs were further divided into two types—shoulder and face shrugs. All coding began by first identifying codable events, which were defined as a continuous excursion and return to the start point of any of the

target movements within the time period of responding to each of the questions during the interviews. For head nods, the target movements were any up and down movement of the head. For headshakes, the target movements were side-to-side head movement around the vertical axis of the head. Face shrugs were defined as a pushing up of the lower lip, a pointing of the corners of the mouth down, with or without an upper lip raise. Shoulder shrugs were defined as up to down movements of the shoulders, on one or both sides.

Interrater reliability was initially established among four raters who coded one third of the entire sample of videos; the average reliability among the coders at this point was .76. Coders were then assigned to code all remaining videos. To assess reliability mid-coding, the two authors divided two thirds of the total sample of videos between them and individually reviewed each of the coded videos. Interrater reliability between the coders and the authors was .90 across all videos checked.

Frequencies of each of the coded gestures were computed separately for each question within each video. For analyses, shoulder and face shrugs were combined into a single "shrugs" category.

Vocal Data Extraction

Audio Record Preparation

All interview videos were logged to denote the onset and offset of each question for both interviewers and participants. Question onset was defined as when the interviewer started a question; offset for each question was defined as onset time of the next question. Within these two time points, participant onset was defined as when the participant began talking in relation to the question asked; participant offset was defined as when the participant completed talking. Within each question, sometimes multiple segments of participant speech was identified (due to prompts by the interviewer, a pause in the conversation, etc.). Extraneous background noises were removed from the records prior to analyses. Video files were converted into audio files, and the following variables were extracted using PRAAT, an open source software:

Pitch

Pitch (frequency) is the rate at which amplitude cycles from positive values to negative values to positive values again in 1 s and is measured in hertz (Hz). Pitch values for the participants were obtained for each segment within each question; mean pitch (Pitch M) was obtained by computing an average across all segments for each question, weighted by segment duration. To obtain indices of variability in pitch, we also computed the standard deviation (SD), minimum (Min), maximum (Max), and range of pitch values in a similar fashion. Analyses involving these variability indices produced similar results; below, we report results using pitch range.

Intensity

Intensity is the difference between the voltage levels of a recorded sound and that of background noise recorded with the sound and is measured in decibels. Intensity values for the participants were obtained for each segment within each question; mean intensity (Intensity M) was obtained by computing a weighted average across all segments for each question. We also obtained the same indices of the variability in intensity in the participant's voices as above; below, we report results using intensity range.

Duration

Duration was defined as the amount of time the participant spent talking and was computed from the time logs generated.

Voice data were also extracted for unfilled pauses, response latency, and speech and articulation rates. These variables, however, were not used in the analyses below.

Coding for Interviewer Contamination

In order to eliminate potential confounds related to the process or quality of the interviews, we coded the interviews in two ways. First, we identified interviews when it was apparent that a participant did not understand the relevant question being posed. Examples included a participant asking the interviewer to repeat the question multiple times or a participant providing a response that clearly did not answer the question. Second, we identified instances when the interviewer impeded or negatively influenced the interview process, potentially causing the participant to provide inaccurate information (and thus influence the produced NVB). Examples were when the interviewer misstated or rearranged the words of the question so that it altered its original meaning; interrupted a participant when he or she was responding; interjected words during a participant's response such as "keep going," "go on;" or volunteered words to help a participant complete a response. Specific questions for which interviewer contamination occurred were identified (coded yes or no).

Two coders, who had several decades of law enforcement experience and extensive experience in analyzing word usage in real-life investigative settings, independently coded transcripts from 30 cases. Both were blind to the condition assignment of all cases and experimental hypotheses. Reliabilities were high and acceptable for both participant did not understand and interviewer contamination codes ($r = .97$ and .83, respectively). One coder then coded the remainder of the cases.

Results

Preliminary Analyses

Assessing clusters of NVB raises questions about their intercorrelations. Table 1 reports the intercorrelations among the variables summed or averaged across all questions. Many pairs of variables were intercorrelated (e.g., anger and disgust, fear and happiness, sadness and surprise, headshakes and shrugs, pitch and pitch range, intensity and intensity range), but many were not. These findings were consistent with the notion that the various channels of NVB produced during the interviews are sometimes related to each other and sometimes not, as suggested in the "Introduction" section.

To test for the existence of an underlying factor structure, we computed principal component analyses on the means of all variables across all questions (to eliminate question effects). Kaiser criterion indicated the existence of six factors that accounted for 66.45% of the variance. We identified scales with variables with factor loadings ≥ .30 and computed Cronbach's alphas; with the exception of the first scale, however, all were low ($\alpha = .83$, .64, .34, .08, .29, .28). The scree plot did not indicate any discernible number of factors. Analyses forcing three-, four-, and five-factor solutions also did not produce interpretable structures. We concluded that a reliable factor structure was not present and proceeded with the remaining analyses utilizing the variables separately.

Hypothesis 1: Differences in NVB as a Function of Question Type

We summed the six facial emotions and three gesture variables, and computed the means of pitch mean, pitch range, intensity mean, and intensity range within each of the three question types (open-ended, direct, and indicator), and computed a question type (3) × ethnicity (3) × veracity condition (2) × gender (2) mixed MANOVA, using the 14 NVB as dependent variables, and filtering all data for any interview contamination. As predicted, the main effect of question type was significant, $F(28, 320) = 13.63$, $p < .000$, $\eta_p^2 = .544$. We decomposed this effect by a series of univariate F tests on each of the 14 NVB measures, followed by orthogonal Helmert contrasts comparing direct questions to the combination of open-ended and indicator questions, and then comparing open-ended and indicator questions (Table 2). As predicted, direct questions produced significantly less happiness, surprise, head nods, headshakes, shrugs, pitch range, intensity, intensity range, and response durations. Additionally, indicator questions produced more happiness, head nods, headshakes, and shrugs, but less pitch range, intensity range, and response durations than open-ended questions. These effects were not moderated by ethnicity or gender, $F(84, 984) = .81$, $p = .892$, $\eta_p^2 = .065$; and $F(28, 320) = .80$, $p = 756$, $\eta_p^2 = .065$, respectively.

Because open-ended questions produced the longest response durations, which confounded the data, we recomputed the scores by dividing by response duration,

Table 1 Intercorrelations among the nonverbal behaviors

		Face					Gesture			Voice				
		DI	FE	HA	SA	SU	HN	HS	SH	PI	PR	IN	IR	DU
Face	AN	.17*	−.02	.01	.15*	−.07	−.09	−.07	−.00	−.14*	−.11	.02	−.03	−.02
	DI		.16*	.14*	.14	−.01	−.08	−.09	−.06	.01	−.03	−.03	−.01	.01
	FE			.22**	.16	.03	.34**	−.08	−.08	.05	.03	.01	.04	−.03
	HA				.02	−.03	.05	.01	−.02	−.04	−.08	.09	.12	−.07
	SA					.17*	.09	.01	.02	−.01	−.04	−.11	−.05	−.02
	SU						.02	.08	.08	.05	.05	−.04	−.02	−.08
Gesture	HN							.28**	.17*	.16*	.06	−.08	−.01	.02
	HS								.40**	.00	−.07	−.03	−.03	−.03
	SH									.13	.04	.01	.00	.00
Voice	PI										.58**	.06	.21**	−.00
	PR											.26**	.33**	.08
	IN												.71**	−.08
	IR													.02

Abbreviations: *AN* total facial anger, *DI* total facial disgust, *FE* total facial fear, *HA* total facial happiness, *SA* total facial sadness, *SU* total facial surprise, *HN* total head nods, *HS* total headshakes, *SH* total shrugs, *PI* mean vocal pitch, *PR* mean vocal pitch range, *IN* mean vocal intensity, *IR* mean vocal intensity range, *DU* mean vocal duration

*$p < .05$ two-tailed; **$p < .01$ two-tailed $N = 211$

Table 2 Significant Helmert contrasts comparing question types for each type of NVB measure

NVB measure	Descriptives			Contrast	$F(1, 86)$	p	η_p^2
	Open-ended	Direct	Indicator				
Facial happiness	.82 (1.70)	.56 (1.32)	1.60 (2.57)	1	15.22	< .001	.150
				2	12.45	.001	.126
Facial surprise	1.44 (2.61)	.27 (.82)	1.05 (2.15)	1	25.50	< .001	.229
				2	1.52	.221	.017
Head nods	9.74 (10.81)	5.75 (6.50)	14.12 (11.83)	1	35.34	< .001	.291
				2	9.40	.003	.099
Headshakes	5.47 (6.77)	4.19 (4.58)	11.85 (12.52)	1	25.11	< .001	.234
				2	26.21	< .001	.234
Shrugs	1.24 (2.18)	.70 (1.15)	4.30 (3.81)	1	62.53	< .001	.421
				2	42.84	< .001	.333
Vocal pitch range	180.62 (92.88)	132.24 (103.96)	148.57 (85.30)	1	11.85	.113	.121
				2	17.24	< .001	.167
Vocal intensity	60.54 (11.30)	59.68 (12.04)	60.12 (11.65)	1	5.69	.019	.062
				2	3.11	.081	.035
Vocal intensity range	29.12 (4.67)	24.09 (6.39)	25.99 (5.48)	1	37.28	< .001	.302
				2	23.95	< .001	.218
Response duration	36.84 (27.28)	9.05 (11.07)	15.18 (9.65)	1	63.79	< .001	.426
				2	52.45	< .001	.379

Contrast 1—direct vs. open-ended and indicator

Contrast 2—open-ended vs. indicator

thereby controlling it. We then computed a question type (3) × ethnicity (3) × veracity condition (2) × gender (2) mixed MANOVA, using the remaining 13 NVB as dependent variables, and filtering all data for any interview contamination. The main effect of question type was still significant, $F(26, 322) = 7.15$, $p < .000$, $\eta_p^2 = .366$. The same orthogonal Helmert contrasts as above, however, produced quite a different picture (Table 3). Direct questions produced more head nods and headshakes, and higher pitch, pitch range, intensity, and intensity range per second than did the other two types of questions. Additionally, indicator questions produced more happiness, surprise, head nods, headshakes, shrugs, pitch, pitch range, intensity, and intensity range than did open-ended questions. These effects were not moderated by gender, $F(26, 322) = .73$, $p = .829$, $\eta_p^2 = .056$.

These interpretations were, however, qualified by a significant ethnicity by question type interaction, $F(78, 990) = 1.36$, $p = .024$, $\eta_p^2 = .097$. We computed the same repeated-measures MANOVA using question type as the factor, separately for each of the four ethnic groups. The same Helmert contrasts as above produced the same findings in the same directions, suggesting a difference in degree, not direction. Cumulatively, these findings provided compelling evidence that different types of questions produce different amounts of NVB.

Hypotheses 2a and 2b: Differentiating Truth Tellers from Liars by Question Type

Question Type Analyses To handle multicollinearity among the NVB, we computed logistic regressions using veracity condition as the dependent variable and the 14 NVB as covariates, using sums (for the six facial emotions and three gesture variables) or means (for the five voice variables) for each variable across the questions within each question type, separately for open-ended, direct, and indicator questions. We utilized backward conditional entry, reckoning that the behaviors were the pool from which any predictive behavior may occur. The remaining behaviors would therefore reflect the cluster that could differentiate truth tellers from liars. For all analyses, we filtered the data for which there was no evidence of interview contamination. We selected for each analysis a final model that was statistically significant and associated with the highest overall classification accuracy rates.

As predicted, the final models were significant for open-ended and indicator questions and accounted for 67.9 and 65.4% overall classification accuracy rates, respectively (the classification rates for liars only are also presented; Table 4). In response to open-ended questions, liars showed more fear and sadness, had lower voice pitches, greater range of vocal intensity, and shorter response durations than did truth tellers. In response to indicator questions, liars displayed more anger,

Table 3 Significant Helmert contrasts comparing question types for each type of NVB measure corrected for response duration

NVB measure	Descriptives			Contrast	$F(1, 86)$	p	η_p^2
	Open-ended	Direct	Indicator				
Facial happiness	.03 (.08)	.14 (.46)	.16 (.33)	1	.90	.345	.010
				2	9.52	.003	.100
Facial surprise	.06 (.11)	.07 (.28)	.11 (.26)	1	.03	.867	.000
				2	4.61	.035	.051
Head nods	.37 (.48)	1.58 (2.47)	1.49 (1.68)	1	9.77	.002	.102
				2	32.03	< .001	.271
Headshakes	.24 (.41)	1.26 (2.20)	1.04 (1.34)	1	4.95	.029	.054
				2	38.08	< .001	.307
Shrugs	.05 (.12)	.21 (.46)	.39 (.44)	1	.10	.753	.001
				2	44.61	< .001	.342
Vocal pitch	6.25 (4.38	44.20 (37.88)	15.15 (9.80)	1	74.69	< .001	.465
				2	69.47	< .001	.447
Vocal pitch range	6.87 (5.16)	31.10 (43.11)	12.85 (10.10)	1	29.45	< .001	.255
				2	43.46	< .001	.336
Vocal intensity	2.64 (2.24)	16.92 (13.47)	6.15 (4.80)	1	84.14	< .001	.495
				2	51.18	< .001	.373
Vocal intensity range	1.22 (.93)	7.71 (5.43)	2.52 (1.70)	1	68.15	< .001	.442
				2	46.24	< .001	.350

Contrast 1—direct vs. open-ended and indicator

Contrast 2—open-ended vs. indicator

disgust, fear, sadness, and surprise, smiled less, and had lower voice pitch.

Analyses to Guard against Type I Error To mitigate the risk of type I error with 14 covariates, we recomputed the above analyses on four randomly selected samples of approximately 66% of the data set, using the same criteria as reported above. These analyses were computed on open-ended and indicator questions only, as direct questions did not produce a significant result. For open-ended questions, all four analyses produced significant results, $\chi^2(5,68) = 12.61$, $p = .027$; $\chi^2(5,72) = 19.85$, $p = .001$; $\chi^2(5,70) = 21.81$, $p = .001$; and $\chi^2(5,64) = 15.37$, $p = .009$, with overall classification rates ranging from 68.8 to 71.4%. All variables in the final equation reported in Table 4 were produced in at least three of the four

Table 4 Results of logistic regressions according to question types

Question type	Final model	Overall classification	Lie classification	Variables in final equation
Open-ended	$\chi^2(5121) = 26.07$, $p < .001$	67.9%	74.1%	Fear Sadness Pitch (−) Intensity Range Duration (−)
Direct	*Ns*			
Indicator	$\chi^2(7121) = 18.61$, $p = .010$	65.4%	70.7%	Anger Disgust Fear Happiness (−) Sadness Surprise Pitch (−)

analyses; intensity range survived in all four analyses. For indicator questions, all four analyses produced statistically significant results, $\chi^2(10,72) = 20.13$, $p = .028$; $\chi^2(10,69) = 19.57$, $p = .034$; $\chi^2(8,71) = 15.96$, $p = .026$; and $\chi^2(6,64) = 12.64$, $p = .049$, with overall classification rates ranging from 64.8 to 75.0%. Disgust, sadness, and pitch survived all four analyses; anger, fear, happiness, and surprise survived at least three of the four analyses.

To further assess the risk of type I error in log regressions with 14 covariates, we created a dataset with totally random data inserted for the 14 NVB variables for each of the three question types and then recomputed the log regression analyses. For open-ended and direct questions, no model was statistically significant. For indicator questions, the analyses did produce a significant model, $\chi^2(6121) = 13.26$, $p = .034$, but the overall classification rate was lower (59.7%). Thus, random data produced no significant results for two of the question types and a lower classification rate for one.

Question-Specific Analyses Averaging across specific questions may have reduced the ability of the NVB to differentiate truth tellers from liars, because NVB is reflective of transient mental states that should be different for each question. Thus, we reanalyzed the data separately for each question using logistic regressions with the same criteria. Of the six questions, the analyses produced five significant models, with overall classification accuracy rates ranging from 62.6 to 72.5% (Table 5). The final cluster of variables in each model represented a combination of facial emotions, gestures, and/or voice variables, speaking to the diverse nature of the predictive nonverbal behavior. The only question to which the nonverbal behavior did not differentiate truth tellers from liars was the direct question asking whether or not the interviewee took the check (question 10).

Post Hoc Analyses

The ethnicity main effect in the overall MANOVA reported earlier was significant, $F(39, 228) = 1.81$, $p < .001$, $\eta_p^2 = .236$. Separate tests of each NVB indicated that there were significant ethnic group differences on pitch range and vocal intensity, $F(3, 86) = 8.37$, $p < .001$, $\eta_p^2 = .226$ and $F(3, 86) = 5.21$, $p = .002$, $\eta_p^2 = .154$, respectively. We followed each using Scheffe post hoc tests. Middle Easterners had significantly higher vocal intensities and greater pitch range than did Chinese or European Americans; Hispanics also had higher vocal intensities than Chinese and European Americans (Table 6). To our knowledge, these findings are new to the literature.

The gender main effect in the overall MANOVA was not significant, $F(13, 74) = .45$, $p = .946$, $\eta_p^2 = .073$.

Discussion

The findings provided broad support for the hypotheses. As predicted, with few exceptions, open-ended questions produced more NVB than did other types of questions. But when response duration was controlled for, direct questions actually produced the most NVB per second, followed by indicator questions. Also as predicted, clusters of NVB differentiated truth tellers from liars, and specific clusters were moderated by question. Accuracy classification rates were well above chance and above deception detection rates by observers (54%; see Bond and DePaulo 2006) and random data.

These findings were not generated without limitations. Although the immigrant participants were either first or second generation, all interviews were conducted in English. It was possible that the non-findings concerning ethnic differences on veracity occurred because the participants used English and its use diluted the possibility of finding ethnic differences. Literature suggesting code frame switching (Hong et al. 2000) among bilinguals is supportive of such a possibility. Regardless, we included these groups because the GEQ data indicated they were culturally different than the European Americans, and their inclusion was ecologically valid as there are many non-native English-speaking individuals who engage with the criminal justice system in English.

Another limitation concerned the exact questions used. While these questions had ecological validity and were grounded in the literature, responses were inextricably tied to the questions asked; thus, the findings were limited to those questions. If different questions were posed, different responses would have been given, producing different findings. A related limitation concerns the number of questions analyzed. Increasing the number of questions (and NVB) tested increases type I error. We attempted to mitigate this concern by cross-validating the findings using randomly selected subsamples of data and by analyses using random data. The main findings were largely reproduced, somewhat reducing the concern for type I error. Still, replication of the findings is necessary, and readers are cautioned to interpret the findings with these caveats.

A third limitation had to do with the differences in sample sizes across the ethnicities and especially the smaller size for Middle Eastern participants. This was particularly true when data were filtered for interviewer contamination. Differences in the sample sizes made statistical comparisons among the ethnicities difficult. Although some ethnicity differences were found in mean levels of two variables, which were new to the literature, these findings ought to be followed in the future.

That Differences in NVB were produced as a function of different types of questions has important implications for their use in investigative interviews and the criminal justice process. Open-ended questions clearly provide investigators with the

Table 5 Results of logistic regressions according to specific interview questions

Question #	Question type	Final model	Overall classification	Lie classification	Variables in final equation
4	Open-ended	$\chi^2(12,121) = 24.07$, $p = .020$	69.1%	71.2%	Anger (−)
					Disgust
					Fear
					Happiness (−)
					Sadness
					Surprise
					Head nods (−)
					Pitch (−)
					Pitch range
					Intensity (−)
					Intensity range
					Duration (−)
6	Open-ended	$\chi^2(13,121) = 24.48$, $p = .027$	72.5%	70.0%	Anger (−)
					Disgust
					Fear
					Happiness (−)
					Surprise
					Head nods (−)
					Headshakes
					Shrugs
					Pitch (−)
					Pitch range
					Intensity (−)
					Intensity range
					Duration (−)
10	Direct	*Ns*			
11	Indicator	$\chi^2(7121) = 19.04$, $p = .008$	66.0%	69.1%	Anger
					Disgust
					Happiness (−)
					Sadness
					Head nods (−)
					Pitch (−)
					Duration
12	Indicator	$\chi^2(8121) = 18.08$, $p = .021$	70.1%	81.0%	Anger
					Disgust
					Happiness (−)
					Surprise
					Head nods (−)
					Headshakes
					Pitch range (−)
					Intensity
13	Indicator	$\chi^2(6121) = 14.34$, $p = .026$	62.6%	67.2%	Anger
					Disgust
					Sadness
					Head nods (−)
					Headshakes
					Pitch (−)

Table 6 Significant ethnicity main effects, descriptives (mean and SD), and results of Scheffe post hoc comparisons

Variable	Chinese	European American	Hispanic	Middle Eastern	Scheffe
Pitch range	135.67 (71.89)	115.89 (69.69)	181.15 (77.84)	221.34 (55.48)	Middle Easterners > Chinese = European Americans
Vocal intensity	55.56 (8.25)	53.01 (7.92)	72.38 (2.31)	72.00 (2.58)	Middle Easterners = Hispanics > Chinese = European Americans

ability to observe the greatest amount and range of NVB for clues of veracity or deception, especially when considered in conjunction with the verbal statements made. But the analyses also showed that other questions, even direct, closed-ended questions that require simple bipolar responses, are pregnant with cognition and emotion, and on a per second basis produced greater amounts of NVB than did open-ended questions, contrary to expectation. Practically, these findings suggest that investigators pay close attention to NVB even when asking questions that require brief verbal responses, as these may actually be associated with much cognition and emotion. Empirically, these findings suggest the development of taxonomies of question types in the future, and for the continued explication of differential patterns of responses to those taxonomies.

No single NVB differentiated truths from lies across all questions. In fact, when data were averaged across questions, the pattern of NVB that differentiated truths from lies was slightly different than the clusters that emerged across the analysis of individual questions (cf, compare the findings reported in Tables 4 and 5). Still, there were consistencies across the analyses, and identifying NVB that consistently differentiated truths from lies across individual questions may be a more powerful and accurate way of assessing which clusters aided in evaluating truth from deception. Examining consistencies across individual questions may be advantageous because averaging across questions can artificially eliminate, or produce, differences. Conducting multivariate (as opposed to univariate) analyses also accounted for the intercorrelations among the predictors. Across the two open-ended questions, liars showed less anger and happiness and more disgust, fear, and surprise. Liars also had fewer head nods; lower voice pitch, intensity, and duration; and greater pitch range and intensity range. Across the three indicator questions, liars produced more facial anger and disgust and less head nods.

These clusters reflected diverse cognitive embodiments and emotional expressions. For instance, liars produced lower pitch when responding to open-ended questions may have been related to efforts to control their heightened levels of emotional arousal, which would have been betrayed by greater pitch range (see discussions concerning emotional control by Hurley and Frank 2011; Vrij 2008). Their fewer head nods may have been associated with less positive affirmations of their statements and/or fewer illustrators of their speech (which also would be consistent with attempts at greater control). The greater anger and disgust, higher pitch, and fewer head nods by liars in response to indicator questions likely reflected greater overall emotionality to such questions compared to than truth tellers, which is indicative of indicator questions.

While much previous research has tested whether single NVB can differentiate truths from lies, examining clusters of NVB is relatively new to the field, and if these findings can be replicated, they would have practical and empirical implications. Practically, they suggest that investigators should be aware of differential patterns of nonverbal responses vis-à-vis different types of questions during investigative interviews. Interviewers could develop detailed knowledge of different types of questions and their associated responses and develop strategies and techniques to prepare for and execute interviews more effectively. Interviewers can also lead and follow the contents and flow of an interview more carefully and have room for unexpected behavioral reactions with less confusion.

These findings also suggest that interviewers need operational knowledge and observational skills related to multiple behavioral channels. Practitioners discriminate lies from truths in real time and within a limited window of opportunity, and for this reason can approach a broader level of observation of interviewees than applying a single sign of deception detection. Practitioners may combine their own database developed from case studies with empirically based scientific findings such as these (and others) to develop and further refine a systematic, evidence-based approach to evaluating truthfulness and detecting deception.

Empirically, these findings suggest that researchers pay greater attention to clusters of NVB that reflect the diversity of the types of cognitive embodiments and emotional expressions that can occur during investigative interviews. The pattern of intercorrelations reported above and the lack of an underlying factor structure suggest that it may be difficult to pinpoint a smaller set of NVB for future testing. Yet, further research can aim to identify smaller sets, continuing the very broad sampling of behaviors we observed in this study. These findings also strongly suggest that researchers pay close

attention in the future to how the type of question asked can moderate the NVB produced and how NVB closely interacts with verbal statements.

The post hoc analyses produced interesting findings that deserve comment. Hispanics and Middle Easterners talked more loudly and with a greater pitch range than did European Americans and Chinese; the Middle Easterners had greater pitch range than did the Chinese or European Americans. These effects may be associated with overall communicative styles of the groups and likely lend themselves to interpersonal and intergroup perceptions and stereotypes. To be sure, these effects did not interact with veracity condition and thus were not signals of deception. But they are typically considered as signs of deception by many (The Global Deception Research Team 2006). Thus, these ethnocultural communication style differences may lead others to believe that speakers are less credible than what one is accustomed to. This may be an important lesson for investigators. Nonverbal differences may also lead to other interesting biases in person and group perceptions, prejudices, and stereotypes, an important line of potential inquiry in the future, with important ramifications for the criminal justice system and investigation processes.

Funding This work was funded in part by the High-Value Detainee Interrogation Group contract J-FBI-12-197 awarded to Humintell LLC. Statements of fact, opinion, and analysis in the paper are those of the authors and do not reflect the official policy or position of the FBI or the US Government.

Compliance with Ethical Standards

Conflict of Interest Both authors are employees of Humintell, to whom the grant was awarded to support this project.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants in the study.

References

Anolli L, Ciceri R (1997) The voice of deception: vocal strategies of naive and able liars. J Nonverbal Behav 21(4):259–284. https://doi.org/10.1023/A:1024916214403

Baumeister RF, Masicampo E (2010) Conscious thought is for facilitating social and cultural interactions: how mental simulations serve the animal–culture interface. Psychol Rev 117(3):945–971. https://doi.org/10.1037/a0019393

Bijlstra G, Dotsch R (2011) FaceReader 4 emotion classification performance in images from the Radboud Faces Database. Retrieved from http://www.gijsbijlstra.nl/ and http://ron.dotsch.org/

Bond CF, DePaulo BM (2006) Accuracy of deception judgments. Personal Soc Psychol Rev 10(3):214–234. https://doi.org/10.1207/s15327957pspr1003_2

Cartmill EA, Goldin-Meadow S (2016) Gesture. In: Matsumoto D, Hwang HC, Frank MG (eds) APA Handbook of nonverbal communication (pp. TBD). American Psychological Association, Washington, DC. https://doi.org/10.1037/14669-012

Chentsova-Dutton YE, Tsai JL (2010) Self-focused attention and emotional reactivity: the role of culture. J Pers Soc Psychol 98(3):507–519. https://doi.org/10.1037/a0018534

Christie R (1970) Scale construction. In: Christie R, Geis FL (eds) Studies in machiavellianism. Academic Press, New York, pp 10–34. https://doi.org/10.1016/B978-0-12-174450-2.50007-5

Davis M, Markus KA, Walters SB, Vorus N, Connors B (2005) Behavioral cues to deception vs. topic incriminating potential in criminal confessions. Law Hum Behav 29(6):683–704. https://doi.org/10.1007/s10979-005-7370-z

Deeb H, Vrij A, Hope L, Mann S, Granhag PA, Lancaster GL (2017) Suspects' consistency in statements concerning two events when different question formats are used. J Investig Psychol Offender Profiling 14(1):74–87. https://doi.org/10.1002/jip.1464

DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H (2003) Cues to deception. Psychol Bull 129(1):74–118. https://doi.org/10.1037/0033-2909.129.1.74

Ekman P (1985) Telling lies: clues to deceit in the marketplace, politics, and marriage, 1st edn. Norton, New York

Ekman P, Friesen WV, O'Sullivan M (1988) Smiles when lying. J Pers Soc Psychol 54(3):414–420. https://doi.org/10.1037/0022-3514.54.3.414

Ekman P, O'Sullivan M, Friesen WV, Scherer KR (1991) Invited article: face, voice, and body in detecting deceit. J Nonverbal Behav 15(2):125–135. https://doi.org/10.1007/BF00998267

Frank MG (2009) Thoughts, feelings, and deception. In: Harrington B (ed) Deception: methods, motives, context and consequences. Stanford University Press, Palo Alto, pp 55–73

Harley JM, Bouchet F, Azevedo R (2012) Measuring learner's co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor. In: Cerri SA, Clancey WJ, Papadourakis G, Panourgia K-K (eds) Intelligent tutoring systems: proceedings of the 11th international conference, ITS 2012, vol 7315. Springer, Crete, pp 40–45. https://doi.org/10.1007/978-3-642-30950-2_5

Hartwig M, Granhag PA, Stromwall LA, Vrij A (2005) Detecting deception via strategic disclosure of evidence. Law Hum Behav 29(4):469–484. https://doi.org/10.1007/s10979-005-5521-x

Hartwig M, Granhag PA, Stromwall LA, Kronkvist O (2006) Strategic use of evidence during police interviews: when training to detect deception works. Law Hum Behav 30(5):603–619. https://doi.org/10.1007/s10979-006-9053-9

Hirschberg J (2002) Communication and prosody: functional aspects of prosody. Speech Comm 36(1-2):31–43. https://doi.org/10.1016/S0167-6393(01)00024-3

Hocking JE, Leathers DG (1980) Nonverbal indicators of deception: a new theoretical perspective. Commun Monogr 47(2):119–131. https://doi.org/10.1080/03637758009376025

Hong Y-Y, Morris M, Chiu C-Y, Benet-Martinez V (2000) Multicultural minds: a dynamic constructivist approach to culture and cognition. Am Psychol 55(7):709–720. https://doi.org/10.1037/0003-066X.55.7.709

Hurley CM, Frank MG (2011) Executing facial control during deception situations. J Nonverbal Behav 35(2):119–131. https://doi.org/10.1007/s10919-010-0102-1

Hwang HC, Matsumoto D (2016) Facial expressions. In: Matsumoto D, Hwang HC, Frank MG (eds) APA handbook of nonverbal communication. American Psychological Association, Washington, DC, pp 257–287. https://doi.org/10.1037/14669-010

Hwang HC, Matsumoto D, Sandoval VA (2016) Linguistic cues of deception across multiple language groups in a mock crime context. J Investig Psychol Offender Profiling 13(1):56–69. https://doi.org/10.1002/jip.1442

Johnson MK (1988) Reality monitoring: an experimental phenomenological approach. J Exp Psychol Gen 117(4):390–394. https://doi.org/10.1037/0096-3445.117.4.390

Johnson MK, Raye CL (1981) Reality monitoring. Psychol Rev 88(1):67–85. https://doi.org/10.1037/0033-295X.88.1.67

Klaver JR, Lee Z, Hart SD (2007) Psychopathy and nonverbal indicators of deception in offenders. Law Hum Behav 31(4):337–351. https://doi.org/10.1007/s10979-006-9063-7

Lambie JA, Marcel AJ (2002) Consciousness and the varieties of emotion experience: a theoretical framework. Psychol Rev 109(2):219–259. https://doi.org/10.1037/0033-295X.109.2.219

Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A (2010) Presentation and validation of the Radboud Face Database. Cognit Emot 24(8):1377–1388. https://doi.org/10.1080/02699930903485076

Matsumoto D (1993) Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. Motiv Emot 17(2):107–123. https://doi.org/10.1007/BF00995188

Matsumoto D, Hwang HC (2015) Differences in word usage by truth tellers and liars in written statements and an investigative interview after a mock crime. J Investig Psychol Offender Profiling 12:199–216. First published online 23 July 2014. https://doi.org/10.1002/jip.1423

Matsumoto D, Juang LP (2016) Culture and psychology, 6th edn. Cengage, Belmont

Matsumoto D, Frank MG, Hwang HS (2013) Nonverbal communication: science and applications. SAGE Publications, Thousand Oaks

Matsumoto D, Hwang HC, Sandoval VA (2015a) Cross-language applicability of linguistic features associated with veracity and deception. J Police Crim Psychol 30(4):229–241. https://doi.org/10.1007/s11896-014-9155-0

Matsumoto D, Hwang HC, Sandoval VA (2015b) Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario. J Police Crim Psychol 30(1):15–26. https://doi.org/10.1007/s11896-013-9137-7

Matsumoto D, Hwang HC, Frank MG (2016) The body: postures, gait, proxemics, and haptics. In: Matsumoto D, Hwang HC, Frank MG (eds) APA handbook of nonverbal communication. American Psychological Association, Washington, DC, pp 387–400. https://doi.org/10.1037/14669-015

Mehrabian A (1971) Nonverbal betrayal of feeling. J Exp Res Pers 5:64–73

Murphy ST, Zajonc RB (1993) Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. J Pers Soc Psychol 64(5):723–739. https://doi.org/10.1037/0022-3514.64.5.723

Noldus Information Technology (2013) FaceReader 5 (Version 5): Noldus Information Technology. Retrieved from http://www.noldus.com/facereader/facereader-5-automatic-facial-expression-analysis-and-emotion-detection

Reynolds E, Rendle-Short J (2011) Cues to deception in context: response latency/gaps in denials and blame shifting. Br J Soc Psychol 50(3):431–449. https://doi.org/10.1348/014466610X520104

Scott S, McGettigan C (2016) The voice: from identity to interactions. In: Matsumoto D, Hwang HC, Frank MG (eds) APA handbook of nonverbal communication. American Psychological Association, Washington, DC, pp 289–306. https://doi.org/10.1037/14669-011

Snyder M (1974) Self-monitoring of expressive behavior. J Pers Soc Psychol 30(4):526–537. https://doi.org/10.1037/h0037039.

Sporer SL, Schwandt B (2006) Paraverbal indicators of deception: a meta-analytic synthesis. Appl Cogn Psychol 20(4):421–446. https://doi.org/10.1002/acp.1190

Sporer SL, Schwandt B (2007) Moderators of nonverbal indicators of deception: a meta-analytic synthesis. Psychol Public Policy Law 13(1):1–34. https://doi.org/10.1037/1076-8971.13.1.1

Streeter LA, Krauss RM, Geller V, Olson C, Apple W (1977) Pitch changes during attempted deception. J Pers Soc Psychol 35(5):345–350. https://doi.org/10.1037/0022-3514.35.5.345

Terzis V, Moridis CN, Economides AA (2012) The effect of emotional feedback on behavioral intention to use computer based assessment. Comput Educ 59(2):710–721. https://doi.org/10.1016/j.compedu.2012.03.003

The Global Deception Research Team (2006) A world of lies. J Cross-Cult Psychol 37(1):60–74. https://doi.org/10.1177/0022022105282295

Tsai JL, Levenson RW (1997) Cultural influences of emotional responding: Chinese American and European American dating couples during interpersonal conflict. J Cross-Cult Psychol 28(5):600–625. https://doi.org/10.1177/0022022197285006

Tsai JL, Levenson RW, Carstensen LL (2000a) Autonomic, expressive, and subjective responses to emotional films in older and younger Chinese American and European American adults. Psychol Aging 15(4):684–693. https://doi.org/10.1037/0882-7974.15.4.684

Tsai JL, Ying Y-W, Lee PA (2000b) The meaning of "being Chinese" and "being American": variation among Chinese-American young adults. J Cross-Cult Psychol 31(3):302–332. https://doi.org/10.1177/0022022100031003002

Vrij A (2008) Detecting lies and deceit: pitfalls and opportunities. Wiley, Chichester

Vrij A, Edward K, Roberts KP, Bull R (2000) Detecting deceit via analysis of verbal and nonverbal behavior. J Nonverbal Behav 24(4):239–263. https://doi.org/10.1023/A:1006610329284

Vrij A, Mann S, Kristen S, Fisher RP (2007) Cues to deception and ability to detect lies as a function of police interview styles. Law Hum Behav 31(5):499–518. https://doi.org/10.1007/s10979-006-9066-4

Vrij A, Leal S, Mann SA, Granhag PA (2011) A comparison between lying about intentions and past activities: verbal cues and detection accuracy. Appl Cogn Psychol 25(2):212–218. https://doi.org/10.1002/acp.1665