Law Hum Behav (2008) 32:339–351
DOI 10.1007/s10979-007-9110-z

ORIGINAL ARTICLE

Deception Detection Expertise

Gary D. Bond

Published online: 10 August 2007

Abstract A lively debate between Bond and Uysal (2007, *Law and Human Behavior, 31*, 109–115) and O'Sullivan (2007, *Law and Human Behavior, 31*, 117–123) concerns whether there are experts in deception detection. Two experiments sought to (a) identify expert(s) in detection and assess them twice with four tests, and (b) study their detection behavior using eye tracking. Paroled felons produced videotaped statements that were presented to students and law enforcement personnel. Two experts were identified, both female Native American BIA correctional officers. Experts were over 80% accurate in the first assessment, and scored at 90% accuracy in the second assessment. In Signal Detection analyses, experts showed high discrimination, and did not evidence biased responding. They exploited nonverbal cues to make fast, accurate decisions. These highly-accurate individuals can be characterized as experts in deception detection.

Keywords Dynamic eye tracking · Deception detection · Nonverbal cues · Expertise · Novices · Cues

Practitioners in all stages of the judicial process understand that deception occurs frequently in those multiple contexts (Strömwall and Granhag 2004). While conducting research with prisoners in my research programs, prisoners informally noted that police detectives, lawyers, correctional officers, and other prisoners lied to them quite frequently. Thus, the forensic context is a unique setting for studying deception and deception detection on many levels from multiple perspectives. Arguably, one forensic context in which lie detection is most important to the judicial process occurs in the law enforcement interviewing context. This research specifically investigated law enforcement practitioners' expertise in detecting deception from paroled felons.

In deception detection studies, Bond and DePaulo (2006) reported that detection accuracy is close to 54%. Detection rates reflect the average of above-chance truth detection (70–80%) and below-chance lie detection (35–40%) over many studies that used university undergraduates, primarily, as detectors (Park and Levine 2001). In studies conducted by Ekman and colleagues, findings suggest that detectors score at a little above chance in their detection accuracy with some exceptions. Some Secret Service personnel (Ekman and O'Sullivan 1991) and other law enforcement practitioners have been reported to detect deception at above-chance rates (Ekman et al. 1999; O'Sullivan and Ekman 2004; C.F. Bond submitted). Thus, there are reports that some people possess greater accuracy in deception detection (Ekman and O'Sullivan 1991), and some may even be designated as experts or "wizards" of detection due to their extraordinary skills (O'Sullivan and Ekman 2004).

In the Ekman and O'Sullivan (1991) study, of the 34 members of the Forensic Services Division of the Secret Service that participated, over half were described as being more than 70% accurate at detecting statements told by undergraduate nursing students videotaped in a previous study (Ekman and Friesen 1974). Of the detectors who were more than 70% accurate, almost one-third of that

G. D. Bond (✉)
Department of Social Sciences, Winston-Salem State University,
115 Coltrane Hall, 601 W. Martin Luther King Jr. Drive,
Winston-Salem, NC 27110, USA
e-mail: gary.bond@gmail.com; bondga@wssu.edu

G. D. Bond
Department of Psychology, Winston-Salem State University,
115 Coltrane Hall, 601 Martin Luther King, Jr. Drive,
Winston-Salem, NC 27110, USA

sample showed greater than 80% accuracy. In a subsequent study (Ekman et al. 1999), a group of federal officers were 80% accurate in detecting lies about opinions told by male undergraduates (Frank and Ekman 1997).

Recently, Bond and Uysal (2007) and C.F. Bond (submitted) challenged the methods and interpretations of findings that Ekman and colleagues reported from their studies, and O'Sullivan (2007) published a reply to the Bond and Uysal paper. Bond and Uysal suggested that the methods used by Ekman and colleagues to score their professionals' deception detection tests were problematic. Detectors attended workshops on detecting deception, where participants judged 10 truthful and deceptive opinion segments (Frank and Ekman 1997). Participants were given the correct answers to the test, and researchers asked participants to raise their hands and report how many correct items they had. If participants reported achieving 90% accuracy or greater, they were invited to take two further tests (Bond and Uysal 2007). As noted by C.F. Bond (submitted), Ekman confirms in a 1992 book that he and O'Sullivan routinely used the self scoring method on the opinion test (pp. 282–285). Bond and Uysal noted that tests should be "supervised and scored by a third party, ideally an individual who does not know the correct answers to test items" (Bond and Uysal, p. 114). O'Sullivan responded that Bond and Uysal are "…erroneous [when they] suggest…that tests should be scored by individuals who do not know the correct answers" (O'Sullivan 2007, p. 122).

Given the lively scholarly debate, researchers outside the Ekman team should independently conduct studies of expertise to investigate whether some people have discovered methods to detect deception that are not identified in the extensive deception literature. If there are experts—and Ekman and O'Sullivan present evidence that there might be—experts may have abilities that defy measurement using conventional methods employed in past research. Thus, two primary goals in this research included (a) identifying experts in forensically-relevant deception detection assessments; and (b) studying their detection behavior using eye tracking. If expert detectors are members of a law enforcement sample, then researchers should employ forensically-relevant tests to accurately assess their detection skill (Bond and Uysal 2007). Also, the measurement of expert detectors' attention should contribute to our understanding of cues that are diagnostic to experts who detect deception in tasks such as these.

Directions for this Research

This investigation assessed whether deception detection expertise would be shown on four tests given at two different times. If detectors in this research showed overall accuracy at or greater than 80% over time, then they would be classified as experts, and their decision reaction time and eye fixation behaviors should be valuable in providing a first step toward describing the decisional and attentional behavior of expert detectors. In think aloud sessions after eye tracking, experts provided descriptions of cues that they used to successfully detect statement veracity. A two-experiment study examined law enforcement personnel and a comparison group of undergraduates. In Experiment 1, detectors participated in large groups in a detection task, and in Experiment 2, the most accurate detectors were studied a second time on a different day. Experiment 2 provided reaction time measures, eye fixations at time of decision, and think aloud data that revealed cues used in detecting deception.

Experiment 1: Group Detection

Following arguments made by Vrij et al. (2006) that stimuli used in past research to assess detection expertise are not ecologically valid (e. g., undergraduates lying and telling the truth about attitudes, beliefs, and opinions have been presented to law enforcement groups), detectors in Experiments 1 and 2 viewed videotapes of paroled felons who lied and told the truth in four conditions. Felons (a) answered interrogation questions concerning a mock crime, (b) answered questions about their past work history in a job interview, (c) spoke about people who had a positive or negative impact upon their lives, and (d) described videos that they viewed.

Since Ekman and colleagues reported that they identified experts in practitioner samples in their past research, law enforcement and comparison undergraduate groups were included in Experiment 1. From the group detection sessions, people who exhibited accuracy rates greater than 80% (expert) and 37.5% or greater (comparisons) were selected for Experiment 2 to compare decision making accuracy, reaction time, eye fixation at moment of decision, and behavioral cues that they used to make decisions.

Method

Stimuli Creation for Detection

Participants Ten paroled felons participated as message presenters. Four Caucasian (2 males, 2 females), four Hispanic (3 males, 1 female), and two African Americans (1 male, 1 female) participated. Participants ranged in age from 21 to 36 (M = 26.7 years, SD = 5.2). Self-reported convictions were for murder, aggravated stalking, robbery, possession with intent to distribute narcotics, possession

with intent to distribute marijuana, child abuse, aggravated assault resulting in great bodily harm on a peace officer, theft, rape, and burglary. Time served in prison ranged from 1-year and 4 months to 17.5 years in prison (M = 6.1 years, SD = 4.7; Mdn = 5.33 years), and time after release from prison to time of participation ranged from 1 week to 3 years (M = .96 years, SD = 1.2; Mdn = .13 years). Each participant received $40. Experimental sessions lasted 3 h with breaks. Interrogators were male graduate students. Interrogators also acted as job interviewers, and in both conditions, followed a memorized script.

General Procedure Participants reported age, years of education, ethnicity, convictions, sentence lengths, lengths of prison term(s) served, time since latest release from prison, and total number of times incarcerated. Experimenters read instructions to participants before each of the interrogation, interview, person description, and video description sessions. Participants produced statements in the four conditions. Statements were edited to one minute for presentation to detectors. Participants consented to have their edited presentations shown to detectors.

Procedure-interrogation Session: Participants were instructed to go to a Professor's office. The participant was asked to wait in the hall until all people in the Professor's office left to take a break. The participant was to enter the office, take a specific book from the bookshelf, and hide it under the Professor's desk. Participants received further instructions about being interrogated. In counterbalanced orders, the participant was told to tell the complete truth about what he or she did, and was told to make up a complete fabrication about the book removal.

Procedure-interview: Participants filled out a generic employment application and were told that they would be interviewed for a mock position at a vacation resort. In counterbalanced orders, participants were told to tell the truth and completely fabricate their job history to the interviewer.

Procedure-affective displacement: Following DePaulo and Rosenthal (1979), participants were asked to talk about a person that they truly liked in a truthful and in a deceptive way. They also spoke truthfully and deceptively about a person that they disliked. Each participant produced four statements (two truthful, two deceptive) in the affective displacement condition.

Procedure-video description: Participants told the truth and lied (in counterbalanced order) about each of the three videos described in *Stimuli* below. They watched a video, and in the truth condition, were told to speak for 5 min describing everything they could remember about the video they watched. In the lie condition, they were told to take 5 min while describing the video they watched in a completely deceptive way. They were told to retain the theme of the video (talk about people who were using Internet chat; talk about two college students who were in a situation together at school; talk about an adult male who was involved in a situation with an 11 year old girl), but were told to completely fabricate their statements about the videos that they watched in the lie condition. Each participant produced three truthful and three deceptive statements in the video description condition.

Procedure-deceptive and truthful statement verification: After each statement was made, felons were asked whether they had lied or told the truth, and were extensively interviewed about the veracity and details of their messages to ensure that stimuli would be accurately classified as deceptive or truthful. In the interrogation, interview, and video conditions, experimenters verified that the statements produced as deceptive were not truthful when compared with details of the mock crime, interview questionnaires, and video stimuli.

Procedure-participant statement videotaping: Participants were videotaped with a Sony Digital Handycam mounted on a tripod. Participants sat in a chair in front of a white wall partially covered in acoustic soundproofing material. Participants' full bodies were captured on the videos, although on a few videos, participants' feet moved outside of the camera angle.

Stimuli Stimuli for the video description condition consisted of three 5-min videos. Segments were taken from *Stalkers: Against the Law* that depicted a male college student who allegedly stalked a female student, a man who stalked an 11-year-old girl, and Internet stalking (*CBS News* 1997). Segment presentations were counterbalanced across participants.

Group Detection

Participants Detectors were recruited from the Federal Law Enforcement Training Center (FLETC) program at Artesia, NM. Other detectors were recruited from southwestern United States federal and local law enforcement groups. Local law enforcement participants were from sheriff and police departments, and federal officers were from Central Intelligence Agency (CIA), Federal Bureau of Investigation (FBI), Bureau of Indian Affairs (BIA), and Border Patrol (BP). There were 64 males and 48 females in the law enforcement group (112 total), ages 23–52 (average 31.4 years, SD = 6.1). There were 55 Caucasians (49.1%),

30 Hispanics (26.8%), 4 African Americans (3.6%), and 23 Native Americans (20.5%). Law enforcement participants reported an average of 15.0 years of education (*SD* = 1.4) and an average of 2.6 years on their current job (*SD* = 1.7). Participants were offered $10 for group sessions, and were advised that if they were 80% or more accurate in group detection, they could participate in a second experiment for $100.

One hundred twenty-two detectors were undergraduate students at a southwestern university who received partial or extra course credit for participating. There were 49 males and 73 females, ages 18–32 (*M* = 20.3 years, *SD* = 2.9); 56 Caucasians, 50 Hispanics, 9 African Americans, 6 Native Americans, and 1 Asian American, with an average 13.5 years of education (*SD* = 1.3). Undergraduates were told that their identification numbers would be drawn from a hat and those chosen would participate in a second session for $100.

Stimuli The 10 paroled felons produced 14 statements each (2 in the book removal condition, 2 in the job description, 4 in the person description, and 6 in the video description conditions). Counterbalanced videotapes were edited from videos to present to detectors. From the 140 videotaped statements, 16 truthful and 16 lie statements were selected for presentation to detector groups in Experiment 1. Each segment was approximately one minute in duration. Equal numbers of truthful and deceptive statements came from interrogations, interviews, person descriptions, and video descriptions. Slides were inserted before videos which cued detectors to condition. The text on slides, for example, read as follows: "*Participant 3, Job Interview*," or *"Participant 26, Person Description."* Conditions were described in detection instructions, which participants read prior to the detection session.

Procedure Groups from 1 to 14 persons assembled in classrooms at FLETC and at a southwestern university. Detectors made binary judgments (truth or lie). They also rated the relative truthfulness or deceptiveness of statements on a 1–7 scale. After each segment, the video was paused for 30 s, and then the next video was shown. A questionnaire was also given, which included questions about ethnicity, age, years of education, time on current job, agency or office where employed, and job title and duties.

Results

All analyses reported in the following sections were conducted with a significance level set at $p < .05$ with Bonferroni adjustments.

Detection in Conditions and Across Presentations

Accuracy scores between law enforcement and undergraduate student groups were analyzed in each of the four presentation conditions and for overall accuracy across conditions. The range of overall accuracy for law enforcement participants was from 31.25% to 93.75% correct. Undergraduates' overall accuracy ranged from 34.38% to 62.5% correct. Table 1 depicts means, standard deviations, and significant results of the analyses. Intercorrelations among the four scenario tests within veracity conditions are depicted in Table 2.

Response Bias: Signal Detection Analysis

Past studies that have computed discrimination accuracy and response bias measures based upon Signal Detection

Table 1 Group detection accuracy: Experiment 1

Stimulus condition	Group					
	Undergraduate students			Law enforcement		
	Truth M (*SD*)	Lie M (*SD*)	Overall M (*SD*)	Truth M (*SD*)	Lie M (*SD*)	Overall M (*SD*)
Book removal	.54 (.28)	.35 (.23)	.44 (.15)	.51 (.30)	.39 (.28)	.45 (.20)
Job interview	.83** (.17)	.30 (.20)	.57 (.12)	.72 (.17)	.44** (.26)	.58 (.16)
Person description	.67 (.22)	.44 (.28)	.56 (.18)	.65 (.25)	.45 (.28)	.55 (.22)
Video description	.59** (.22)	.45 (.33)	.52 (.21)	.50 (.24)	.53 (.31)	.52 (.22)
Across conditions	.66** (.12)	.39 (.14)	.52 (.07)	.59 (.16)	.45** (.17)	.52 (.14)

Notes: *$p < .05$; **$p < .01$. Job interview lie detection: $F(1, 232) = 21.30$, $p = .0001$, $\eta^2 = .08$. Job interview truth condition: $F(1, 232) = 22.27$, $p = .0001$, $\eta^2 = .09$. Video description truth condition: $F(1, 232) = 8.77$, $p = .003$, $\eta^2 = .04$. Groups differed on lie accuracy, $F(1, 232) = 9.97$, $p = .002$, $\eta^2 = .04$, but both performed significantly below chance in lie detection. Groups differed on truth accuracy, $F(1, 232) = 11.02$, $p = .001$, $\eta^2 = .05$, but both groups performed significantly above chance in truth detection

Table 2 Intercorrelations among stimulus and veracity conditions: Experiment 1

Condition/veracity	Interr. T	Interr. L	Intrvw. T	Intrvw. L	Pers. T	Pers. L	Vid. T	Vid. L
Interr. T	–							
Interr. L	−.16*	–						
Intrvw. T	.08	−.15*	–					
Intrvw. L	−.15*	.36**	−.10	–				
Pers. T	.04	.04	−.02	−.22**	–			
Pers. L	−.43**	.31**	.01	.29**	−.20**	–		
Vid. T	.26**	−.01	.36**	.04	.25**	.04	–	
Vid. L	−.34**	−.07	.04	−.02	−.16*	−.05	−.20**	–

Notes: *$p < .05$, **$p < .01$; $N = 234$. T = truth, L = lie. Interr. = Book removal condition. Intrvw. = Job interview condition. Pers. = Person description condition. Vid. = Video Description condition. 32 presentations, 16 in each veracity condition

Theory suggest that law enforcement officers' training and experience produces a bias to respond "lie" rather than producing improvements in measures of detection performance (Meissner and Kassin 2002). In the detection analyses, law enforcement showed significantly more correct lie judgments than students. This trend suggests that they might have held a liberal bias in responding, rather than greater discrimination. In order to characterize decision making bias in the groups, Signal Detection Theory (SDT; Green and Swets 1966) analyses were conducted. SDT provides a measure of discrimination (d') independent from observer bias (β). If a person was biased toward saying that most messages are deceptive (lie-bias; Bond et al. 2005), the number of hits and false alarms would be high, and β would be lower than if a person tended to believe most messages (truth bias; Zuckerman et al. 1984).

Groups were analyzed on hits (saying "lie" when a deceptive presentation was given) and false alarms (saying "lie" when the truth was presented). Results showed differences between groups in both cases. Results for hits showed $F(1, 232) = 9.97$, $p = .002$, $\eta^2 = .04$; $M_{\text{law enforcement}} = .59$, $SD = .3$, and $M_{\text{undergraduate}} = .28$, $SD = .16$. For false alarms, the test revealed $F(1, 232) = 10.10$, $p = .002$, $\eta^2 = .04$; $M_{\text{law enforcement}} = .50$, $SD = .37$, and $M_{\text{undergraduate}} = .39$, $SD = .24$.

When hits and false alarms were used as estimates to compute criterion and discrimination, results showed that groups were different in criterion-setting, $F(1, 232) = 5.96$, $p = .02$, $\eta^2 = .03$. Law enforcement set a more liberal criterion when making decisions, $M = .29$, $SD = .5$. Students' criterion averaged .43, $SD = .36$. There was no difference, however, in discrimination, $F(1, 232) = .24$, *ns* ($M_{\text{law enforcement}} = .15$, $SD = .8$, $M_{\text{undergraduate}} = .11$, $SD = .43$). Results indicated that the law enforcement group showed a bias to respond "lie" and not greater discrimination of message veracity.

Detection in the four conditions was assessed to determine whether law enforcement would show more liberal criterion-setting behaviors in experience-related areas, particularly for conditions in which felons responded to questions (book removal and job interviews). In the book removal condition, undergraduates set a criterion (β) of .09, $d' = -.31$; and the law enforcement group exhibited $\beta = .01$, $d' = -.27$. Neither β nor d' parameters were different between groups. Undergraduates ($\beta = .95$, $d' = .43$) and law enforcement ($\beta = .59$, $d' = .55$), when contrasted in job interview detection, showed differences in criterion-setting, $F(1, 232) = 25.59$, $p = .0001$, $\eta^2 = .10$; but d' was not different between groups. In person descriptions, undergraduates ($\beta = .43$, $d' = .29$) and law enforcement ($\beta = .43$, $d' = .30$) showed no differences on β and d'. In video descriptions, undergraduates ($\beta = .22$, $d' = .11$) and law enforcement ($\beta = .00$, and $d' = .08$) were different in criterion-setting, $F(1, 232) = 5.56$, $p = .02$, $\eta^2 = .02$; but d' was not different between groups.

Participants Greater than 80% Accuracy

After 112 law enforcement personnel participated, 11 showed overall accuracy scores ranging from 81.25% (26 correct out of 32 total decisions) to 93.75% (30/32; $SD = .04$). Of the 11, the range for lie detection was from 68.75% to 93.75%, ($SD = .07$), and the range for truth detection was from 75% to 100% ($SD = .08$). Only one high scorer showed a liberal responding bias ($\beta = 0$); while criterion-setting behaviors for the other high scorers ranged from a criterion of .81 to 3.09 ($M = 1.27$, $SD = .75$). With the individual removed, averaged group criterion was 1.39 ($SD = .66$). The high scorer with the liberal responding bias showed $d' = .74$, but the other high scorers showed a range from $d' = 1.66$ through 4.14 (mean $d' = 2.27$, $SD = .86$). Undergraduates who scored at 37.5% or greater were assigned a 4-digit ID number, and after 122 students participated, 11 who scored at or above the specified accuracy percentage were chosen from 120 numbers. Scores for the

persons selected to participate in Experiment 2 ranged from 43.75% (14/32 correct decisions) to 62.5% (20/32 correct; *SD* = .07).

Experiment 2: Eye Tracking

Results in Experiment 1 showed that 11 persons scored at 80% or greater in detecting truth and deception from messages produced by paroled felons. Those high-performers were all members of the law enforcement group. If expertise was replicated by some or all of the law enforcement personnel in Experiment 2 (defined as overall detection accuracy at or greater than 80%), then their accuracy, reaction times, eye fixations at point of decision, and think aloud data (cues to deception) would be valuable in characterizing expertise in detecting deception.

Reaction Time, Eye Movements, and Cues

Experts were predicted to quickly integrate behavioral information to come to a veracity conclusion, because experts fixate salient locations quickly (Charness et al. 2001). Less-experienced observers require more sensory information to create an overall representation of a display (Abernathy 1990). It was expected that experts would be significantly faster at making decisions, as measured by reaction time (RT), due to faster identification of the most diagnostic behavioral information with less sensory input (*H1*). It was also important to distinguish the fixation location on the area of the face or body at the precise time when an expert made his or her judgment of "truth" or "lie." *RQ1* examined this question by describing the area fixated upon at the time of decision for each expert participant. Cues that detectors used were reported in think aloud sessions after each stimulus was viewed. *RQ2* examined whether experts reported using more paralinguistic, verbal, nonverbal, or intuitive information to make decisions.

Method

Participants

Detectors who scored at 80% accuracy or greater in the law enforcement group were invited to participate in one 3-h session on a different day. Two female participants' data were excluded due to eye tracking software malfunctions and one male declined to participate in the second session. Of 112 law enforcement participants, 11 persons scored at 80% accuracy or greater in group detection, but only 8 were included in analyses. The final participants were five males and three females, ages 26–30 (*M* = 28.13 years, *SD* = 1.6). There were 3 Caucasians (CIA, Border Patrol, and local law enforcement), 3 Hispanics (one Border Patrol and two local law enforcement), and 2 Native Americans (both BIA correctional officers). Data from three females in the undergraduate group were excluded due to software malfunctions, and replacements were randomly chosen. The final student detectors were four males and four females, ages 19–27 (*M* = 20.2 years, *SD* = 1.3); 3 Caucasians, 2 Hispanics, 1 African American, and 2 Native Americans.

Materials and Apparati

A 42″ (diagonal measurement) plasma video monitor was used for eye tracking experiments. The size of the monitor controlled for simultaneous fixation of the mouth and eye areas (Vatikiotis-Bateson et al. 1998). Eye movements were measured with an Eye-Link II eye tracker. The system was calibrated before each video was shown, where gaze position error was less than .5°, and temporal resolution of the system was approximately equal to 4 ms. After calibration, the Eye-Link II automatically chose which eye showed the most accurate fixation behavior, and eye movements were measured monocularly in that trial with that eye. Two dedicated PCs were used. A host computer controlled the eye tracking equipment, and a display computer presented each message on the plasma monitor. Eye tracking presentations were programmed using SR Research Experiment Builder 1.2©. The program allowed for presentations of ten .avi videos synchronized with .wav sound and with eye recording. Text screens were inserted between videos to provide participants with condition instructions. Ten stimulus presentations that were not used in Experiment 1 were presented in Experiment 2 (2 book interrogations, 2 job interviews, 4 person descriptions, and 2 video descriptions) from each of the 10 paroled felon participants.

Procedure

Participants watched 10 presentations, and before they began viewing, they were asked to report immediately when they thought the person was telling the truth or lying. Participants were not told that they were viewing convicted felons. At the end of each video, participants made relative deceptiveness/truthfulness and confidence ratings. Participants viewed 10 counterbalanced videos, five of which were truthful, and five deceptive statements. Participants viewed each video a second time for an audio taped think-

aloud session. Participants were told to verbally report everything they were thinking as they watched the stimulus (Ericsson and Simon 1980, 1993). The specific instruction given was as follows: "You are going to watch the video a second time. I would like you to talk aloud about everything you are thinking as you watch the video again." If participants stopped talking for a period of 10 s, they were prompted to "please continue talking aloud about everything you are thinking as you watch the video." Participants were not provided a "warm-up" procedure prior to the first talk-aloud session.

Results

All analyses reported in the following sections were conducted with a significance level set at $p < .05$ with Bonferroni adjustments.

Detection in Presentation Conditions

Accuracy scores of law enforcement and undergraduate student groups were analyzed in each of the book removal, job interview, person description, and video description presentation conditions. Groups did not differ in any of the presentation conditions.

Response Bias: Signal Detection Analysis

Detection performance was analyzed with SDT. Detection in each stimulus condition was assessed. Neither β nor d' parameters were different between groups in any condition.

Detection Accuracy

Overall accuracy across presentation conditions was investigated, and results are presented in Table 3. When Experiment 2 was completed, two people showed overall accuracy scores which continued to exceed 80%. Of the two, lie and truth accuracy ranged from 80% to 100%. The two experts (referred to as expert A and B hereafter) were female Native American BIA correctional officers. They were 30 and 28 years old, respectively. Table 4 shows the two experts' accuracy within veracity conditions in Experiments 1 and 2 and across experiments.

Table 3 Detection accuracy across conditions: Experiment 2

Condition	Group					
	Undergraduate students			Law enforcement		
	Truth M (*SD*)	Lie M (*SD*)	Overall M (*SD*)	Truth M (*SD*)	Lie M (*SD*)	Overall M (*SD*)
Across conditions	.63 (.2)	.33* (.15)	.48 (.07)	.55 (.33)	.63* (.27)	.59 (.27)

Notes: *$p < .05$; **$p < .01$. Lie, truth, and overall accuracy rates were not different from chance for the law enforcement group (all *t*s < 1.30, *ns*). Undergraduate lie detection was below chance, $t(7) = -3.33$, $p = .01$; truth detection and overall accuracy was not different from chance. Groups differed in lie detection accuracy, $F(1, 14) = 7.52$, $p = .02$, $\eta^2 = .35$

Table 4 Expert detection: Experiments 1 and 2 and across experiments

Expert	Detection				
	Truth % accuracy	Lie % accuracy	Overall % accuracy	β	d'
Experiment 1					
A	93.75%	93.75%	93.75%	1.53	3.07
B	81.25%	87.50%	84.38%	.89	2.04
Experiment 2					
A	100.00%	80.00%	90.00%	3.09	3.93
B	80.00%	100.00%	90.00%	.84	3.93
Across Experiments					
A	95.23%[a]	90.48%	92.86%	1.67	2.98
B	80.95%	90.48%	85.71%	.88	2.19

[a] Averaged accuracy for Expert A across experiments was 92.86% (39/42 presentations), and Expert B's average across experiments was 85.71% (36/42)

Reaction Time Analysis (Time to Decision): H1

A 1,000 Hz pure tone was sounded at the beginning of each video presentation, and reaction time (RT) was measured from the cessation of the pure tone to the onset at which the detector said "lie," "lying," "truth," or "truthful," as his or her response. RT was measured by extracting audio from the videotape made during each eye tracking session, and importing the audio into a GoldWave© version 5.08 program (GoldWave, Inc., 2004). RT was calculated in seconds, with time less than a second rounded to two decimals (e. g., 52.79 s).

A log10 transform of RT was conducted to compare (a) the law enforcement group with the undergraduate group in decision-making, and (b) in an exploratory analysis, the two experts' RT behaviors with undergraduate participants to investigate differences. When law enforcement was compared to undergraduates in overall RT across veracity conditions, results were $F(1, 14) = 2.67$, $p = .12$, *ns*, $\eta^2 = .16$ ($M_{\text{law enforcement}} = 1.48$, $SD = .30$; $M_{\text{undergraduate}} = 1.67$, $SD = .15$). RT did not differ in lie or truth conditions. The two experts from the law enforcement group were compared to undergraduates in overall RT across veracity conditions. Experts were significantly faster than undergraduates in making decisions in all conditions. Table 5 depicts untransformed RT for experts, law enforcement, and undergraduates.

Eye Tracking Analysis

The EyeLink II Data Viewer was used to analyze eye fixations at moment of decision. *RQ1* examined fixation location on the area of the face or body at the moment when an expert detector made a judgment of "truth" or "lie." The Data Viewer has an Animation Viewer which was accessed for viewing fixation at time of decision. Expert's decisions were investigated by reviewing the playback to determine exact locations on the face and body upon which they looked when making their decisions. Their fixations/decisions appear in Fig. 1. Expert A looked more at face areas and Expert B looked more at arm/torso areas when making decisions.

Think Aloud Cue Coding and Analysis

The think aloud data were transcribed and coded to analyze *RQ2*. One research assistant transcribed think aloud sessions, and a second assistant reviewed and corrected inaccuracies. Two other assistants blind to veracity conditions coded all phrases in the transcripts. Each phrase was coded into (a) paralinguistic cues, (b) message content cues, (c) nonverbal cues, and (d) intuitive cues. Paralinguistic cues followed definitions in research by Kasl and Mahl (1965) and by DePaulo et al. (1982). If detectors mentioned nonfluencies (e. g., stutters, pauses, sentence incompletions); um's, er's, ah's, or pitch, loudness, or rate in their phrases, those instances were coded as paralanguage. Think aloud reports that mentioned aspects of verbal language or content (e. g., inconsistencies, contradictions, plausibility) were coded as message content cues. Nonverbal cues were movements, gaze, facial expressions, posture, eye behaviors, etc. Intuitive cues were reports of feelings, gut reactions, and instinctive feelings (e. g., "I just felt like he was lying"). Together, coders practiced dividing three of the transcripts into phrases. Separately, they coded each phrase in the transcripts. Interrater reliability was $\alpha = .74$ for paralinguistic cues, $\alpha = .80$ for nonverbal cues, $\alpha = .78$ for verbal cues, and $\alpha = .86$ for intuitive cues. Percentages were derived from the number of cues in each category divided by total cues mentioned in the sessions.

An analysis of variance revealed a difference between undergraduates and law enforcement on mentioning paralinguistic cues, $F(1, 14) = 22.04$, $p = .0001$, $\eta^2 = .61$. The proportion of paralinguistic cues mentioned by undergraduates was .12, $SD = .03$; $M_{\text{law enforcement}} = .05$, $SD = .03$. There were no differences on verbal, nonverbal, or intuitive cues. The data were also examined to determine if

Table 5 Experiment 2: decision making reaction times

Veracity condition	Group		
	Experts M (*SD*)	Law enforcement M (*SD*)	Undergraduate students M (*SD*)
Truth	12.27** (1.25)	36.08 (16.47)	52.24 (11.14)
Lie	10.75** (3.52)	36.29 (16.98)	50.09 (12.99)
All conditions	11.51** (1.13)	36.18 (16.51)	51.17 (11.32)

Notes: *$p < .05$; **$p < .01$. Expert $N = 2$; Law Enforcement $N = 8$; Undergraduate Students' $N = 8$. In all conditions, Expert RT was faster than Undergraduate Students, $F(1, 8) = 34.65$, $p = .0001$, $\eta^2 = .81$. In lie and truth conditions, Expert RT was faster, $F(1, 8) = 24.04$, $p = .001$, $\eta^2 = .75$; $F(1, 8) = 43.00$, $p = .0001$, $\eta^2 = .84$, respectively

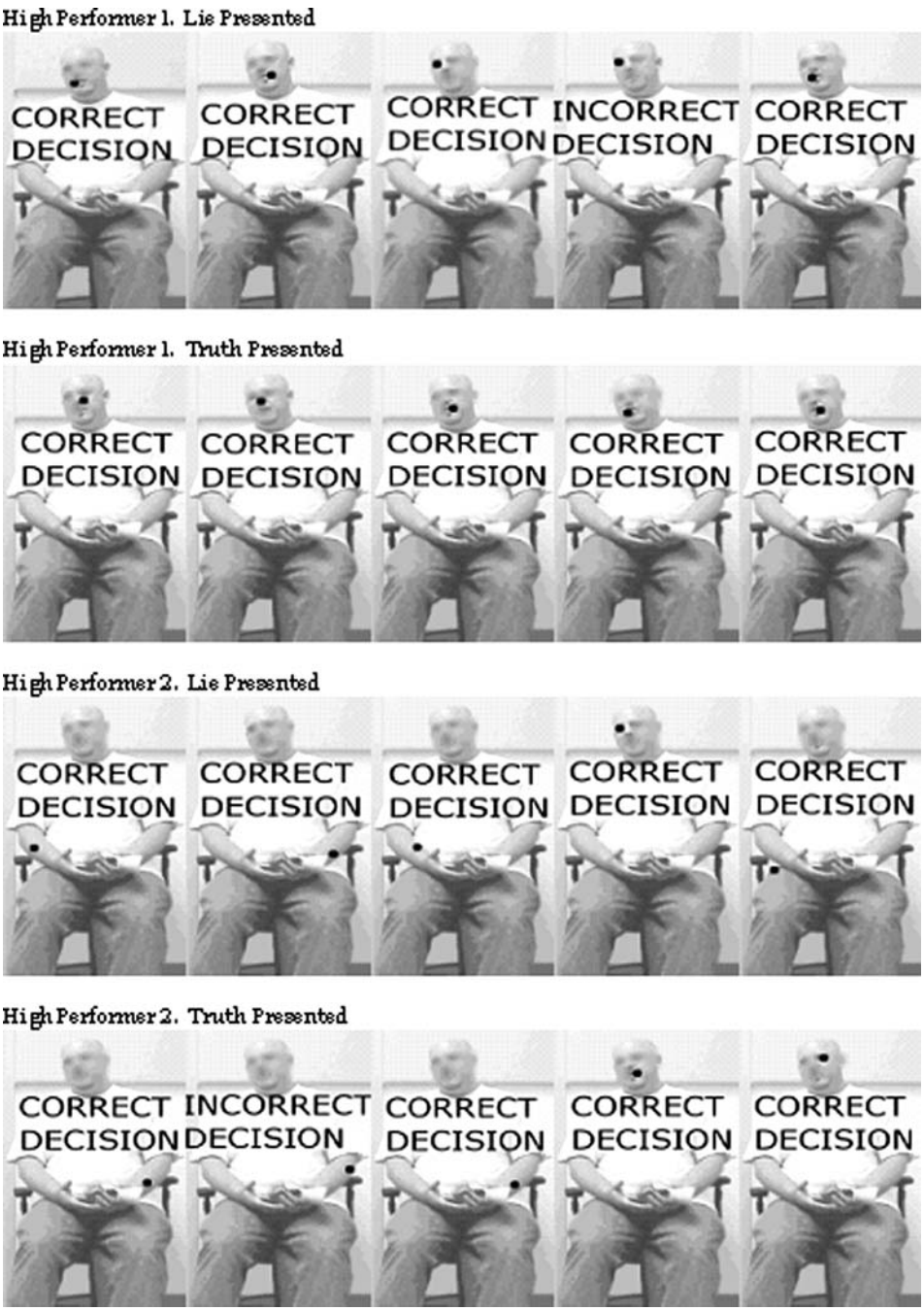

Fig. 1 Experts' fixations at time of decision

differences existed between experts A and B and undergraduate detectors. The reporting of paralinguistic cues was higher for undergraduates than for experts, $F(1, 8) = 11.45$, $p = .01$, $\eta^2 = .59$. Fewer verbal cues were mentioned by experts than in the comparison group, $F(1, 8) = 9.642$, $p = .02$, $\eta^2 = .55$. However, nonverbal cues were mentioned more by experts than undergraduates, $F(1, 8) = 9.10$, $p = .02$, $\eta^2 = .53$. No difference was found on intuitive cues. Undergraduates reported more verbal cues, and experts reported more nonverbal cues.

Discussion

O'Sullivan and Ekman (2004) have reported that a small number of people can be classified as "wizards" in deception detection. The current study was conducted with a sample of screened law enforcement personnel and undergraduate students, and just two people, both from law enforcement, might be characterized as experts in deception detection. The two experts were over 80% accurate in a first assessment which consisted of four different tests,

and scored at 90% accuracy in detection in a second assessment with the same four tests. Although methods, stimuli, and criterion for expertise employed in this research are different from those used by the O'Sullivan and Ekman research group, results of this study indicate that there are a few individuals who detect deception at expert levels.

This research improves upon some limitations in the O'Sullivan and Ekman (2004) study. While one of O'Sullivan and Ekman's methods sometimes involve mailing stimulus videotapes to people for follow-up testing of potential "wizards" (O'Sullivan and Ekman 2004), for example, all detectors in this study were screened and assessed under controlled conditions in the field and in the laboratory. Second, methods that O'Sullivan and Ekman use to identify experts include stimuli presented by undergraduates who lie and tell the truth about feelings, opinions, and a mock theft. Arguably, felons lie and tell the truth differently from students, as suggested in research by Bond and Lee in U. S. prisons (2005). Expert detectors have been identified by Ekman and others from law enforcement backgrounds, and practitioners regularly interact with recidivists and others who have violated society's laws. Vrij and colleagues (2006) suggest that suspects' deceptive language and behaviors in police interviews are different than messages presented by undergraduates (pp. 742–743). With these ideas in mind, felons were used as stimuli in this study for screening expertise.

In future studies, expert detectors may show that they primarily use nonverbal processing of suspect behaviors at a greater rate than intuition to determine the veracity of messages. Experts in this study attended to nonverbal information, and were faster at identifying areas of diagnostic behavioral information with less sensory input than novices. Those data suggest that experts are actively processing behavioral information, rather than relying on intuition to make decisions.

Expert or Statistical Anomaly?

Are the two experts identified in this study "virtuosos" of lie detection (O'Sullivan 2007) or statistical "flukes" (Bond and Uysal 2007)? Bond and Uysal (2007) questioned whether O'Sullivan and Ekman (2004) had identified 29 true wizards out of 12,000 tested, suggesting that participants might have achieved "wizard" performance by chance. In this research, two persons were identified as experts.

First, we need to understand the criterion for expertise, as set by O'Sullivan and Ekman (2004), in order to ascertain whether we have identified experts in this study. Experts in the O'Sullivan and Ekman investigation needed to score at least 90% on their first test (opinions), and be at least 80% accurate on both of two subsequent tests (mock crime or emotions; O'Sullivan 2007). Only 1/3 of the wizards were 80% or more accurate on both subsequent tests, but 2/3 showed a profession-specific error. On two follow-up tests, the latter participants performed better (O'Sullivan, personal communication). In this study, participants were to score at least 80% on four different tests (interrogation, job interview, person description, and video description), and score at least 80% on the four tests a second time. O'Sullivan and Ekman (2004) indicated that there was an extremely low probability (.000025) that their participants had achieved accuracy rates at expert levels by chance alone. Bond and DePaulo (2006) indicated that over many detection studies, detection rates are 54%, and so a research-based model would indicate that the probability of people achieving their rates by chance would be .00016 (O'Sullivan 2007).

Let us consider whether the 2 experts identified out of 112 law enforcement participants in this research might have achieved their high accuracy rates by chance. Suppose each of 112 law enforcement judges takes a 42-item lie/truth test and the probability of each judge being correct on each item is .52 (the accuracy average of law enforcement participants in this research). All law enforcement personnel did not respond to all 42 items (most only responded to the first 32-item test and not to the second 10-item test), so we will use a mean of .526 for a research-based model (the mean accuracy of all participants in this study) for a more stringent test. The probability that the very best detector (out of 112) would achieve 39 or more of the 42 items correct (Expert A's score) is .000002.

Biased Responding

The large group of law enforcement personnel showed a bias to respond "lie," yet their lie detection accuracy as a group was less than 50%. O'Sullivan (2007) reports that some law enforcement groups only score at chance levels, while other "selected groups" of deception interested personnel yield above-average detection rates. This study included local law enforcement and federal law enforcement detectors, and no differences were found between the two groups on accuracy. However, the group of federal law enforcement did yield two highly accurate experts. Experts' criterion-setting was not liberal, so they did not show the bias to respond "lie" that has been found in other studies of law enforcement personnel.

Eye Tracking, RT, Cues Used in Detection

Expert detectors made perceptually-fast, highly-accurate decisions about message veracity. Their time from initial

inspection of a presentation to decision was swift. The experts seemed to hold developed schemas based on past encounters with deceptive suspects or other people, and they were ready to use those schemas to actively detect deception or truth from nonverbal behaviors from the onset of the presentations. Expert A made decisions while fixating face areas (lips, eye, nose, and cheek). Those areas were areas of high visual interest because she had probably used those areas successfully in the past to detect deception. Expert B showed greater idiosyncrasy than Expert A, making her decisions when looking at areas near the torso (right and left arm, top of right leg). However, she also looked at the face and made decisions while looking at lip and eye areas as well. Expert B made accurate decisions while fixating arm torso-leg areas, and so movements of the arms and legs were particularly diagnostic for her. Eye tracking should be used in future studies not only to identify attention at time of decision, but to gather information about the way experts scan nonverbal elements of a suspect's dynamic presentation. Those data were collected in this experiment, but will be reported separately.

Experts were primarily nonverbal detectors. Although Ekman and O'Sullivan (1991) indicate that participants who mention both speech cues and nonverbal cues obtain higher total accuracy scores than those who mention only speech cues or only nonverbal cues, that particular pattern of multiple cue usage was not found in the behavioral coding of experts' detection behaviors in this study.

Limitations

The apparent low-stakes nature of the book removal task might have limited detection performance of law enforcement personnel. As Frank and Ekman (1997) and Ekman (1985) indicate, the importance of stakes (higher are better) in scenarios will set the stage for leakages and other cues to deception that would be low or non-existent in lower-stakes situations. However, accuracy rates of most of the law enforcement personnel in this study are comparable to rates obtained in past studies which included higher-stakes scenarios (e.g., DePaulo and Pfeifer 1986; Ekman et al. 1999). Arguably the book removal method might not be as stress-provoking as when participants commit a more serious mock crime, such as breaking and entering or vandalism (e.g., Kassin and Fong 1999).

One limitation suggested by reviewers was employing graduate students to act as interviewers in the book removal condition. The persons selected for the task were males who were former university athletes, and they were both experienced speakers. They practiced interviews many times before experiments began, and they exhibited the demeanor of "professors" upset about an important textbook being removed from their offices. Interviews were realistic and the paroled felons sometimes became defensive in their behavior and in the language that they used in their responses.

The Experiment 2 tests contrasting two experts with others were exploratory; in order to generalize these findings, larger sample of experts and comparisons are needed. Another limitation to the research is that 10 presenters produced all of messages (32 in the first experiment, 10 in the second experiment). If a detector was correct in "reading" someone on one item, it is possible that the detector would be accurate when shown that same person on a second item.

Finally, some scholars believe that there are inherent problems associated with verbal reports such as think aloud protocols. Participants said everything that they were thinking as they viewed presentations a second time in Experiment 2. As Ericsson and Simon (1980) note, "verbal reports have been suspect as data," and "verbal behavior[s] are frequently dismissed as variants of the discredited process of introspection" (p. 216). Instructions in this research were undirected, and so the method did not induce artificial generation of information regarding cues, where in a directed probe method, participants might try to infer or guess what the experimenter wanted, or might "alter their normal mode of processing in order to be able to give the requested information to the experimenter on subsequent trials" (Ericsson and Simon 1980, p. 222). Participants understood that their task was lie detection, but they provided a range of information regarding presentations. They provided cues they used to detect from presentations in almost all instances.

Future Directions

The number of Native Americans and Hispanics in the study was large, since the research was conducted in the southwestern United States. Experts were young adult Native American females with relatively brief on-the-job experience in correctional settings. Standardized personality and cognitive measures were not administered in this research, and so many commonalities or dissimilarities could not be assessed. Experts' supervisors and coworkers were not interviewed to provide data on the job performance of the two. Their families were not contacted to provide reports which might describe life experiences or other factors which made these two unique people effective in detecting deception from nonverbal behaviors. O'Sullivan and Ekman (2004) report that they are collecting some of the data mentioned above, and those data should be collected in future studies of expertise.

Most importantly, truly-expert deception detectors, when identified, should not just be tested on a few occasions to assess their abilities. As Vrij (2004) points out, a better way to measure expert ability is to employ studies in which people are tested on several occasions. In this way, accuracy and cues that experts mention and use behaviorally might be better assessed. Speeded judgments (Vrij et al. 2004) should also be included as a methodology in assessing expertise. Since expert decisions in this study were fast and highly-accurate, judgments should be obtained from short stimuli presentations.

Stimuli for future detection experiments should be comprised of samples from the population that evidences a criminal history, especially when assessing expertise of participants in forensic contexts. This limitation in past research has resulted in findings that are probably not generalizable to the population of people who actually commit many of the crimes that law enforcement officers investigate. For example, the average suspect in a police interrogation, according to Gudjonsson (2003) is intellectually disadvantaged, has an average IQ of 82 (Hartwig et al. 2005) and is fundamentally different from undergraduate students who have been used in studies to act as "suspects," in order to present stimuli to deception detectors. Thus, it is important to psychology and law professionals to critically evaluate past deception detection research which involves student presentations. This research was conducted with paroled felons' stimuli to contrast professionals' and students' deception detection accuracy with the idea in mind that real-life suspects are fundamentally different from students in the way that they react to questioning.

Acknowledgments The author wishes to thank Drs. Adrienne Lee, Anne Hubbell, Douglas Gillan, Dominic Simon, Timothy Ketelaar, and Daniel Malloy for assistance. Special thanks to Dr. Charles Bond, Jr., who provided statistical assistance and comments; and to Dr. Maureen O'Sullivan for comments. Thanks to graduate students Carlo González and Johnny Ramirez, and undergraduates Caroline Encarnacion, Timothy Dixon, and Ryan Brewer (New Mexico State University); Jennifer Johnson, Lassiter Speller, Amaris Lyles, Deidre Herring, Kristin Peoples, and Deandra Keys (Winston-Salem State University). Thanks to the Federal Law Enforcement Training Center in Artesia for their assistance.

References

Abernathy, B. (1990). Expertise, visual search, and information pick-up in squash. *Perception, 19*, 63–77.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214–234.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*, 313–329.

Bond, G. D., Malloy, D. M., Arias, E. A., Nunn, S. N., & Thompson, L. A. (2005). Lie-biased decision making in prison. *Communication Reports, 18*, 9–19.

Bond, C. F., & Uysal, A. (2007). On lie detection "wizards". *Law and Human Behavior, 31*, 109–115.

CBS News. (Producer). (1997). *Stalkers: Against the law* [Television broadcast]. (Available from CBS News, 524 W. 57th Street, New York, NY 10019).

Charness, N., Reingold, E. M, Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory and Cognition, 29*, 1146–1152.

DePaulo, B. M., & Pfeiffer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology, 16*, 249–267.

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology, 37*, 1713–1722.

DePaulo, B. M., Rosenthal, R., Rosenkrantz, J., & Green, C. (1982). Actual and perceived cues to deception: A closer look at speech. *Basic and Applied Social Psychology, 3*, 291–312.

Ekman, P. (1985). *Telling lies: Clues to deceit in the marketplace, marriage, and politics*. New York: W. W. Norton.

Ekman, P. (1992). *Telling lies: Clues to deceit in the marketplace, marriage, and politics* (2nd ed.). New York: W. W. Norton.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology, 29*, 288–298.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*, 913–920.

Ekman, P., O'Sullivan, M., & Frank, M. (1999). A few can catch a liar. *Psychological Science, 10*, 263–266.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data. Revised edition*. Cambridge, MA: The MIT Press.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high stake lies. *Journal of Personality and Social Psychology, 72*, 1429–1439.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford: Wiley.

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. Chichester: Wiley.

Hartwig, M., Granhag, P. A., & Vrij, A. (2005). Police interrogation from a social psychology perspective. *Policing and Society, 15*, 379–399.

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.

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior, 23*, 499–516.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469–480.

O'Sullivan, M. (2007). Unicorns or Tiger Woods: Are lie detection experts myths or realities? A response to *On Lie Detection Wizards* by Bond and Uysal. *Law and Human Behavior, 31*, 117–123.

O'Sullivan, M., & Ekman, P. (2004). The wizards of deception detection. In P. A. Granhag & L. Strömwell (Eds.), *The detection of deception in forensic contexts*. London: Cambridge University Press.

Park, H. S., & Levine, T. R. (2001). A probability model of accuracy in deception detection experiments. *Communication Monographs, 68*, 201–210.

Strömwall, L. A., & Granhag, P. A. (2004). *The detection of deception in forensic contexts*. Cambridge: Cambridge University Press.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics, 60*, 926–940.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology, 9*, 159–181.

Vrij, A., Evans, H., Akehurst, L., & Mann, S. (2004). Rapid judgments in assessing verbal and nonverbal cues: Their potential for deception researchers and lie detection. *Applied Cognitive Psychology, 16*, 283–296.

Vrij, A., Mann, S., Robbins, E., & Robinson, M. (2006). Police officers ability to detect deception in high stakes situations and in repeated lie detection tests. *Applied Cognitive Psychology, 20*, 741–755.

Zuckerman, M., Koestner, R., Colella, M. J., & Alton, A. O. (1984). Anchoring in the detection of deception and leakage. *Journal of Personality and Social Psychology, 47*, 301–311.