Journal of Applied Psychology
2004, Vol. 89, No. 1, 137–149

Copyright 2004 by the American Psychological Association, Inc.
0021-9010/04/$12.00 DOI: 10.1037/0021-9010.89.1.137

Detecting True Lies: Police Officers' Ability to Detect Suspects' Lies

Samantha Mann, Aldert Vrij, and Ray Bull
University of Portsmouth

Ninety-nine police officers, not identified in previous research as belonging to groups that are superior in lie detection, attempted to detect truths and lies told by suspects during their videotaped police interviews. Accuracy rates were higher than those typically found in deception research and reached levels similar to those obtained by specialized lie detectors in previous research. Accuracy was positively correlated with perceived experience in interviewing suspects and with mentioning cues to detecting deceit that relate to a suspect's story. Accuracy was negatively correlated with popular stereotypical cues such as gaze aversion and fidgeting. As in previous research, accuracy and confidence were not significantly correlated, but the level of confidence was dependent on whether officers judged actual truths or actual lies and on the method by which confidence was measured.

Overview

Police manuals give the impression that experienced police detectives make good lie detectors (Inbau, Reid, Buckley, & Jayne, 2001), though this claim has not been supported by previous research. The present study is unique, as we tested police officers' ability to distinguish between truths and lies in a realistic setting (during police interviews with suspects), rather than in an artificial laboratory setting. This provides us with a more valid test of Inbau et al.'s claim. Apart from testing truth and lie detection ability, we also examined what characterizes good and poor lie detectors. On the basis of the available deception research, we argue that paying attention to cues promoted in police manuals (gaze aversion, fidgeting, etc.) actually hampers ability to detect truths and lies.

Accuracy Rates and Their Relationships With Background Characteristics

In scientific studies concerning the detection of deception, observers are typically given videotaped or audiotaped statements from a number of people who are either lying or telling the truth. After each statement, observers are asked to judge whether the statement is true or false. In a review of all the literature available at the time, Kraut (1980) found an accuracy rate (percentage of correct answers) of 57%, which is a low score because 50% accuracy can be expected by chance alone. (Guessing whether someone is lying or not gives a 50% chance of being correct.) Vrij (2000a) reviewed an additional 39 studies that were published after 1980 (the year of Kraut's publication) and found an almost identical accuracy rate of 56.6%. In a minority of studies, accuracy in detecting lies was computed separately from accuracy in detecting truth. Where this did occur, results showed a *truth bias*; that is, judges are more likely to consider that messages are truthful than deceptive and, as a result, truthful messages are identified with relatively high accuracy (67%) and deceptive messages with relatively low accuracy (44%). In fact, 44% is below the level of chance, and people would be more accurate at detecting lies if they simply guessed. One explanation for the truth bias is that in daily life, most people are more often confronted with truthful than with deceptive statements and so are therefore more inclined to assume that the behavior they observe is honest (the so-called availability heuristic; O'Sullivan, Ekman, & Friesen, 1988).

Both reviews (Kraut, 1980; Vrij, 2000a) included studies in which college students tried to detect lies and truths in people they were not familiar with. It could be argued that college students are not habitually called on to detect deception. Perhaps professional lie catchers, such as police officers or customs officers, would obtain higher accuracy rates than laypersons. In several studies, professional lie catchers were exposed to videotaped footage of liars and truth tellers and their ability to detect lies was tested (see Vrij & Mann, 2001b, for a review). Three findings emerged from these studies. First, most total accuracy rates were similar to those found in studies with college students as observers, falling in the 45%–60% range. DePaulo and Pfeifer (1986), Meissner and Kassin (2002), and Vrij and Graham (1997) found that police officers were as (un)successful as university students in detecting deception (accuracy rates around 50%). Ekman and O'Sullivan (1991) found that police officers and polygraph examiners obtained similar accuracy rates to university students (accuracy rates around 55%). Second, some groups seem to be better than others. Ekman's research has shown that members of the Secret Service (64% accuracy rate), Central Intelligence Agency (73% accuracy rates), and sheriffs (67% accuracy rates) were better lie detectors than other groups of lie detectors (Ekman & O'Sullivan, 1991; Ekman, O'Sullivan, & Frank, 1999). Third, the truth bias, consistently found in studies with students as observers, is much less profound, or perhaps even lacking, in studies with professional lie catchers (Ekman et al., 1999; Meissner & Kassin, 2002; Porter, Woodworth, & Birt, 2000). Perhaps the nature of their work makes professional lie catchers more wary about the possibility that they are being lied to.

Samantha Mann, Aldert Vrij, and Ray Bull, Psychology Department, University of Portsmouth, Portsmouth, England.

This study was sponsored by Economic and Social Research Council Grant R00429734727.

Correspondence concerning this article should be addressed to Aldert Vrij, Psychology Department, University of Portsmouth, King Henry Building, King Henry 1 Street, Portsmouth, PO1 2DY, United Kingdom. E-mail: aldert.vrij@port.ac.uk

In summary, even the accuracy rates for most professional lie catchers are modest, raising serious doubt about their ability to detect deceit. However, these disappointing accuracy levels may be the result of an artifact. In typical deception studies, including those with professional lie catchers, observers detect truths and lies told by college students who are asked to lie and tell the truth for the sake of the experiment in university laboratories. Perhaps in these laboratory studies the stakes (negative consequences of being caught and positive consequences of getting away with the lie) are not high enough for the liar to exhibit clear deceptive cues to deception (Miller & Stiff, 1993), which makes the lie detection task virtually impossible for the observer.

To raise the stakes in laboratory experiments, participants are offered money if they successfully get away with their lies (Vrij, 1995), or participants (e.g., nursing students) are told that being a good liar is an important indicator of success in a future career (Ekman & Friesen, 1974; Vrij, Edward, & Bull, 2001a, 2001b). In some studies, participants are told that they will be observed by a peer who will judge their sincerity (DePaulo, Stone, & Lassiter, 1985b). In a series of experiments in which the stakes were manipulated, researchers found that such "high-stakes" lies were easier to detect than low-stakes lies (Bond & Atoum, 2000; DePaulo, Kirkendol, Tang, & O'Brien, 1988; DePaulo, Lanier, & Davis, 1983; DePaulo, LeMay, & Epstein, 1991; DePaulo et al., 1985b; Feeley & deTurck, 1998; Forrest & Feldman, 2000; Heinrich & Borkenau, 1998; Lane & DePaulo, 1999; Vrij, 2000b; Vrij, Harden, Terry, Edward, & Bull, 2001).

In an attempt to raise the stakes even further, participants in Frank and Ekman's (1997) study were given the opportunity to "steal" $50. If they could convince the interviewer that they had not taken the money, they could keep all of it. If they took the money and the interviewer judged them as lying, they had to give back the $50 in addition to their $10 per hour participation fee. Moreover, some participants faced an additional punishment if they were found to be lying. They were told that they would have to sit on a cold metal chair inside a cramped, darkened room ominously labeled *XXX*, where they would have to endure anything from 10 to 40 randomly sequenced, 110-decibel starting blasts of white noise over the course of 1 hr.

A deception study like this probably borders on unethical, and yet the stakes are still not comparable with the stakes in real-life situations in which professional lie catchers operate, such as during police interviews. Therefore, one might argue that the only valid way to investigate police officers' true ability to detect deceit is to examine their skills when they detect lies and truths that are told in real-life criminal investigation settings. Vrij and Mann (2001a, 2001b) were the first researchers to do this. Vrij and Mann (2001a) exposed police officers to fragments of a videotaped police interview with a man suspected of murder. However, that study had two limitations. First, fragments of only one suspect were shown, and second, the police officers could not understand the suspect because he spoke a foreign language (suspect and police officers were of different nationalities). Vrij and Mann (2001b) later exposed judges to videotaped press conferences of people who were asking the general public for help in finding either their missing relatives or the murderers of their relatives. They all lied during these press conferences, and they were all subsequently found guilty of having killed the "missing person" themselves. This study had limitations as well. First, the judges were only subjected to lies, and, second, again the lie detectors and liars spoke in different languages, as they were from different nationalities.

We overcame these limitations in the present experiment. We exposed British police officers to fragments of videotaped real-life police interviews with English-speaking suspects and asked them to detect truths and lies told by these suspects during these interviews. We expected truth and lie accuracy rates to be significantly above the level of chance (which is 50%), and, as a consequence of this, expected lie accuracy rates to be significantly higher than those typically found in previous research (44%; Hypothesis 1). In view of the fact that police officers in the present study were assessing the veracity of suspects, a group that is likely to arouse heightened skepticism in a police officer (Moston, Stephenson, & Williamson, 1992), a truth bias is unlikely to occur.

We also expected individual differences, with some police officers being more skilled at detecting truths and lies than others. We predicted that the reported experience in interviewing suspects would be positively correlated with truth and lie accuracy (Hypothesis 2). This background characteristic has not been examined in deception research before, but we expected it to be related to accuracy, as it is this particular aspect of police work that gives police officers experience in detecting lies and truths. Previous research has focused on the relationship between length of service/years of job experience and accuracy, and a significant relationship between the two was not found (Ekman & O'Sullivan, 1991; Porter et al., 2000; Vrij & Mann, 2001b). This is not surprising, as an officer who has served in the police force for many years will not necessarily have a great deal of experience in interviewing suspects, and vice versa. Other background characteristics, such as age and gender, have generally not been found to be related to accuracy (DePaulo, Epstein, & Wyer, 1993; Ekman & O'Sullivan, 1991; Ekman et al., 1999; Hurd & Noller, 1988; Köhnken, 1987; Manstead, Wagner, & MacDonald, 1986; Porter et al., 2000; Vrij & Mann, 2001b).

Cues Used to Detect Deceit

We asked lie detectors to indicate which verbal and nonverbal cues they typically use to decide whether someone is lying, so-called beliefs about cues associated with deception (DePaulo, Stone, & Lassiter, 1985a; Zuckerman, DePaulo, & Rosenthal, 1981). We expected good lie detectors to mention speech-related cues significantly more often than poor lie detectors (Hypothesis 3). In part, this is because research has shown that the intellectual ability of suspects who are interviewed by the police is often rather low. Gudjonsson (1994) measured intellectual functioning with three subtests of the Wechsler Adult Intelligence Scale—Revised (WAIS–R; Wechsler, 1981)—Vocabulary, Comprehension, and Picture Completion—and found a mean IQ of 82, with a range of 61–131. It might well be that people with a low IQ will find it hard to tell a lie that sounds plausible and convincing (Ekman & Frank, 1993). Moreover, in their review of detection of deception research, DePaulo et al. (1985a) found that lie detectors who read transcripts only (and are therefore "forced" to focus on story cues) are typically better lie detectors than those who are exposed to the actual person (speech, sound, and behavior; see also Wiseman, 1995).

Stereotypical views typically held among professional lie catchers (and also laypersons) is that liars look away and fidget (Akehurst, Köhnken, Vrij, & Bull, 1996; Vrij & Semin, 1996). These cues, however, are unrelated to deception (see DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Vrij, 2000a, for reviews about nonverbal and verbal cues to deceit). We therefore expected negative correlations between mentioning such cues and accuracy rates; in other words, the more of these cues the officers reported to look at, the lower their accuracy rates would become (Hypothesis 4).

In their influential manual about police interviewing, *Criminal Interrogation and Confessions*, Inbau, Reid, and Buckley (1986; a new edition was recently published; Inbau et al., 2001) described in detail how, in their view, liars behave. As evidence, the authors included showing gaze aversion, displaying unnatural posture changes, exhibiting self-manipulations, and placing the hand over the mouth or eyes when speaking. None of these behaviors have been found to be reliably related to lying in deception research. It is therefore not surprising that participants in a deception detection study by Kassin and Fong (1999), who were trained to look at the cues Inbau and colleagues claim to be related to deception, actually performed worse than naive observers who did not receive any information about deceptive behavior. In the present study, we expected negative correlations between reporting "Inbau cues" and accuracy. In other words, the more of these Inbau cues that police officers mentioned that they use to detect deceit, the worse we expected them to be at distinguishing between truths and lies (Hypothesis 5).

We also examined whether the cues lie detectors used to make their veracity judgments were related to the behaviors shown by the suspects in the videotape (so-called cues to perceived deception; Zuckerman et al., 1981).[1] We predicted that poor lie detectors would be significantly more guided by invalid cues, such as gaze aversion, than good lie detectors (Hypothesis 6).

Accuracy–Confidence Relationship

Studies investigating lie detectors' confidence in their decision making typically reveal three findings. First, there is usually no significant relationship between confidence and accuracy (see DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997, for a meta-analysis). Second, confidence scores among professional lie catchers are typically high (Allwood & Granhag, 1999; DePaulo & Pfeifer, 1986; Strömwall, 2001; Vrij, 1993), and police officers are sometimes found to be more confident than laypersons (Allwood & Granhag, 1999; DePaulo & Pfeifer, 1986). Furthermore, DePaulo et al. (1997) found an "overconfidence effect"; that is, judges' confidence is typically higher than their accuracy. Third, observers tend to have higher levels of confidence when judging truthful statements than when judging deceptive statements, irrespective of whether they judge the statement as a truth or a lie (DePaulo et al., 1997).

In the present study, confidence was investigated in two different ways. First, it was investigated in the traditional way, by asking observers after each veracity judgment how confident they were of their decision. Second, we also asked participants at the end of the lie detection experiment how well they thought they had done at the task. This latter method of measuring confidence may well result in more accurate confidence levels (less prone to an overconfidence effect), because at that stage lie detectors have insight into their overall performance and are asked to judge this overall performance. For this reason the latter method may even result in a positive relationship between accuracy and confidence. This issue was explored in the present study.

Method

Participants

Ninety-nine Kent County Police Officers (Kent, England) participated. Of these, 24 were women and 75 were men. Ages ranged from 22 years to 52 years, with a mean average of 34.3 years ($SD = 7.40$ years). Seventy-eight participants were from the Criminal Investigation Department (CID), 8 were police trainers, 4 were traffic officers, and the remaining 9 were uniform response officers. Although different groups of police officers participated, none of these groups are the specialized groups that are identified by Ekman and his colleagues as particularly good lie detectors (Ekman & O'Sullivan, 1991; Ekman et al., 1999). As some of the group sizes are rather small, differences between groups are not discussed in the main text.

Length of service on the job ranged from 1 year to 30 years, with a mean average of 11.2 years ($SD = 7.31$ years). The distribution of this variable differed significantly from a normal distribution ($z = 1.83$, $p < .01$, skewness = .94, $Mdn = 9$ years).

Materials

Participants in this study were asked to judge the veracity of people in real-life high-stakes situations. More specifically, participants saw video clips of 14 suspects (of whom 12 were men, 4 of whom were juvenile, and 2 were women) in their police interviews. The interview rooms were fitted with a fixed camera, which produces the main color picture and is aimed at the suspect's chair, and a small insert picture, produced by a wide lens camera. The picture in the small insert was not of good quality and displayed the whole interview room from the view taken at the back of the suspect. The purpose of the wide lens insert is to show how many people are present in the room and any larger movements made by any person present (therefore proving or disproving that the officers might have physically threatened or coerced the suspect in some way).[2] The quality of

[1] Investigating *beliefs about cues associated with deception* provides insight into which cues people think they use when detecting deceit, but it does not necessarily mean that they actually use these cues when they try to detect deceit. For example, people may indicate that they use gaze aversion as a cue for deceit, but it still may be the case that they subsequently judge someone who shows gaze aversion to be truthful. Investigating *cues to perceived deception* provides insight into which cues lie detectors actually use to indicate deception, but it is not certain whether they actually realize this. For example, when there is a tendency among lie detectors to judge those who moved a great deal as more deceptive than those who made few movements, it can be concluded that they used making movements as a cue to detect deception. It is, however, unclear whether lie detectors realized that they used making movements as a cue to detect deceit. The combination of those two methods therefore provides the most complete insight.

[2] The picture in the small insert was not clear enough to enable the viewer to see any detail like, for example, the expressions of the interviewer. It is therefore unlikely that the participants paid any attention to this small insert picture (nobody mentioned that they did), and so it is unlikely that participants have been guided by the behavior or demeanor of the interviewer when judging the veracity of the suspects. (When the participants were asked afterward to indicate what made them decide whether the suspect on the screen was lying, nobody mentioned that they had been influenced by the interviewer.)

the main picture was good enough to code the occurrences of eye blinks, but not good enough to see subtler facial changes. Sound quality was good in all interviews. The positioning of the cameras varied slightly, depending on which interview room the interview was conducted, but in all cases the suspect's upper torso could be seen. However, in some cases the lower torso could not be seen, hence leg and foot movements were not analyzed.[3] In the main picture only the suspect was visible. Crimes about which the suspects were being interviewed included theft (9), arson (2), attempted rape (1), and murder (2). Cases had been chosen in which other sources (reliable, independent witness statements and forensic evidence) provided evidence that the suspect told the truth and lied at various points within the interview. Once a case had been selected, only those particular clips in which each word was known to be a truth or a lie were selected. The truths that were selected were chosen so as to be as comparable as possible in nature to the lies (e.g., a truthful response to an easy question such as giving a name and address is not comparable to a deceitful response regarding whether the suspect had committed a murder. Video footage about names and addresses were therefore not included as truths in this study). The following account is an example of one of the cases used: The suspect (a juvenile) spent the night in a derelict building with a friend. With the friend, he shot at windows of a neighboring house with his air rifle and then stole items from a local shop. The suspect denied involvement in any of those activities and provided an alibi. His friend (the alibi), however, immediately admitted to both his and the suspect's part in the offenses. The suspect's alibi fell through, and so the suspect confessed to the crimes and told police of the whereabouts of the stolen goods, his gun, and from where he purchased it. The suspect admitted guilt and was charged accordingly. Lies included in clips were the initial denials of any involvement in the crimes. It is important to point out that, rather than take the form of a straightforward "No, I didn't do it" and "Yes, I did do it," all clips used in this study contained story elements that were true and false. So in the above example, in the denial the suspect gave an alternative story of the events of the day to those that actually occurred (that he went over to another friend's house, etc.), and in the confession he gave a true version of events, not all of which was necessarily incriminating. Therefore, a participant watching the clips, who does not know the facts of the case, would not easily be able to tell what are snippets of denial and what are snippets of confession. See Mann, Vrij, and Bull (2002) for further details.[4]

The length of each clip unavoidably varied considerably (from 6 s to 145 s). There were 54 clips total (23 truthful clips and 31 deceptive clips), and the number of clips for each suspect varied between a minimum of 2 and a maximum of 8 clips (each suspect with at least one example of a truth and a lie). The total length of the video clips of all 14 suspects was approximately 1 hr. Clearly it would be impossible to show each participant all the clips because of logistical constraints and fatigue. Therefore, the clips were divided between four tapes of roughly equal length, and 24–25 participants saw each clip. As mentioned above, the length of the clips varied, and so each of the four tapes contained between 10 and 16 clips (Clip 1: $n = 15$, 6 truths and 9 lies; Clip 2: $n = 16$, 6 truths and 10 lies; Clip 3: $n = 10$, 5 truths and 5 lies; Clip 4: $n = 13$, 6 truths and 7 lies). Those suspects for whom there were several clips may have had clips spread over several of the tapes. However, for each suspect there was always at least one example of a lie and a truth present on each tape on which they appeared. Clips were presented on the tapes in random order so that the same suspect did not appear in consecutive clips. Two analyses of variance (ANOVAs), with tape as the between-subjects factor and lie accuracy and truth accuracy as dependent variables were conducted to examine possible differences in accuracy between the four tapes. Neither of the two ANOVAs were significant for truth accuracy, $F(3, 95) = 0.20$, *ns*, $\eta^2 = .00$; or for lie accuracy, $F(3, 95) = 1.57$, *ns*, $\eta^2 = .05$. Hence, the fact that participants did not all judge exactly the same clips was not considered an issue, and accuracy scores were collapsed over the four tapes in all subsequent analyses.

Procedure

Permission to approach police officers was granted by the Chief Constable in the first instance, and then by appropriate superintendents. Participants were recruited on duty from either the training college where they were attending courses or various police stations within Kent. Participants were approached and asked whether they would participate in a study about police officers' ability to detect deception and informed that their participation would be anonymous. Participants completed the deception detection task individually. Before attempting the task, participants filled out a questionnaire. This included details such as age, gender, length of service, division, perceived level of experience in interviewing suspects (1 = *totally inexperienced*, 5 = *highly experienced*; $M = 3.75$, $SD = 0.85$), and the verbal or nonverbal cues they use to decide whether another person is lying or telling the truth. After completion of this section, each participant was then read the following instructions: "You are about to see a selection of clips of suspects who are either lying or telling the truth. The clips vary considerably in length, and the suspects may appear on several occasions. This is irrelevant. They will be either lying the whole length of the clip or truth-telling for the length of the clip. After viewing each clip I would like you to indicate whether you think the suspect is lying or telling the truth (measured with a dichotomous scale), and how confident you are of your decision, on a seven-point scale. If you recognize any of the suspects please bring it to my attention." (This latter point was not an issue.) Participants were not informed of how many clips they were going to see, or of how many instances of lies and truths they would see.

After completing the task, participants answered a remaining few questions on the questionnaire. These included questions about what behaviors they had used to guide them in making veracity judgments and questions measuring their confidence. Depending on the participant, participation time lasted between 45 and 90 min. After each veracity judgment, participants were shown each clip again and were asked several questions about the clip. This (time-consuming) part of the experiment is beyond the scope of this article and is therefore not addressed further (see Mann, 2001, for further details about this aspect of the study). The variation in participation time was the result of several factors. Some participants took longer than others to complete their forms; some participants took slightly longer to reach a decision, but the largest time range was in the amount of time taken, and detail given, in responding to the questions that were asked about each clip after the veracity judgment had been given.

Dependent Variables

The dependent variables for this study were the accuracy scores, the behaviors that participants associated with deception before and after the task, cues to perceived deception, and their confidence scores during and after the task.

Accuracy was calculated by assigning a score of 1 when the participant correctly identified a truth or a lie, and assigning a score of 0 when the participant was incorrect. The lie accuracy score was calculated by dividing the number of correctly classified lies by the number of lies shown on the

[3] Although it is unfortunate that sometimes the lower torso could not be seen, this is not atypical for detection of deception research, as in many studies, including Ekman et al. (1999), only the head and shoulders are visible.

[4] Mann et al. (2002) examined the behaviors of 16 suspects. However, 2 of those suspects were omitted for the purpose of this study. Those 2 were too well-known to show the clips to participants, as they were higher profile cases that received some media attention. We did not want participants to know the cases that they were seeing, as obviously this would give them an advantage, and they may score high accuracy, not on the merits of the task, but purely on facts that they already knew.

tape, and the truth accuracy score was calculated by dividing the number of correctly classified truths by the number of truths shown on the tape.

The behaviors that participants typically use to detect deception were investigated with the open-ended question, "What verbal or nonverbal cues do you use to decide whether another person is lying or telling the truth?" The behaviors that participants said they used in the present lie detection task were investigated with the open-ended question, "What verbal or nonverbal cues did you use in this task to decide whether the people on the screen were lying or telling the truth?" In other words, cues to deception both prior to and after the deception task were investigated. A similar procedure was used by Ekman and O'Sullivan (1991). Asking this question twice enabled us to explore whether, in our deception task, the police officers paid attention to cues they typically consider they pay attention to. In case they did, the answers they would give to the "prior" and "after" questions would be similar, whereas the answers would be different in case they did not. We expected similar responses. We had no reason to believe that the police officers would find the responses of the suspects atypical (they were a random sample of suspects' responses in connection with serious crimes), nor did we think that officers would change their mind about beliefs about deception on the basis of a single lie detection task. We return to this issue in the Discussion section. These two open-ended questions were coded by two coders into 30 different cues. Appendix A shows the list of 30 cues.

This list was the result of sorting and tallying all participants' comments into various groups and combining them as much as possible within specific headings to make the system as manageable as possible. Once the coding system was created, the creator coded for each participant every behavior that they mentioned on the questionnaire before and after the task. Another independent coder then used the coding system also to code each behavior mentioned by participants on the questionnaires to determine the reliability of the coding system. For each participant, each code was used a maximum of one time before the task and one time after the task. So, for example, if a participant said before the task "eyes looking up, looking away from interviewer, high-pitched voice, vocally loud," then just the codes gaze and voice would be recorded, even though, in effect, the participant mentioned two aspects of gaze and two aspects of voice. For the 99 participants, 677 behaviors mentioned on the questionnaires before and after completing the task were coded. Hence, each participant mentioned a mean of 6.84 behaviors. In 651 (96.2%) of the 677 mentioned behaviors, the two coders agreed; any disagreements were resolved by discussion. Twenty-nine of those categories could be clustered into four categories: story, vocal, body, and conduct (see the second column of Appendix A). Those four categories have been introduced by Feeley and Young (2000). The total number of times each participant mentioned behaviors in each group was calculated. So, for example, if a participant mentioned gaze aversion and movements, that participant would obtain a score of 2 for the body category. As a result, the scores for story cues could range from 0 to 6, the scores for vocal cues from 0 to 5, the scores for body cues from 0 to 14, and the scores for conduct cues from 0 to 4. One cue, gut feeling, could not be included in any of these categories, and we therefore analyze the data for this cue separately.

To compare good and poor lie detectors, we followed Ekman and O'Sullivan's (1991) procedure and divided the lie detectors into two ability groups. Good lie detectors ($n = 27$) were those who had scored above the mean for lie clips (66.16%, see the Results section) and above the mean for truth clips (63.61%, see the Results section). Poor lie detectors ($n = 72$) were those who remained, who may well have scored very well on either truth clips or lie clips, but did not score above the mean for both.

To test Hypothesis 4, we constructed new variables: "popular stereotypical beliefs" (one variable was created for cues mentioned before the task and one for cues mentioned after the task). These variables included three cues (gaze, fidget, and self-manipulation) and could range from 0 to 3.

To test Hypothesis 5, two further new variables were created: Inbau cues (again, separate variables were created for cues mentioned before the task and cues mentioned after the task). Inbau cues included the following five cues: posture, cover, gaze, fidget, and self-manipulation; they could range from 0 to 5.

To examine cues to perceived deception, 13 behaviors of the suspects in the clips were scored by two independent coders with a coding scheme used previously by Vrij and colleagues (Vrij, 1995; Vrij et al., 2001a, 2001b; Vrij, Edward, Roberts, & Bull, 2000; Vrij, Semin, & Bull, 1996; Vrij & Winkel, 1991). An overview of these behaviors and the interrater agreement rates between the two coders (Pearson correlations) are reported in Appendix B. Differences between truth tellers and liars regarding these behaviors have been discussed elsewhere in detail (Mann et al., 2002). To summarize the findings, liars blinked less and included more pauses in their speech.

Confidence was measured in two ways. First, participants indicated after each veracity judgment how confident they were in their decision (1 = *not at all confident*, and 7 = *very confident*). Second, after completing the lie detection task, the participants were also asked to answer the open-ended question, "What percentage of answers do you think you answered correctly?"

Results

Accuracy Rates and Their Relationships With Background Characteristics

For the whole sample, the mean lie accuracy was 66.16% ($SD = 17.0$), and the mean truth accuracy was 63.61% ($SD = 22.5$). The difference between lie and truth accuracy was not significant, $t(98) = 0.87$, *ns*, $d = 0.09$; neither were lie and truth accuracy significantly correlated with each other, $r(99) = .08$, *ns*.[5]

Both accuracy rates were significantly higher than the level of chance, which is 50%; truth accuracy, $t(98) = 6.02$, $p < .01$, $d = 0.60$; lie accuracy, $t(98) = 9.43$, $p < .01$, $d = 0.95$. (See Clark-Carter, 1997, for conducting *t* tests when the standard deviation of the sample is unknown.) Moreover, the lie accuracy rate was significantly higher than the average lie accuracy rate that was found in Vrij's (2000a) review of previous research (lie accuracy: $M = 66.16\%$ vs. 44.00%), $t(98) = 12.93$, $p < .01$, $d = 1.30$. Truth accuracy did not differ significantly from what has previously been found (63.61% vs. 67.00%), $t(98) = 1.50$, *ns*, $d = 0.15$. This supports Hypothesis 1.

Pearson correlations revealed that experience in interviewing, however, was significantly correlated with truth accuracy, $r(99) = .20$, $p < .05$. The correlation with lie accuracy was $r(99) = .18$, $p = .07$. These positive correlations indicate that the more experienced the police officers perceived themselves to be in interviewing suspects, the better they were in the lie detection task. This supports Hypothesis 2. Age and length of service were unrelated to lie accuracy, $r(99) = -.09$, *ns*, and $r(99) = -.04$, *ns*, respectively; and truth accuracy, $r(99) = .01$, *ns*, and $r(99) = -.07$, *ns*, respectively. Age and length of service were strongly correlated,

[5] Separate analyses for poor and good lie detectors showed that the truth–lie accuracy correlation was significant for poor lie detectors, $r(72) = -.35$, $p < .01$, but not for good lie detectors, $r(27) = -.21$, *ns*. A negative correlation means that the better poor lie detectors were at detecting truths, the worse they were at detecting lies, and vice versa. However, truth and lie accuracy did not differ significantly from each other for poor lie detectors (truth accuracy: $M = 57.61$, $SD = 22.7$; lie accuracy: $M = 61.37$, $SD = 16.6$), $t(71) = 0.97$, *ns*, and good lie detectors (truth accuracy: $M = 79.61$, $SD = 11.5$; lie accuracy: $M = 78.94$, $SD = 10.4$), $t(26) = 0.20$, *ns*.

$r(99) = .80$, $p < .01$; whereas age and experience in interviewing, $r(99) = .34$, $p < .01$, and experience in interviewing and length of service, $r(99) = .46$, $p < .01$, were moderately correlated.

Men were significantly better at detecting truths ($M = 66.61\%$, $SD = 21.9$) than women ($M = 54.22\%$, $SD = 22.3$), $t(97) = 2.40$, $p < .05$, $d = 0.56$; but no differences were found for detecting lies, $t(97)= .41$, *ns*, $d = 0.09$ ($M = 66.56\%$, $SD = 17.0$ vs. $M = 64.92\%$, $SD = 17.7$).[6]

Cues Used to Detect Deceit

Appendix A shows how many police officers mentioned that they use the cues to detect deceit before and after the task. The most frequently mentioned cue was gaze, with 73% of the officers ($n = 72$) mentioning the cue before the task and 78% ($n = 77$) after the task. The second most frequently mentioned cue was movements, which was mentioned by 25 police officers before the task and by 31 officers after the task. Also vagueness, contradictions, miscellaneous speech (a category for speech-related cues that does not fit into other categories, e.g., pleading/minimizing offense or uncertain replies; all story cues), and fidgeting were relatively frequently mentioned.

ANOVAs comparing how many cues were mentioned in each category (ANOVA, with cue category—story, vocal, body, and conduct—as the single within-subjects factor) showed significant differences in the number of cues mentioned, both before the task, $F(3, 96) = 58.56$, $p < .01$, $\eta^2 = .65$; and after the task, $F(3, 96) = 85.61$, $p < .01$, $\eta^2 = .73$. Before the task, police officers mentioned a mean of 1.84 body cues ($SD = 1.05$; see also Table 1). Tukey's honestly significant difference test revealed that this is significantly more than any of the other three categories of cues. They also mentioned significantly more story cues than conduct and vocal cues before the task. The latter two categories did not differ significantly from each other. Exactly the same pattern emerged for cues mentioned after the task.

To compare the number of cues mentioned before and after the task, we conducted a multivariate analysis of variance (MANOVA),with time (before or after) as the within-subjects factor and the four categories of cues as dependent variables. At a multivariate level, the test revealed a nonsignificant effect, $F(4, 95) = 1.91$, *ns*, $\eta^2 = .07$. In other words, the lie detection task did not influence the police officers' ideas about which cues to attend to in order to detect deceit. However, if we look at individual cues (see Appendix A), rather than the categories, differences emerged regarding some cues. Sharp increases in cues mentioned (between before and after the task) occurred for self-corrections, miscellaneous speech, hand movements and head movements; and sharp decreases were found for contradictions, evidence, facial cues and physiological cues; the latter findings might be the result of the experimental setting. For example, noticing physiological cues might be difficult when watching a tape, hence, in this lie detection task, participants did not look for such cues to detect deceit.

Table 1
Overview of the Total Number of Times Each Participant Mentioned Story, Vocal, Body, and Conduct Cues Before and After the Task

	Before the task		After the task	
Cue	*M*	*SD*	*M*	*SD*
Story	0.68_b	0.62	0.78_b	0.72
Vocal	0.40_a	0.59	0.48_a	0.61
Body	1.84_c	1.05	2.01_c	0.96
Conduct	0.32_a	0.49	0.28_a	0.50

Note. Only mean scores in the columns with a different subscript differ significantly from each other.

To investigate and compare behaviors mentioned by good and poor lie detectors, we conducted ANOVAs, with skill (good or poor lie detector) as the between-subjects factor and the four cue categories as dependent variables. As predicted in Hypothesis 3, good lie detectors were more inclined to claim that they focused on story cues ($M = .89$, $SD = .60$) than poor lie detectors ($M = .60$, $SD = .60$), $F(1, 97) = 4.50$, $p < .05$, $\eta^2 = .04$; although this effect only occurred for the story cues mentioned before the task. All other effects were not significant.[7]

For the remaining cue, gut feeling, we used chi-square analyses to compare responses from good and poor lie detectors. After the lie detection task, none of the 72 poor lie detectors said that they had relied on gut feeling, whereas 11% ($n = 3$) of the good lie detectors claimed to have relied on such intuitive feelings, $\chi^2(1, N = 99) = 8.05$, $p < .01$, $\phi = .29$. The analysis for mentioning gut

[6] A 2 (veracity) × 2 (gender of suspect) × 2 (gender of observer) ANOVA, with a mixed factorial design (the first two factors were within-subjects factors), was carried out to investigate the gender issue in more detail. In this analysis, only 74 participants (58 men and 16 women) were included because on one tape, no female suspects appeared. Apart from a significant gender of observers effect, $F(1, 72) = 5.51$, $p < .05$, $\eta^2 = .07$ (indicating that male accuracy was superior, $M = 68\%$, $SD = .13$, to female accuracy, $M = 58\%$, $SD = .20$), a Deception × Gender of Suspect effect occurred, $F(1, 72) = 15.09$, $p < .01$, $\eta^2 = .17$. In male suspects, lies ($M = 72.00$, $SD = 19.11$) were more easily detected than truths ($M = 59.73$, $SD = 24.44$), whereas in female suspects, truths ($M = 79.73$, $SD = 40.48$) were more easily detected than lies ($M = 52.03$, $SD = 43.44$). However, participants only saw three deceptive clips of female suspects (and six truthful clips), so conclusions have to be drawn with caution. A significant difference between the four groups was found for detecting lies, $F(3, 95) = 4.44$, $p < .01$, $\eta^2 = .12$.

The 4 traffic officers who participated were highly accurate ($M = .95$, $SD = .58$), and Tukey's honestly significant difference test revealed that they were more accurate than any of the other three groups of participants (which did not differ significantly from each other). Because only 4 traffic officers participated, it would be presumptuous to assume that this sample is representative of all traffic officers and claim that all officers in this area of specialty would be more accurate at detecting deception. However, reasons why traffic officers may be more accurate than officers from other divisions include that they are more used to making snap judgments (e.g., highway patrol officers) about whether a person is drinking, or is lying about their involvement in a crash, and so on. Also they may speak to more people on a daily basis, because many traffic offenses are fairly quick to deal with, and hence traffic officers are more practiced in making veracity judgments than officers in other departments.

[7] These findings are available from Aldert Vrij.

feeling before the task was not significant, $\chi^2(1, N = 99) = 2.63$, *ns*, $\phi = .17$.[8]

ANOVAs further revealed that good and poor lie detectors did not differ significantly from each other on the newly created variables popular stereotypical beliefs and Inbau cues. A disadvantage of using a dichotomization procedure (i.e., dividing the lie detectors into two groups) is a loss in data measurement, because many participants are treated alike in a dichotomy when in fact they are different. An alternative method is to keep the continuous lie and truth accuracy scores. Pearson correlations (see Table 2) revealed that mentioning popular stereotypical beliefs and Inbau cues (both before and after the task) were negatively correlated with accuracy. This supports Hypotheses 4 and 5.[9]

To investigate cues to perception, we conducted multiple stepwise regression analyses. The units of analysis were the 54 different clips. The criterion was the percentage of police officers who judged the suspect in the clip as lying. The predictors were the behaviors displayed by the suspects (13 behaviors were entered, see the Method section), the veracity of the suspects' statements, and age (adult or juvenile) and gender of the suspects. Different analyses were carried out for good and poor lie detectors. The analysis for good lie detectors revealed two predictors, which explained 61% of the variance, $F(2, 51) = 40.16$, $p < .01$; these were veracity of the clip ($R = .76$, $\beta = .74$), $t(53) = 8.48$, $p < .01$, and illustrators ($R = .02$, $\beta = -.19$), $t(53) = 2.18$, $p < .05$. Participants were most likely to judge the clip as deceptive if the clip was in fact a lie, and the fewer illustrators the suspects made, the more likely it was that they were judged as deceptive.

The analysis for poor lie detectors revealed four predictors, which explained 51% of the variance, $F(4, 49) = 12.60$, $p < .01$. These were gender of the suspect ($R = .38$, $\beta = .57$), $t(53) = 5.14$, $p < .01$; veracity of the clip ($R = .18$, $\beta = .42$), $t(53) = 4.06$, $p < .01$; gaze aversion ($R = .10$, $\beta = .41$), $t(53) = 4.02$, $p < .01$; and head nods ($R = .04$, $\beta = .25$), $t(53) = 2.28$, $p < .05$. Participants were most likely to judge the clip as deceptive if the suspect was a man and if the clip was in fact a lie. Moreover, the more gaze aversion and the more head nods the suspects made, the more likely it was that they were judged as deceptive.

The correlation between displaying gaze aversion and judging the person as deceptive was significant for poor lie detectors, $r(54) = .39$, $p < .01$, but not significant for good lie detectors, $r(54) = .06$, *ns*. These two correlations differed significantly from each other ($z = 1.78$, $p < .05$, one-tailed). This supports Hypothesis 6 (poor lie detectors would be more guided by gaze aversion than good lie detectors).

Table 2
Pearson Correlations Between Popular Stereotypical Beliefs and Inbau Cues With Truth Accuracy and Lie Accuracy

	Before the task		After the task	
Variable	Truth accuracy	Lie accuracy	Truth accuracy	Lie accuracy
Popular stereotypical beliefs	−.21*	−.02	−.22*	−.08
Inbau cues	−.23*	−.06	−.23*	−.05

Note. * $p < .05$.

Accuracy–Confidence Relationship

Participants were significantly more confident after they saw a truthful clip ($M = 4.55$, $SD = 0.92$) than after watching a deceptive clip ($M = 4.38$, $SD = 0.95$), $t(98) = 3.08$, $p < .01$, $d = 0.18$. Those two confidence measures were significantly correlated with each other, $r(99) = .82$, $p < .01$. The police officers estimated their percentage of correct answers ("posttask estimated accuracy," measured after the lie detection task) very modestly ($M = 49.98\%$, $SD = 15.08$). This percentage was significantly lower than the actual truth accuracy ($M = 63.61$, $SD = 22.50$), $t(98) = 5.65$, $p < .01$, $d = 0.52$, and lie accuracy ($M = 66.16$, $SD = 17.05$), $t(98) = 7.05$, $p < .01$, $d = 0.74$, obtained in the lie detection task.

Neither the truth accuracy–truth confidence correlation, $r(99) = .10$, nor the lie accuracy–lie confidence correlation, $r(99) = .03$, were significant. Neither was the posttask estimated accuracy significantly correlated with the actual lie accuracy, $r(99) = -.07$, or actual truth accuracy, $r(99) = .17$. Age, length of service, and experience in interviewing suspects were not significantly correlated with truth confidence, lie confidence, or posttask estimated accuracy. Neither were there significant differences found between men and women on any of these three variables, although the difference between men and women for posttask estimated accuracy was marginally significant, $t(97) = 1.93$, $p = .056$, $d = 0.47$ (women were more skeptical about their performance, $M = 44.38\%$, $SD = 13.21$, than men, $M = 51.12\%$, $SD = 15.35$).[10]

Discussion

Accuracy Rates and Their Relationships With Background Characteristics

In the present study, 99 police officers, who did not belong to a group that has been identified as specialized in lie detection,

[8] Following Anderson, DePaulo, Ansfield, Tickle, and Green (1999), who found gender differences in cues mentioned, we conducted ANOVAs and chi-square analyses, with gender as the between-subjects factor and the four categories and gut feeling as dependent variables. We only found one significant difference: Before the task, female participants mentioned more body cues ($M = 2.21$, $SD = 1.18$) than male participants ($M = 1.72$, $SD = .98$), $F(1, 97) = 4.08$, $p < .05$, $\eta^2 = .04$.

[9] To explore gender differences in how often popular stereotypical beliefs and Inbau cues were mentioned, ANOVAs were carried out, with gender as the between-subjects factor and popular stereotypical beliefs and Inbau cues as dependent variables. Before the task, women mentioned popular stereotypical beliefs significantly more often ($M = 1.29$, $SD = 0.69$) than men ($M = 0.89$, $SD = 0.65$), $F(1, 97) = 6.65$, $p < .05$, $\eta^2 = .06$. Also after the task, women mentioned these cues more often ($M = 1.17$, $SD = 0.76$) than men ($M = 0.88$, $SD = 0.59$); although the difference was borderline significant, $F(1, 97) = 3.69$, $p = .058$, $\eta^2 = .03$.

Before the task, women mentioned Inbau cues significantly more often ($M = 1.46$, $SD = 0.78$) than men ($M = 1.00$, $SD = 0.72$), $F(1, 97) = 7.13$, $p < .01$, $\eta^2 = .07$. No gender differences emerged regarding the mention of Inbau cues after the task (men: $M = 1.09$, $SD = 0.74$; women: $M = 1.38$, $SD = 0.82$), $F(1, 97) = 2.50$, *ns*, $\eta^2 = .03$.

[10] Differences between the four groups (Criminal Investigation Department, police trainers, traffic officers, and uniform response officers) were not found on any of the three (truth confidence, lie confidence, and posttask estimated accuracy) confidence scores (all *p*s > .32).

attempted to detect lies and truths told by suspects during their police interviews. Regarding accuracy, two main findings emerged. First, truth accuracy and lie accuracy were both around 65% in this study, which was higher than was found in most previous deception detection studies. It is also the highest accuracy rate ever found for a group of "ordinary" police officers. The accuracy rates found in this sample of ordinary police officers were comparable to those found among specialized groups of lie detectors in previous studies (Ekman & O'Sullivan, 1991; Ekman et al., 1999). In other words, ordinary police officers might well be better at detecting truths and lies than was previously suggested. Although the accuracy rates were significantly higher than the average accuracy scores obtained by laypersons (mostly college students) in previous research, we cannot conclude that police officers are actually better lie detectors than laypersons, because the latter were not included in this study. Had they been included as participants, it is possible that laypersons would have scored similarly to police officers. Unfortunately, inclusion of a group of laypersons was not possible, as (understandably) the police would not give us permission to show the highly sensitive stimulus material (fragments of real-life police interviews) to laypersons.

Second, findings showed a modest but significant relationship between experience in interviewing suspects and truth accuracy, with the more experience police officers reported in interviewing suspects (a self-report measure), the higher truth accuracy scores they obtained. This finding suggests that experience does make police officers better able to distinguish between truths and lies, a finding typically not found in deception studies with professionals as observers (DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991; Porter et al., 2000). We believe that this finding is affected by the way we measured experience. Other researchers use length of service/years of job experience as a measurement for experience (DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991; Porter et al., 2000). Such a measurement is unfortunate, as it says little about the officers' actual experience in situations in which they will attempt to detect deceit such as interviewing suspects. There is little reason to suggest that a police officer who had worked for many years in a managerial or administrative position within the police force would be a better lie detector than someone with a similar position outside the police force. Therefore, perhaps unsurprisingly, the present study also did not reveal significant correlations between length of service and accuracy. In other words, experience may benefit truth and lie detection only if the relevant experience is taken into account. Perhaps a weakness of our experience measure is that it is a self-report rather than an objective measure. It would be interesting to see whether an objective measure of experience in interviewing suspects (e.g., the number of suspect interviews a police officer has conducted) would correlate with accuracy as well. This would strengthen our argument. Unfortunately, the police do not record objective measures of experience with interviewing suspects.

The findings further revealed that men were better at detecting truths than women. We discuss this further below.

Theoretically, the higher than usual accuracy rates obtained in this study could be explained in several ways. First, as previously discussed in the introduction, the stakes for liars and truth tellers were higher in this study than in previous studies, and high-stakes lies were easier to detect than low-stakes lies. Second, the police officers were exposed to truths and lies told by the sort of people they are familiar with, namely police suspects, and familiarity with this group of people might have increased the accuracy rates. Third, police officers were exposed to truths and lies in a setting that is familiar to them, namely during police interviews, and familiarity with the setting might have increased accuracy rates. Probably all three factors contributed to the high accuracy rates found in this study. Therefore, these explanations have two theoretical implications. First, the obtained findings might well be situation and person specific and we therefore cannot guarantee that exposing police officers to high-stakes lies in situations that they are not familiar with (such as lies told by businesspeople in negotiations, by salespersons to clients, by politicians during interviews, or between romantic partners, etc.) would lead to similar accuracy rates as those found in this study. Similarly, we cannot guarantee that police officers will be any good at detecting low-stakes lies told by suspects. Second, to obtain insight into police officers' skills to detect deceit, exposing them to ecologically valid material (high-stakes lies told by suspects in police interviews) is crucial. This ecologically valid argument also applies to the measurement of relevant background variables, such as measuring police officers' experience with interviewing suspects.

Cues Used to Detect Deceit

The majority of police officers claimed that looking at gaze is a useful tool to detect deceit. This discovery was in agreement with previous findings (Akehurst et al., 1996; Vrij & Semin, 1996). On the one hand, this finding is surprising given that deception research has convincingly demonstrated that gaze behavior is not related to deception (DePaulo et al., 2003; Vrij, 2000a). Nor was gaze related to deception in the present stimulus material (Mann et al., 2002). On the other hand, this finding is not so surprising given that police manuals, including Inbau's manual, which is widely used, claim that suspects typically show gaze aversion when they lie (Gordon & Fleisher, 2002; Hess, 1997; Inbau et al., 1986, 2001). In other words, police officers are taught to look for these incorrect cues.

Several (modest) relationships occurred between cues mentioned by the officers as useful to detect deceit and their accuracy in truth and lie detection. First, good lie detectors mentioned story cues more often than poor lie detectors. Second, the more popular stereotypical belief cues participants mentioned (gaze, fidget, and self-manipulations), and the more they endorsed Inbau's view on cues to deception (liars show gaze aversion, display unnatural posture changes, exhibit self-manipulations, and place the hand over the mouth or eyes when speaking), the worse they became at distinguishing between truths and lies. In other words, looking at Inbau et al.'s (1986, 2001) cues is counterproductive. This is not surprising, as deception research has not supported Inbau's views (DePaulo et al., 2003; Vrij, 2000a). Female participants claimed to look more at Inbau cues than male participants, which might explain why female participants were poorer at detecting truths than male participants.

When we, by means of a regression analysis, compared the veracity judgments made by good and poor lie detectors with the behaviors actually shown by the suspects in the stimulus material (so-called cues to perceived deception), we found that poor lie detectors associated an increase in gaze aversion and an increase in head nods with deception. However, good lie detectors associated a decrease in illustrators with deception. Research has demonstrated that a decrease in illustrators is a much more valid cue to deception than gaze aversion or head nods (Ekman & Friesen, 1972; see DePaulo et al., 2003, and Vrij, 2000a, for reviews of such literature). The regression analysis further showed that poor lie detectors were guided by the gender of the suspect: Female suspects were considered less suspicious than male suspects. Obviously, such a generalized approach has nothing to do with sophisticated truth and lie detection.

Police officers were asked both prior to and after the lie detection task which cues they pay attention to in order to detect deceit. The results revealed that, with a few exceptions, the officers mentioned the same cues before and after the task. The exceptions are easy to explain. For example, officers mentioned physiological cues more often prior to the task. This is unsurprising, as such cues are difficult to notice when someone watches a videotape. Moreover, they mentioned looking for facts more often prior to the task than after. This is also unsurprising, as facts about the cases were not made available to the lie detectors in this study. The fact that a big overlap emerged between cues mentioned before and after the task has a theoretical implication. It suggests that the cues police officers rely on are more general rather than idiosyncratic. Moreover, these general views could then be used to predict police officers' lie detection ability in future situations. Our results support this idea. Mentioning popular stereotypical beliefs and mentioning Inbau's cues prior to the task was negatively correlated with accuracy.

Finally, apart from relying on different cues, the results revealed one further difference between poor and good lie detectors. For poor lie detectors, a significant negative correlation emerged between lie and truth accuracy, whereas such a significant correlation did not emerge for good lie detectors. This implies that for poor lie detectors, increased success at one aspect of the task (success at either lie detection or truth detection) hampers success at the other aspect of the task.

Accuracy–Confidence Relationship

Our analyses regarding the accuracy–confidence relationship revealed three major findings. First, as many researchers before us (see DePaulo et al., 1997, for a review), we did not find a significant relationship between accuracy and confidence. Even our alternative method of measuring confidence (measuring confidence after completing the whole lie detection task instead of after each veracity judgment) did not lead to any significant relationships. Second, participants were more confident when they were rating actual truths compared with when they were rating actual lies. This same effect has been found before (DePaulo et al., 1997), including in several recent studies (Anderson, Ansfield, & DePaulo,1999; Vrij & Baxter, 2000; Vrij et al., 2001). However, the reason for this is unclear. Possibly, when judges observe lies, there is something going on in the presentation that raises their doubts. Perhaps there is not enough to indicate the person as a liar, but enough to raise doubts about their subsequent judgment.

Most important, participants' estimated performance in the lie detection task (investigated after the task was completed) was significantly lower than their actual performance. This contradicts the overconfidence effect typically found in deception studies (DePaulo et al., 1997). Perhaps the overconfidence is an artifact. People are typically asked to express their confidence after each veracity judgment they make. One might argue that this is a very difficult task that could easily lead to overconfidence. Participants may believe that some veracity judgments they make during a lie detection task are correct. They then will probably give themselves confidence levels of above 50% for these judgments. For each judgment in which they are uncertain, they will probably give themselves a 50% chance of being correct, because why would they think that they have less than a 50% chance of being correct for each individual judgment? A confidence score above 50% is the likely result of this strategy.

Methodological Issues

Two methodological issues merit attention. First, police officers were exposed to an unbalanced number of truths and lies. This made it impossible to calculate a total accuracy score (accuracies of truths and lies combined) in this study, as that score cannot be unambiguously interpreted. For example, if an observer thinks that everyone was lying, that person would have a high total accuracy score in the event that he or she watched Tape 1 because that tape included nine lies and six truths. However, in this example, there would be no lie detecting ability, only a lie bias. We overcame this problem in two different ways, first by calculating truth and lie accuracy scores separately. The results showed that the difference between lie and truth accuracy was not significant, indicating that the sample as a whole did not show a truth or lie bias. We found that experience in interviewing was positively correlated with both truth accuracy and lie accuracy (although the latter correlation was only marginally significant). The fact that both correlations were positive indicates that experienced officers were most accurate and rules out the consideration that they were more biased. If they had a lie bias, then the experience–truth accuracy correlation would have been negative and vice versa; if they had a truth bias, then the experience–lie accuracy correlation would have been negative. The same reasoning applies to the other correlational findings. For example, mentioning Inbau et al.'s (1986, 2001) cues was negatively correlated with both truth and lie accuracy (although the latter correlation was not significant), hence, looking at those cues makes observers less accurate and not more biased. Moreover, we found that men were significantly better at detecting truths than women, whereas no significant gender difference emerged for detecting lies. Again, this demonstrates that men were more accurate at detecting truths and not more biased. In other analyses, in which the group of police officers were divided into two ability groups (poor lie detectors and good lie detectors), good lie detectors were those who scored both above the mean for lie clips (66.16%) and above the mean for truth clips (63.61%). This rules out that any of the good lie detectors could have been biased, as a lie bias would have resulted in a low truth accuracy score and a truth bias would have resulted in a low lie accuracy score.

Second, although the lie detection task was very realistic, it still differs in some aspects from real-life lie detection in police interviews. For example, normally the police officers would conduct the interview, and not just watch it. However, research has shown that conducting the interview is not necessarily advantageous in lie detection. Several researchers compared the accuracy scores of observers who actually interviewed potential liars with those who passively observed the interviews but did not actually interview the potential liars (Buller, Strzyzewski, & Hunsaker, 1991; Feeley & deTurck, 1998; Granhag & Strömwall, 2001). In all three studies, researchers found that passive observers were more accurate in detecting truths and lies than were interviewers. These findings suggest that merely observing is actually an advantage, not a disadvantage, in detecting deceit.

Moreover, ordinarily the police officer would see a much larger section, if not the whole interview(s), than they were exposed to in this experiment. Showing the whole interview would not have worked in this experiment, because without cutting out the majority of the interview, the footage would contain a huge amount of information that the experimenter could not be sure was true or false. Additionally, the experimenters were not asking participants to determine whether the suspect was guilty, as the truth–lie did not necessarily specifically relate to whether the suspect committed the crime under investigation, as mentioned earlier.

Also, in real life, officers may know some facts of the case. Although we could have provided our participants with the available evidence facts, we found this undesirable, as it would have made detecting some lies (those of which the suspect's statement contradicts the available evidence) too easy.

Finally, although participants on the whole were very willing to participate in the task, and keen to achieve high accuracy levels, this experiment does not have the same motivating consequences for them that judging the veracity of suspects in real life has. However, DePaulo, Anderson, and Cooper (1999) demonstrated that motivation does not improve performance in a lie detection task.

Conclusion

Police manuals typically give the impression that police officers who are experienced in interviewing suspects are good lie detectors (Inbau et al., 1986, 2001). Although previous research could not support this view whatsoever, our study, superior in terms of ecological validity over previous research, revealed that these claims are true to a limited extent. Police officers can detect truths and lies above the level of chance, and accuracy is related to experience with interviewing suspects. However, the results also revealed serious shortcomings in police work. First, accuracy rates, although above the level of chance, were far from perfect, and errors in truth–lie detection were frequently made. Second, police officers had a tendency to pay attention to cues that are not diagnostic cues to deceit, particularly body cues, such as gaze aversion. There may be various reasons why these nondiagnostic cues are so popular, one of which may be the discussion of these cues as diagnostic cues to deception in popular police manuals, such as the manual published by Inbau and colleagues. In fact, our research revealed that the more police officers followed their advice, the worse they were in their ability to distinguish between truths and lies.

References

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology, 10,* 461–471.

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgments in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 123–146). Mahwah, NJ: Erlbaum.

Anderson, D. E., Ansfield, M. E., & DePaulo, B. M. (1999). Love's best habit: Deception in the context of relationships. In P. Philippot, R. S. Feldman, & E. J. Coats (Eds.), *The social context of nonverbal behavior* (pp. 372–409). Cambridge, England: Cambridge University Press.

Anderson, D. E., DePaulo, B. M., Ansfield, M. E., Tickle, J. J., & Green, E. (1999). Beliefs about cues to deception: Mindless stereotypes or untapped wisdom? *Journal of Nonverbal Behaviour, 23,* 67–89.

Bond, C. F., & Atoum, A. O. (2000). International deception. *Personality and Social Psychology Bulletin, 26,* 385–395.

Buller, D. B., Strzyzewski, K. D., & Hunsaker, F. G. (1991). Interpersonal deception II: The inferiority of conversational participants as deception detectors. *Communication Monographs, 58,* 25–40.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report.* Hove, England: Psychology Press.

DePaulo, B. M., Anderson, D. E., & Cooper, H. (1999, October). *Explicit and implicit deception detection.* Paper presented at the Society of Experimental Social Psychologists, St. Louis, MO.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. L., & Muhlenbruck, L. (1997). *Personality and Social Psychology Review, 1,* 346–357.

DePaulo, B. M., Epstein, J. A., & Wyer, M. M. (1993). Sex differences in lying: How women and men deal with the dilemma of deceit. In M. Lewis & C. Saarni (Eds.), *Lying and deception in everyday life* (pp. 126–147). New York: Guilford Press.

DePaulo, B. M., Kirkendol, S. E., Tang, J., & O'Brien, T. P. (1988). The motivational impairment effect in the communication of deception: Replications and extensions. *Journal of Nonverbal Behavior, 12,* 177–201.

DePaulo, B. M., Lanier, K., & Davis, T. (1983). Detecting the deceit of the motivated liar. *Journal of Personality and Social Psychology, 45,* 1096–1103.

DePaulo, B. M., LeMay, C. S., & Epstein, J. A. (1991). Effects of importance of success and expectations for success on effectiveness at deceiving. *Personality and Social Psychology Bulletin, 17,* 14–24.

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

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

DePaulo, B. M., Stone, J. L., & Lassiter, G. D. (1985a). Deceiving and detecting deceit. In B. R. Schenkler (Ed.), *The self and social life* (pp. 323–370). New York: McGraw-Hill.

DePaulo, B. M., Stone, J. I., & Lassiter, G. D. (1985b). Telling ingratiating lies: Effects of target sex and target attractiveness on verbal and nonverbal deceptive success. *Journal of Personality and Social Psychology, 48,* 1191–1203.

Ekman, P., & Frank, M. G. (1993). Lies that fail. In M. Lewis & C. Saarni (Eds.), *Lying and deception in everyday life* (pp. 184–201). New York: Guilford Press.

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

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. G. (1999). A few can catch a liar. *Psychological Science, 10,* 263–266.

Feeley, T. H., & deTurck, M. A. (1998). The behavioral correlates of sanctioned and unsanctioned deceptive communication. *Journal of Nonverbal Behavior, 22,* 189–204.

Feeley, T. H., & Young, M. J. (2000). The effects of cognitive capacity on beliefs about deceptive communication. *Communication Quarterly, 48,* 101–119.

Forrest, J. A., & Feldman, R. S. (2000). Detecting deception and judge's involvement: Lower task involvement leads to better lie detection. *Personality and Social Psychology Bulletin, 26,* 118–125.

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.

Gordon, N. J., & Fleisher, W. L. (2002). *Effective interviewing and interrogation techniques.* San Diego, CA: Academic Press.

Granhag, P. A., & Strömwall, L. A. (2001). Detection deception based on repeated interrogations. *Legal and Criminological Psychology, 6,* 85–101.

Gudjonsson, G. H. (1994). Psychological vulnerability: Suspects at risk. In D. Morgan & G. M. Stephenson (Eds.), *Suspicion and silence: The right to silence in criminal investigations* (pp. 91–106). London: Blackstone.

Heinrich, C. A., & Borkenau, P. (1998). Deception and deception detection: The role of cross-modal inconsistency. *Journal of Personality, 66,* 687–712.

Hess, J. E. (1997). *Interviewing and interrogation for law enforcement.* Reading, England: Anderson Publishing Co.

Hurd, K., & Noller, P. (1988). Decoding deception: A look at the process. *Journal of Nonverbal Behavior, 12,* 217–233.

Inbau, F. E., Reid, J. E., & Buckley, J. P. (1986). *Criminal interrogation and confessions* (3rd ed.). Baltimore: Williams & Wilkins.

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2001). *Criminal interrogation and confessions* (4th ed.). Gaithersburg, MD: Aspen Publishers.

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.

Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements. Does it work? *Social Behaviour, 2,* 1–17.

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

Lane, J. D., & DePaulo, B. M. (1999). Completing Coyne's cycle: Dysphorics' ability to detect deception. *Journal of Research in Personality, 33,* 311–329.

Mann, S. (2001). *Suspects, lies and videotape: An investigation into telling and detecting lies in police/suspect interviews.* Unpublished doctoral dissertation, University of Portsmouth, Portsmouth, England.

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

Manstead, A. S. R., Wagner, H. L., & MacDonald, C. J. (1986). Deceptive and nondeceptive communications: Sending experience, modality, and individual abilities. *Journal of Nonverbal Behavior, 10,* 147–167.

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.

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

Moston, S., Stephenson, G. M., & Williamson, T. M. (1992). The effects of case characteristics on suspect behaviour during police questioning. *British Journal of Criminology, 32,* 23–40.

O'Sullivan, M., Ekman, P., & Friesen, W. V. (1988). The effect of comparisons on detecting deceit. *Journal of Nonverbal Behaviour, 12,* 203–216.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24,* 643–658.

Strömwall, L. A. (2001). *Detecting deception: Moderating factors and accuracy.* Unpublished doctoral dissertation, University of Gothenburg, Gothenburg, Sweden.

Vrij, A. (1993). Credibility judgments of detectives: The impact of nonverbal behavior, social skills, and physical characteristics on impression formation. *Journal of Social Psychology, 133,* 601–610.

Vrij, A. (1995). Behavioral correlates of deception in a simulated police interview. *Journal of Psychology, 129,* 15–28.

Vrij, A. (2000a). *Detecting lies and deceit: The psychology of lying and the implications for professional practice.* Chichester, England: Wiley.

Vrij, A. (2000b). Telling and detecting lies as a function of raising the stakes. In C. M. Breur, M. M. Kommer, J. F. Nijboer, & J. M. Reintjes (Eds.), *New trends in criminal investigation and evidence II* (pp. 699–709). Antwerpen, Belgium: Intersentia.

Vrij, A., & Baxter, M. (2000). Accuracy and confidence in detecting truths and lies in elaborations and denials: Truth bias, lie bias and individual differences. *Expert Evidence: The International Digest of Human Behaviour, Science and Law, 7,* 25–36.

Vrij, A., Edward, K., & Bull, R. (2001a). People's insight into their own behaviour and speech content while lying. *British Journal of Psychology, 92,* 373–389.

Vrij, A., Edward, K., & Bull, R. (2001b). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin, 27,* 899–909.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behaviour. *Journal of Nonverbal Behaviour, 24,* 239–263.

Vrij, A., & Graham, S. (1997). Individual differences between liars and the ability to detect lies. *Expert Evidence: The International Digest of Human Behaviour, Science and Law, 5,* 144–148.

Vrij, A., Harden, F., Terry, J., Edward, K., & Bull, R. (2001). The influence of personal characteristics, stakes and lie complexity on the accuracy and confidence to detect deceit. In R. Roesch, R. R. Corrado, & R. J. Dempster (Eds.), *Psychology in the courts: International advances in knowledge* (pp. 289–304). London: Routledge.

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

Vrij, A., & Mann, S. (2001b). Who killed my relative? Police officers' ability to detect real-life high-stake lies. *Psychology, Crime, & Law, 7,* 119–132.

Vrij, A., & Semin, G. R. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behaviour, 20,* 65–80.

Vrij, A., Semin, G. R., & Bull, R. (1996). Insight into behavior displayed during deception. *Human Communication Research, 22,* 544–562.

Vrij, A., & Winkel, F. W. (1991). Cultural patterns in Dutch and Surinam nonverbal behavior: An analysis of simulated police/citizen encounters. *Journal of Nonverbal Behavior, 15,* 169–184.

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale—Revised (*WAIS–R).* New York: Psychological Corporation.

Wiseman, R. (1995). The megalab truth test. *Nature, 373,* 391.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Volume 14* (pp. 1–57). New York: Academic Press.

(*Appendixes follow*)

Appendix A

Cue Categories, Descriptions and Frequency of Cues Mentioned Before and After the Task, and Number of Participants to Mention Each Cue

Cue	Group	Examples (including antonyms)	Before task	After task
Vagueness	Story	Vague reply/lots of detail	19	20
Contradictions	Story	Contradictions in story/consistent	18	10
Speech content	Story	Story content/specific words	9	15
Self-corrections	Story	Corrected self/corrected officer	0	7
Repetitions	Story	Repeating the question/buying time	3	1
Misc. speech	Story	Anything about speech that does not fit into “speech content,” e.g., pleading/minimizing offense or “uncertain replies”	10	22
Evidence	Story	Facts of the case	8	2
Hesitance/pauses	Vocal	Hesitation/pauses in speech/fluent speech	16	27
Voice	Vocal	Voice pitch/volume/harshness/soft	15	16
Stammering	Vocal	Stammered/stuttered	4	0
Speech fillers	Vocal	Lots of “ems” and “ahs”/no “ems”	2	1
Response length	Vocal	Lengthy reply/one-word reply	3	4
Gaze	Body	Averting gaze/eye contact	72	77
Movements	Body	Body language and movements	25	31
Posture	Body	Upright posture/slouched	6	13
Fidgeting	Body	Fidgeting/nervous movements/twiddling	19	11
Covering face	Body	Hands over face/hiding mouth	6	8
Hands	Body	Hand movements/still hands	9	28
Self-manipulation	Body	Touching/fiddling with self—excluding nails	7	6
Facial	Body	Facial expression/smiling/frowning	5	1
Props	Body	Playing with other things, e.g., cup/cigarette	3	2
Nail-biting	Body	Biting the nails/chewing fingers	2	2
Head movements	Body	Shaking/nodding/moving head	0	9
Physiological	Body	Sweating/blushing/blinking	15	5
Emotion	Body	Crying/upset/happy	6	1
Changes	Body	Changes in behavior/attitude	7	5
Demeanor	Conduct	Demeanor/relaxed/attitude	9	10
Defensive	Conduct	Sitting defensively/legs or arms crossed	12	9
Confidence	Conduct	Confidence/nervousness	11	9
Gut feeling	Other	Gut feeling/intuition	1	3
Total			322	355

Note. Misc. = miscellaneous.

Appendix B

Descriptions of the Coded Behaviors Displayed by the Suspects in the Stimulus Material and the Interrater Agreement Scores Between the Two Coders (Pearson Correlations)

1. Gaze aversion: number of seconds in which the participant looked away from the interviewer (two coders, *r* = .86).
2. Smiles: frequency of smiles and laughs (*r* = .98).
3. Blinking: frequency of eye blinks (*r* = .99).
4. Head nods: frequency of head nods for which each upward and downward movement was counted as a separate nod (*r* = .93).
5. Head shakes: frequency of head shakes. Similar to head nods, each sideways movement was counted as a separate shake (*r* = .98).
6. Other head movements: head movements that were not included as head shakes or head nods (e.g., tilting the head to the side, turning the face, etc.; *r* = .95).
7. Shrugs: frequency of where one or both shoulders is briefly raised in an “I don’t know” type gesture (*r* = .99).
8. Self-manipulations: frequency of scratching the head, wrists, etc. (touching the hands was counted as hand/finger movements rather than self-manipulations; *r* = .99).
9. Illustrators: frequency of arm and hand movements which were designed to modify and/or supplement what was being said verbally (*r* = .99).
10. Hand and/or finger movements: any other movements of the hands or fingers without moving the arms (*r* = .99).
11. Speech fillers: (speech fillers and speech errors were scored on the basis of a typed verbatim text) frequency of saying “ah” or “mmm,” etc., between words (*r* = .98).
12. Speech errors: frequency of word and/or sentence repetition, sentence change, sentence incompletion, stutters, etc. (*r* = .97). Deviations from the official English language (e.g., local dialects such as saying “it weren’t me” rather than “it wasn’t me”) were not included as speech errors.
13. Pauses: number of seconds in which there is a noticeable pause in the monologue of the participant (*r* = .55).

Received July 12, 2002
Revision received May 13, 2003
Accepted May 16, 2003 ■