Law Hum Behav (2009) 33:96–109
DOI 10.1007/s10979-008-9137-9

ORIGINAL ARTICLE

The Reliability of Lie Detection Performance

Amy-May Leach · R. C. L. Lindsay · Rachel Koehler · Jennifer L. Beaudry · Nicholas C. Bala · Kang Lee · Victoria Talwar

Published online: 2 July 2008
© American Psychology-Law Society/Division 41 of the American Psychological Association 2008

Abstract We examined whether individuals' ability to detect deception remained stable over time. In two sessions, held one week apart, university students viewed video clips of individuals and attempted to differentiate between the lie-tellers and truth-tellers. Overall, participants had difficulty detecting all types of deception. When viewing children answering yes–no questions about a transgression (*Experiments 1* and *5*), participants' performance was highly reliable. However, rating adults who provided truthful or fabricated accounts did not produce a significant alternate forms correlation (*Experiment 2*). This lack of reliability was not due to the types of deceivers (i.e., children versus adults) or interviews (i.e., closed-ended questions versus extended accounts) (*Experiment 3*). Finally, the type of deceptive scenario (naturalistic vs. experimentally-manipulated) could not account for differences in reliability (*Experiment 4*). Theoretical and legal implications are discussed.

Keywords Deception · Lie detection · Reliability · Individual differences

Individuals, such as customs officers and police officers, often are asked to determine the veracity of people's statements. Although performance in "intuitive" lie detection studies is usually near chance levels, some individuals score quite well whereas others do very poorly (e.g., Leach, Talwar, Lee, Bala, & Lindsay, 2004). This raises a question: Do people have an innate ability, or lack of ability, to detect lies? One criterion for considering behavior to be trait-like is reliable performance over time. However, it is unclear whether measured lie detection performance is stable or due to chance occurrences (e.g., a series of lucky or unlucky guesses). We examined the reliability of individuals' lie detection accuracy with an alternate forms paradigm. Further, we varied the type of deception and deceivers to determine the boundary conditions of reliable performance.

CUES TO DECEPTION

It might be expected that observers can reliably tell when someone is lying because there are cues to deception. Analyses of verbal cues in statements, using Criteria-Based Content Analysis and Reality Monitoring, identify deception well above the level of chance (e.g., Sporer, 1997; Vrij, 2005; Vrij & Mann, 2004). Nonverbal behaviors, which are more difficult to control than the content of speech, have also been used to correctly classify lie-tellers and truth-tellers at accuracy rates approaching 80% (Vrij, Edward, Roberts, & Bull, 2000). In fact, there are over a dozen reliable differences between lie-tellers and truth-tellers (DePaulo et al., 2003).

A.-M. Leach (✉)
Faculty of Criminology, Justice, and Policy Studies, University of Ontario Institute of Technology, Oshawa, ON, Canada L1H 7K4
e-mail: amy.leach@uoit.ca

R. C. L. Lindsay · R. Koehler · J. L. Beaudry · N. C. Bala
Queen's University, Kingston, ON, Canada

K. Lee
Institute of Child Study, University of Toronto, Toronto, ON, Canada

V. Talwar
McGill University, Montreal, QC, Canada

INTUITIVE LIE DETECTION ACCURACY

However, cues must be correctly perceived, interpreted, and employed by observers to be of any use in lie detection. Lay persons and police officers have many incorrect notions about the cues to deception (Akehurst, Kohnken, Vrij, & Bull, 1996). This might account for why the average lie detection accuracy for lay persons is 54% (Bond & DePaulo, 2006) and the majority of law enforcement groups do not perform much better (e.g., Bala, Ramakrishnan, Lindsay, & Lee, 2005; Ekman & O'Sullivan, 1991; Garrido, Masip, & Herrero, 2004; Kraut & Poe, 1980). Although a few groups (i.e., Secret Service agents, CIA agents, sheriffs, and forensic clinical psychologists) have scored above chance levels in some studies, they are the exception rather than the rule (Ekman & O'Sullivan, 1991; Ekman, O'Sullivan, & Frank, 1999). In general, law enforcement groups tend to overestimate their lie detection accuracy (Kassin et al., 2007) and express unwarranted confidence in their decisions (e.g., Leach et al., 2004).

Focusing on the performance of groups may obscure individual differences in lie detection ability. Lie detection performance exists on a continuum. For example, Vrij and Graham (1997) found that the accuracy of untrained individuals ranged from 20% to 70%. There are few explanations for why some individuals are more accurate than others. Attributes, such as self-monitoring, shyness, extraversion, and the "Big 5," are not linked to performance (Porter, Campbell, Stapleton, & Birt, 2002; Vrij & Baxter, 1999; Zuckerman, DePaulo, & Rosenthal, 1981). Although certain characteristics (e.g., heightened social anxiety, dysphoria, aphasia, left-handedness) are related to successful lie detection (DePaulo & Tang, 1994; Etcoff, Ekman, Magee, & Frank, 2000; Lane & DePaulo, 1999; Porter et al., 2002), it does not seem likely that these factors, alone, account for variations in accuracy within the general population.

THE RELIABILITY OF LIE DETECTION

An underlying assumption has been that lie detection is a measurable ability (O'Sullivan, 2007). Yet, researchers have not fully addressed the most fundamental requirement of a valid interpersonal difference: reliability (i.e., measured performance should be consistent over time). Usually, accuracy has been assessed within a single session (e.g., DePaulo & Pfeifer, 1986; Ekman et al., 1999). Although it is possible that the near-chance overall performance that has been observed in most studies masks the innate proficiency of a few individuals, and a counterbalancing lack of proficiency for others, it is also feasible that high and low performers were simply lucky and unlucky guessers. Findings might replicate only because performance is consistently near chance levels. If there is no evidence of reliability, then the notion that people have the ability to detect lies may be incorrect and fluctuations in data can be attributed to random guessing. In turn, it would be unclear as to what previous research has actually been measuring. Thus, establishing the reliability of performance is an important component in lie detection research.

O'Sullivan and Ekman's (2004) Wizards Project provides some evidence about the reliability of deception detection within a small segment of the population. After assessing over 10,000 individuals, these researchers found 14 people who achieved high levels of lie detection accuracy across three types of deceptive scenarios (i.e., featuring lies about emotions, opinions, and mock crimes). Apart from the Wizards Project, other researchers have discovered a few individuals who obtain accuracy rates over 80% across two different lie detection sessions (Bond, in press). Yet, both sets of researchers suggest that even highly proficient individuals' accuracy rates vary with the type of deception. It is unknown whether these fluctuations indicate that performance is not completely reliable or that some tasks are less successful at tapping into an underlying ability. There is some debate about whether the discovery of these highly proficient individuals should be attributed to chance (Bond & DePaulo, in press; Bond & Uysal, 2007), especially given the small number of trials in some cases, or a normal distribution of ability (O'Sullivan, 2007). Also, the majority of these "wizards" are law enforcement officials; there is little sense of the exact distribution of proficient and reliable lie detection performance within the general population.

Few researchers have examined the reliability of lie detection performance over time. Although some researchers have examined changes in accuracy and cue use across sessions, they have not directly analyzed the consistency of performance (Anderson, DePaulo, & Ansfield, 2002; Anderson, DePaulo, Ansfield, Tickle, & Green 1999; Granhag & Strömwall, 2001). Only Vrij, Mann, Robbins, and Robinson (2006) have explicitly tested the reliability of lie detection performance. Across 4 days, police officers attempted to detect the deception of actual suspects. Regardless of proficiency, their performance was not reliable. However, the same suspects were shown multiple times within each experimental session (i.e., at one point on the tape a particular suspect would be telling the truth and, at another, he would be lying). Although officers were instructed to ignore the repetitions, it is plausible that they were unable to do so. If they compared the suspects' behaviors over time, they could have altered their lie detection strategies. Thus, it might be the interference of a relative judgment strategy, rather than a lack of underlying ability, that led to the observed inconsistency in lie detection

performance. Moreover, it remains unknown whether this finding generalizes to other deceptive contexts.

TYPES OF DECEPTION

In addition, it is unknown whether these results generalize to all types of lies. The majority of lie detection studies involve experimentally-manipulated deception. That is, the targets are explicitly informed that they are participating in deception research and then they are instructed to lie or tell the truth about life events (e.g., Anderson et al., 2002; Porter et al., 2002), opinions (e.g., Ekman et al., 1999; Feldman, Jenkins, & Popoola, 1979), or observations (e.g., Ekman & O'Sullivan, 1991; Garrido et al., 2004). A minority of researchers have given individuals the opportunity to engage in mock crimes (e.g., theft) to more closely approximate situations that would be seen by law enforcement officials in the field (e.g., Frank & Ekman, 1997; Kassin & Fong, 1999; Vrij & Graham, 1997). However, even in these scenarios, participants know that they are part of a sanctioned experiment and there are few, if any, negative consequences for their behavior. It is not known whether experimentally-manipulated deception is comparable to more naturalistic scenarios, in which individuals lie of their own volition and believe that their lies might be successful. The handful of studies that used real-life examples of criminal behavior found that accuracy levels of observers are slightly higher than when experimental manipulations of lying are used (e.g., Davis, Markus, & Walters, 2006; Mann & Vrij, 2006; Mann, Vrij, & Bull, 2004; Vrij & Mann, 2001a, 2001b). How the type of deception affects the reliability of lie detection performance is unknown.

Also, the impact of the age of the deceiver on the consistency of lie detection performance is unclear. The large number of children serving as witnesses in the justice system has increased interest in the detection of children's deception (Bruck, Ceci, & Hembrooke, 1998). The few studies that have examined children's ability to lie have produced inconsistent results. In some studies, young children's deception was easier to detect than that of older individuals (Feldman et al., 1979; Feldman & White, 1980). However, more recent research, in which children and adults were instructed to lie or tell the truth about an event, revealed that the ability to deceive did not vary with age (Edelstein, Luten, Ekman, & Goodman, 2006; Vrij, Akehurst, Brown, & Mann, 2006). When children behave naturalistically (i.e., they volitionally lie or tell the truth), observers cannot detect their deception (Lewis, Stanger, & Sullivan, 1989). Replications and extensions of this research have shown that undergraduate students, police officers, customs officers, and parents have difficulty detecting children's lies (Leach et al., 2004; Talwar & Lee, 2002). Although the weight of the evidence suggests that both children and adults are capable of successful deception, it is not clear whether the stability of lie detection performance would remain unaffected by the age of the deceiver.

THE PRESENT RESEARCH

The purpose of the present experiments was to directly assess the reliability of lie detection performance over time across different types of deception and deceivers. First, individuals' accuracy was measured across two separate sessions. Accuracy was expected to be poor (i.e., near the level of chance). However, if researchers (e.g., O'Sullivan, 2007) are correct in asserting that lie detection is an *ability*, then performance should have been reliable over time. It is important to note that accuracy and reliability are independent factors. For example, a student who consistently scores 50% on true/false Introductory Psychology tests exhibits reliable (albeit poor, chance-level) performance. Thus, it is possible to be completely inaccurate, yet reliable.

Second, two different target populations, children and adults, were examined to determine whether the age of the deceiver affected the reliability of observers' lie detection performance. Children's deception could be difficult to detect because their behavior may not conform to that typically thought to be associated with deceivers. Young children are not as familiar with display rules, or heuristics that establish the appropriateness of behaviors in various contexts (Saarni & Von Salisch, 1993). For example, children's eyes might wander during a conversation because they are unaware of the nonverbal behavior (eye contact) that is most suitable in that situation. One initial consequence is that these children might be labeled lie-tellers because adults often list averted gaze as a marker of deception (Akehurst et al., 1996; Vrij & Semin, 1996). However, after viewing several children exhibiting this pattern, observers might recognize that it is unlikely that all of the children are lie-tellers. In turn, they could vacillate between attributing averted gaze to lying or more innocuous explanations (e.g., that the child is shy), leading to unreliable lie detection performance. Conversely, given that observers are more familiar with adult deception, they might be less likely to vary the cues that they employ when judging adults, making their performance more stable over time.

Third, the reliability of lie detection accuracy was assessed using two different types of deception. Naturalistic deception scenarios gave individuals the opportunity to commit a transgression (i.e., peek at a toy or cheat on a

test), but their behavior was voluntary and self-motivated. In order to produce experimentally-manipulated deception, individuals were instructed to lie or tell the truth about a life event (e.g., a visit to the hospital) or a transgression (i.e., cheating on a test). We posited that experimentally-manipulated deception could be more difficult to detect because deceivers might not experience the same levels of physiological arousal, or reveal the same cues, as "real" lie-tellers. If observers modify their cue use when viewing experimentally-manipulated deception, their lie detection performance could be unreliable. On the other hand, when faced with naturalistic deception, observers might be less confused, employ their regular strategies, and perform more consistently.

Finally, large numbers of deceivers and truth-tellers were included, regardless of their ability to deceive. This approach ensured that the examples of deception represented a wide range of believability (as it varies in the real world across good and poor lie-tellers *and* truth-tellers) and a high level of stimulus sampling (Wells & Windshitl, 1999).

EXPERIMENT 1

This study examined the reliability of the detection of children's naturalistic deception. Given previous research with the same stimuli (Leach et al., 2004; Talwar & Lee, 2002), performance was expected to be at chance levels. However, in keeping with the perspective that people have an underlying ability to detect deception, performance was predicted to be reliable.

Method

Participants

Undergraduate students ($N = 58$) participated in the study in exchange for course credit; individuals ($n = 7$) who did not attend both testing sessions were dropped from the study. Preliminary analyses failed to reveal any differences, in first session performance, between individuals who did and did not return—this was true for subsequent studies. Overall, 51 students (37 women and 14 men, M age = 18.82 years, $SD = 0.87$) completed the study.

Materials

Video Clips Video clips of a temptation resistance paradigm were obtained in a previous study (Talwar & Lee, 2002). Children's upper bodies and faces were recorded by a hidden video camera. A female experimenter played a guessing game individually with 3- to 11-year-olds. During this interaction, the experimenter was called out of the room. Prior to her departure, she asked the children not to peek at a hidden toy. Upon her return, she asked the children three questions: (1) "While I was gone, did you turn your head to the side?" (2) "Did you move around in your chair?" and (3) "Did you peek to see who it [the toy] was?" Reponses to these questions were completely spontaneous and produced three groups of children. Lie-tellers peeked at the toy and lied about it. Truth-tellers did not peek and truthfully denied having peeked. Finally, confessors peeked at the toy and admitted to the transgression.

In all, there were 80 video clips (M length = 17.50 s, $SD = 6.66$) produced from the recorded exchanges (featuring all three questions). Two videotapes, each consisting of 40 randomly assigned video clips, were made with the restriction that equal numbers of lie-tellers and truth-tellers were assigned to each tape. Each participant viewed both tapes, approximately one week apart, with the order of presentation counterbalanced.

Responses Participants were asked whether each child was lying or telling the truth using a forced-choice paradigm.

Procedure

A female experimenter randomly assigned participants to one of two groups (the only difference between the two groups was which of the two videotapes they viewed first). In each session, the entire procedure took approximately 45 min to complete. Approximately one week later, participants returned for the second session. The procedure was identical to that of *Session 1*, except that the previously unseen videotape (which featured different lie- and truth-tellers) was shown.

Results

Preliminary analyses in all studies revealed inconsistent effects of participant sex. As no meaningful conclusions could be drawn, data from all observers were combined for all subsequent analyses. As in previous research with this stimulus set (Leach et al., 2004), responses to confessors were not analyzed. These children could be easily classified based solely on their verbal reports (admissions) and their inclusion would have falsely inflated estimates of accuracy. These children's video clips were included as a control to assess participants' level of attention. Participants correctly identified most of these children as telling the truth ($M = .92$, $SD = .15$), suggesting that they were paying attention to the video clips. Eliminating participants who stated that one or more confessors were lying ($n = 23$) did not change any of the findings; therefore, data from all participants were included in the analyses. Only

participants' classifications of 32 lie-tellers and 38 truth-tellers (i.e., 16 lie-tellers and 19 truth-tellers during each session) were used in the analyses.

All responses were recoded to determine accuracy. Each correct decision was awarded a "1", and each incorrect decision was given a "0". These scores were averaged across children (or targets) to yield the overall proportion of accurate decisions for each participant. Analyses were conducted on the mean proportions (possible maximum score = 1.00; minimum score = 0). All effect sizes were reported using meta-analytic *r*, as recommended by Rosenthal (1991). These approaches were employed in all subsequent experiments.

Across all studies, participants' scores were at, or very near, the level of chance. As the primary purpose of the project was to examine reliability, rather than average accuracy scores, only analyses related to reliability will be discussed. Accuracy, discrimination (d') and bias (β) values for each experiment are reported in Table 1.

Lie Detection Reliability

The range in accuracy appeared similar in session one (17–71% correct) and session two (14–86% correct). There was a significant correlation between accuracy in the first and second sessions, $r(50) = 0.67$, $p < .001$.

One possibility is that this finding was due to a few outliers (i.e., individuals who detected deception significantly above or below the level of chance on both occasions). This alternative explanation was examined by excluding these individuals using a binomial theorem analysis (to determine individual performance differing from chance) that is described below. Even when excluding these individuals, there was still a significant correlation between accuracy in the first and second sessions, $r(36) = 0.46$, $p < .01$.

In addition, accuracy reflects two separate dimensions of the participants' decision-making process: (1) discrimination between truth- and lie-tellers (often quantified as d'), and (2) bias (i.e., the tendency to favor a particular response, such as categorizing children as lying, often quantified as β). Increasingly, researchers have underscored the importance of signal detection theory for analyses of lie detection (Meissner & Kassin, 2002). In order to extract this information, additional analyses were conducted using signal detection theory. Discrimination between truth-tellers and lie-tellers in the first and second sessions was significantly correlated, $r(50) = 0.65$, $p < .001$. Also, there was a significant correlation between bias in both sessions, $r(50) = 0.38$, $p < .01$.

Binomial Theorem Analysis

For descriptive purposes, each individual's performance was compared to chance using a binomial theorem analysis. Assuming that people with no ability to detect deception are guessing, the probability of an individual correct decision is .50. It is possible to calculate the minimum number of correct or incorrect decisions required for

Table 1 Participants' average accuracy, discrimination, and bias scores

	Lie-tellers		Truth-tellers		d'		β	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experiments 1 and 5								
Time 1	0.47^{a}	0.12	0.49^{a}	0.13	-0.12^{a}	0.58	1.02^{a}	0.18
Time 2	0.52^{b}	0.14	0.54^{b}	0.16	0.18^{b}	0.67	1.06^{a}	0.30
Experiment 2								
Time 1	0.41^{a}	0.14	0.59^{b}	0.15	0.02^{a}	0.55	1.06^{a}	0.32
Time 2	0.38^{a}	0.15	0.64^{c}	0.16	0.08^{a}	0.61	1.11^{a}	0.18
Experiment 3								
Questions	0.42^{a}	0.15	0.47^{b}	0.16	-0.45^{a}	0.59	0.99^{a}	0.24
Narratives	0.37^{c}	0.15	0.48^{b}	0.16	-0.31^{a}	0.65	1.09^{a}	1.42
Experiment 4								
Time 1								
Experimental	0.54^{ab}	0.20	0.49^{ac}	0.19	0.06^{a}	0.90	1.25^{a}	1.47
Naturalistic	0.37^{d}	0.16	0.64^{e}	0.14	-0.01^{b}	0.62	1.01^{a}	0.31
Time 2								
Experimental	0.49^{ac}	0.21	0.60^{be}	0.18	0.25^{a}	0.80	1.35^{a}	2.02
Naturalistic	0.45^{cd}	0.20	0.54^{abc}	0.19	-0.10^{b}	0.72	1.04^{a}	0.30

For each experiment, only scores with a different superscript differ significantly from each other ($p < .01$)

the performance of the individual participant to be significantly different ($p < .05$) from chance. Performance was significantly above chance if 23 or more of the participant's decisions were correct and significantly below chance if 12 or less of the participant's decisions were correct. This analysis allows for a sense of the distribution of reliability (i.e., whether participants who performed below, above, or at chance were most likely to display stable performance). The performance of the majority of participants ($n = 37$, or 72.5% of the sample) was not significantly different from chance during both sessions. Very few participants performed consistently above ($n = 2$; 4.0%) or below chance ($n = 2$; 4.0%) (see Table 2). The remaining seven participants (13.7%) were different from chance in one, but not both, of the sessions.

Discussion

Participants' performance was stable over time, such that a high correlation existed between accuracy during the first and second sessions. Yet, observers' consistency was not a product of accuracy. In fact only 2 of 51 participants (4%), performed significantly better than chance during both sessions.

Although the observed consistency in lie detection performance is interesting, the type of deception tested is not representative of what is seen in the justice system. For example, it is unlikely that law enforcement officials make decisions based solely on children's yes–no responses. In fact, customs officers are explicitly trained to avoid asking these types of closed-ended questions (Canada Customs and Revenue Agency, 1998). Thus, officials in the justice system may be more accustomed to detecting deception within adults' narratives (e.g., alibis). It is unknown whether the present findings may be attributed to generalized or specific ability. That is, participants could be reliable when detecting lie-tellers and truth-tellers of all ages or only children. As noted, there is some debate about whether children's and adults' deception is similar (e.g., Feldman et al., 1979; Leach et al., 2004). Moreover, the stimuli used in *Experiment 1* featured very brief responses to direct questions. The results may not generalize to the detection of deception in lengthy, open-ended accounts of events. Researchers suggest that open-ended accounts provide more material upon which to make a decision (Vrij & Baxter, 1999). Yes–no responses may be so brief that observers are limited in the numbers of cues that they can use (and, in turn, there is little variability in their performance over time). Conversely with so much information to process, individuals could have more difficulty maintaining a particular strategy. *Experiment* 2 was conducted to test the generalizability of the findings in *Experiment 1* to stimuli that more closely approximate interests in the legal system (i.e., adults' deceptive narratives).

Table 2 Distribution of participants' accuracy during session one and session two in Experiments 1–4

Session two	Session one		
	Below chance	At chance	Above chance
Experiment 1			
Below chance	2	1	0
At chance	4	37	1
Above chance	0	4	2
Experiment 2			
Below chance	0	2	0
At chance	2	43	1
Above chance	0	0	0
Experiment 3			
Below chance	10	29	1
At chance	22	115	11
Above chance	3	6	0
Experiment 4			
Below chance	0	1	0
At chance	0	32	1
Above chance	0	1	0

EXPERIMENT 2

Method

Participants

Undergraduate students ($N = 57$) participated in the study in exchange for course credit, with nine participants excluded from analyses because they did not complete both testing sessions. Overall, 48 students (32 women and 16 men, M age = 18.81 years, $SD = 0.96$) completed the study.

Materials

Video Clips Video clips were prepared of adults giving an account of an event that had actually occurred in their life (e.g., a car accident) and an event that had never occurred (e.g., breaking a limb). Pairs of same-sex adults told the same stories, such that the true narrative for one individual within the pair was the false narrative for the other individual. The storytellers' upper bodies and faces were clearly visible throughout the entire procedure.

In all, 96 video clips were produced (M length = 98.13 s, $SD = 34.30$). Four subsets were created, with 24 clips (featuring 12 men and 12 women) randomly assigned to each set. Within each subset, half of the individuals

discussed real events (i.e., told the truth), whereas half discussed events that had never occurred (i.e., lied). Each participant viewed two subsets, with the restriction that they could not see the same storyteller twice. A computer program showed the stimuli randomly, such that no two participants observed the same order of presentation. In addition, the subset order was counterbalanced so that each subset was presented in the first and second session an equal number of times.

Ratings As the stimuli were longer and more complex than in the previous experiment, participants were asked whether each adult was lying or telling the truth about a particular event (e.g., car accident). That is, the participant was to indicate if the primary theme of the story was honestly portrayed (e.g., the individual actually had been in a car accident).

Procedure

The procedure was similar to that of *Experiment 1* except that the experiment was conducted on a computer and featured the adult narratives as stimuli.

Results

Lie Detection Reliability

The range in accuracy appeared similar in session one (29–75% correct) and session two (29–67% correct). As in *Experiment 1*, a correlation analysis was performed to determine if participants' performance was stable over time. There was no significant relationship between accuracy in sessions one and two, $r(47) = -0.02$, $p = .91$. Further, correlation analyses failed to reveal significant relationships between participants' discrimination (d'), $r(47) = -0.07$, $p = .52$, or bias (β), $r(47) = -0.26$, $p = .07$, in sessions one and two.

Binomial Theorem Analysis

Performance was significantly above chance if 17 or more of the participant's decisions were correct and significantly below chance if 7 or less of the participant's decisions were correct. The majority of participants ($n = 43$; 90.0%) performed at chance (i.e., scored between 8 and 16) during both sessions (see Table 2).

Discussion

Overall, lie detection performance was not reliable. These results differed dramatically from those of *Experiment 1*, which was surprising given the consistency in performance observed in the first experiment. There are several possible reasons why lie detection performance was not reliable. First, the findings of either one of the two experiments could have been the result of random error. Second, the deception stimuli varied across the two experiments. In *Experiment 1*, observers rated children and in *Experiment 2* observers rated adults. The few studies that have examined deception across the lifespan have produced widely different results: one set of findings indicated that deception improves with age (e.g., Feldman & White, 1980), whereas the other suggested that there are no developmental differences (e.g., Talwar & Lee, 2002). The lack of replication between the present experiments might suggest that individuals may not be as reliable when detecting deceptive adults as when detecting deceptive children. Third, the experiments also featured different types of interviews. Specifically, the clips in *Experiment 1* focused on yes–no responses; the clips in *Experiment 2* involved full narratives. Vrij and Baxter (1999) maintain that observers assess elaborations and denials differently due to the information afforded by each type of interview. It is possible that the reliability of performance was adversely affected by the presentation of narratives. In sum, it is unknown whether the lack of replication was due to the different populations or scenarios. *Experiment 3* was conducted to address this issue further.

EXPERIMENT 3

Method

Participants

Undergraduate students ($N = 230$) participated in this study in exchange for course credit; 33 individuals were not included in the analyses because of attrition ($n = 25$) or technical difficulties ($n = 8$). Overall, 197 students (164 women and 33 men, M age = 18.70 years, $SD = 2.99$) completed the study.

Materials

Video Clips Four experimental conditions were created by showing video clips of adults or children either giving lengthy narratives (narrative conditions) or answering yes–no questions (question conditions). The "Adult Narrative" condition contained the same video footage as in *Experiment 2*. At the end of the original discussion (about a true or fabricated event), each storyteller was asked three questions: (1) "Did this really happen?" (2) "Did you make up this story?" and (3) "Did someone tell you to make up this story?" Footage from this closed-ended exchange was shown to participants in the "Adult Question" condition. This condition was designed to examine

the reliability of the detection of adults' deceptive and truthful yes–no responses.

The final two conditions featured 4- to 7-year-old children telling narratives about real or fabricated events (for details see Talwar, Lee, Bala & Lindsay, 2006). Due to the complexity of the task, each child was only asked to discuss one event (either real or fabricated). In addition, due to children's reluctance to provide complete accounts (Ceci & Bruck, 1995), an experimenter prompted each child to provide more information about the event (e.g., by asking, "What else did you do?"). Video footage of the children discussing the events was compiled for the "Child Narrative" condition. As with the adults, children were matched such that one child's true narrative was another child's false narrative. In addition, children were asked three questions about the event: (1) "Did this really happen?" (2) "Did you make up this story?" and (3) "Did you and your mom [or someone else] make up this story?" The children's answers to these three questions were used in the "Child Question" condition.

Ninety-six clips were produced for the "Adult Narrative" (M length = 98.13 s, SD = 34.30) and "Adult Question" (M length = 9.54 s, SD = 4.31) conditions. Again, four subsets were created, with 24 clips (featuring 12 men and 12 women) randomly assigned to each set. Within each subset, half of the individuals told the truth (and half of the individuals lied). Due to the lower number of children's clips (as each child only discussed one event), two subsets of 24 clips were compiled for the "Child Narrative" (M length = 85.17 s, SD = 34.36) and "Child Question" (M length = 15.90 s, SD = 5.19) conditions. Within each set, six children (three boys and three girls) represented each age group, with half lying and half telling the truth.

Each participant was assigned to only one condition and viewed two different subsets of similar stimuli. Participants never saw the same target twice. Again, a computer program presented the stimuli randomly and the subset order was counterbalanced.

Ratings Participants were asked to determine whether each child or adult was lying or telling the truth using a forced-choice paradigm.

Procedure

The procedure was identical to that of *Experiment 2*.

Results

Lie Detection Reliability

The range in accuracy appeared similar in session one (0–79% correct) and session two (8–79% correct). A correlation analysis did not reveal a significant relationship between accuracy in sessions one and two, $r(196) = 0.00$, $p = .99$. Specifically, correlation analyses failed to reveal a significant relationship when participants viewed "Child Narrative," $r(48) = 0.11$, $p = .47$, "Child Question," $r(48) = -0.08$, $p = .59$, "Adult Narrative," $r(48) = -0.02$, $p = .90$, and "Adult Question," $r(49) = -0.00$, $p = .98$, clips.

Correlation analyses of participants' discrimination abilities in sessions one and two failed to reveal a significant relationship in the "Child Narrative," $r(48) = 0.02$, $p = .91$, "Child Question," $r(48) = -0.13$, $p = .39$, "Adult Narrative," $r(48) = 0.04$, $p = .81$, and "Adult Question," $r(49) = 0.01$, $p = 1.00$, conditions. The same was true for bias in the "Child Narrative," $r(48) = -0.08$, $p = .58$, "Child Question," $r(48) = -0.12$, $p = .41$, and "Adult Narrative," $r(48) = -0.04$, $p = .80$ conditions. However, there was a significant relationship between bias in sessions one and two for "Adult Question" clips, $r(49) = 0.51$, $p < .001$.

Binomial Theorem Analysis

Each individual's performance was compared to chance using a binomial theorem analysis. Performance was significantly above chance if 17 or more of the participant's decisions were correct and significantly below chance if 7 or fewer decisions were correct. The majority of participants (n = 115; 58.4%) performed at chance (i.e., scoring between 8/24 and 16/24) during both sessions, 10 (5.1%) performed below chance on both sessions, and none performed above chance in both sessions (see Table 2).

Discussion

Regardless of the age of the deceiver or the type of interview, lie detection performance was not reliable. Although this experiment was designed to account for the differences in reliability between *Experiments 1* and *2*, it did not. Lie detection was not stable when varying the age of the deceiver, nor the type of interview. It is possible that the observed reliability in *Experiment 1* was simply due to random error. Two subsequent experiments have suggested that performance is not reliable. In addition, the present experiment featured children's yes–no responses (i.e., stimuli that were similar to *Experiment 1*) and the reliability effect could not be replicated.

However, there is another, untested, difference between the stimuli used in *Experiment 1* and *Experiments 2* and *3*. Specifically, the first experiment involved naturalistic deception (i.e., lie-tellers and truth-tellers did not know that they were involved in a deception task and their responses were volitional). Both of the other experiments featured

experimentally-manipulated deception (i.e., lie-tellers and truth-tellers were explicitly told that they were involved in a deception task and they were instructed to respond in a particular way). The detection of naturalistic deception may be easier than the detection of experimentally-manipulated deception because it more closely approximates what individuals encounter daily and lie-tellers might have more difficulty concealing their deception due to motivational constraints (e.g., DePaulo & Kirkendol, 1989; Mann et al., 2004). It is reasonable to expect that there might be corresponding effects on reliability. *Experiment 4* explored this issue.

EXPERIMENT 4

Method

Participants

Undergraduate students ($N = 42$) participated in this study; seven students were not included in the analyses because they were not present at both testing sessions. Overall, 35 students (28 women and 7 men, M age = 20.49 years, $SD = 2.03$) completed the study.

Materials

Video Clips Video clips were obtained using a temptation resistance paradigm. Adults were recorded by a hidden video camera. Individually, each participant was escorted into a quiet room. They were seated at a table containing a ticking clock and a pile of marked anagram tests. A female experimenter asked each person to complete a timed anagram test—the same version of the tests that were left out in plain view. The 20 anagrams ranged in difficulty from common words (e.g., sugar) to those that were more obscure (e.g., jocose). The experimenter provided one of two incentives for high performance. Some individuals were told that there was a monetary reward ($50) for the highest score; others were informed that if they solved 75% of the anagrams in the time allotted, they could leave early and that most of their peers had achieved this standard (actually, no one achieved this standard without cheating). Individuals were told that they had ten minutes to complete the exam and that they would be left alone during that time. As the experimenter left the room, she instructed participants to continually refer to the clock (beside the marked tests) and to take a pencil from the box on the table (which was on top of the pile of marked tests). These manipulations were used to draw the person's attention to the tests. When the experimenter returned, she 'noticed' the tests that had been left out and asked the participants two questions: (1) "While I was gone, did you look at the tests?" (2) "Did you cheat on this test?"

There were two versions of this procedure. In the naturalistic condition, the procedure was exactly as described above. Participants were unaware that they were part of a lie detection study: their actions and responses to the two critical questions were completely spontaneous and produced three results. Naturalistic lie-tellers cheated on the test and lied about it. Naturalistic truth-tellers did not cheat and truthfully denied having cheated. Finally, confessors cheated on the test and admitted to the transgression. The latter were excluded from this study because the veracity of their statements could easily be deduced (i.e., it is rare for an individual to falsely confess to a transgression of this sort). In order to produce experimentally-manipulated scenarios, the same procedure was used. However, prior to beginning the anagram task, another experimenter spoke with the participant. This experimenter instructed the participant to perform a certain way (i.e., either cheat on the test or not) and to deny having cheated. Participants in the experimentally-manipulated condition were yoked to participants in the naturalistic condition on the basis of sex and statement type (i.e., lie vs. truth). For example, if a man lied in the naturalistic condition, a man was enlisted to lie in the experimentally-manipulated condition. The experimenter who asked the questions was blind to condition and did not know if the participant had cheated or not.

In all, there were 56 video clips used in *Experiment 4*: 28 involved naturalistic scenarios (M length = 7.44 s, $SD = 1.97$) and 28 featured experimentally-manipulated scenarios (M length = 8.91 s, $SD = 2.72$). Each clip contained the two critical questions and the participants' responses. The 28 clips were randomly assigned to each of the two testing sessions, such that equal numbers of truth-tellers and lie-tellers, men and women, and naturalistic and experimentally-manipulated scenarios appeared in each session.

Ratings Participants were asked to determine whether each adult was lying or telling the truth using a forced-choice paradigm.

Procedure

The procedure was identical to the previous experiments, except that the study was conducted in a classroom setting using the adult cheating paradigm clips.

Results

A review of the data indicated that several participants knew at least one of the targets. As there might be potential biases associated with familiarity, known targets were not included in any of the analyses. It should be noted that this

was a very rare occurrence—only 1.3% (26/1960) of the responses involved known targets. In addition, analyses revealed that the inclusion of these targets did not alter any of the findings.

Lie Detection Reliability

The range in accuracy for naturalistic clips appeared similar in session one (31–71% correct) and session two (21–79% correct). This was also true for experimentally-manipulated scenarios in sessions one (29–92% correct) and two (29–86% correct)

A correlation analysis was performed to determine whether participants' performance was stable over time. Overall, there was no significant relationship between accuracy in sessions one and two, $r(34) = -0.04, p = .82$. Specifically, correlation analyses failed to reveal a significant relationship between accuracy in sessions one and two when participants viewed naturalistic, $r(34) = -0.11$, $p = .53$, or experimentally-manipulated, $r(34) = -0.17, p = .34$, clips.

Additional correlation analyses failed to reveal a significant relationship between discrimination in sessions one and two when participants viewed naturalistic, $r(34) = -0.10$, $p = .56$, or experimentally-manipulated, $r(34) = -0.20, p = .25$, scenarios. Although there was a significant correlation between bias in sessions one and two for naturalistic, $r(34) = -0.39, p < .05$ clips, there was not when participant viewed experimentally-manipulated deception, $r(34) = -0.01, p = .94$.

Binomial Theorem Analysis

Again, for descriptive purposes, each individual's performance was compared to chance using a binomial theorem analysis. For each scenario, performance was significantly above chance if 11 or more of the participant's decisions were correct and significantly below chance if 3 or less of the participant's decisions were correct. The majority of participants ($n = 32$; 91%) performed at chance (i.e., scored between 4 and 10 out of 14) during both sessions (see Table 2).

Discussion

Regardless of the type of scenario (naturalistic or experimentally-manipulated), lie detection performance was not reliable.

EXPERIMENT 5

Although reliable performance was observed in *Experiment 1*, the finding did not generalize to other lie detection contexts (*Experiments 2–4*). This raises the possibility that performance is not, in fact, reliable and that the original finding was merely an artifact. In order to test this possibility, we conducted a direct replication of *Experiment 1*.

Method

Participants and Procedure

University students ($N = 24$) participated in this study in exchange for course credit, with fifteen students (13 women and 2 men, M age = 25.49 years, $SD = 8.65$) completing both sessions and, thus, included in the analyses. The participants were tested, as a group, in a classroom. All other aspects of this study were an exact replication of the first experiment.

Results and Discussion

As in *Experiment 1*, there was a substantial and significant relationship between accuracy of lie detection decisions in sessions one and two, $r(14) = 0.56, p < .05$. This was true even when outliers (i.e., individuals who performed consistently above and below chance) were excluded, $r(9) = 0.63$, $p = .05$, or only the discrimination between truth-tellers and lie-tellers was considered, $r(14) = 0.57$, $p < .05$. Although substantial, the correlation for consistency in bias over time was not significant; however, this is likely to be a power limitation given the small sample size, $r(14) = 0.48$, $p = .07$. Overall, there is additional evidence that the original finding—that observers' lie detection performance is reliable—is robust.

GENERAL DISCUSSION

We examined the stability of university students' lie detection performance. Observers attempted to differentiate between lie-tellers and truth-tellers during two separate sessions. Overall, as expected, people's ability to detect deception was poor (i.e., accuracy hovered around chance levels). Lie detection performance did not appear to be reliable in the majority of experiments. However, accuracy was stable when individuals classified children denying having committed a transgression (*Experiment 1* and *Experiment 5*).

Performance was not reliable when participants viewed adults' experimentally-manipulated narratives (*Experiment 2*). *Experiment 3* ruled out the possibility that the difference in reliability was due to the type of interview (yes-no responses vs. narratives) or the age of the deceiver (adults vs. children). Finally, *Experiment 4* indicated that the type

of scenario (i.e., experimentally-manipulated or naturalistic deception) did not account for differences in reliability.

There are two explanations for these conflicting findings. First, performance may only be reliable under a highly restricted set of conditions and the boundaries of reliability are still unknown. *Experiments 1* and *5* suggest that children's yes–no responses to questions about resisting temptation elicit stable performance. *Experiments 2–4* provide a wider range of situations that do not produce reliable lie detection performance. Future research should explore other circumstances under which performance is reliable. Regardless, the implications of this explanation must be considered. The most forensically relevant variables (specifically, adults' and children's narratives) produced unreliable performance. Although it is difficult to elicit complete accounts of events from children (e.g., Ceci & Bruck, 1995), there are significant problems with obtaining accurate responses to yes–no questions (e.g., Fritzley & Lee, 2003). If performance is only reliable under specific conditions (i.e., when children answer yes–no questions) that are unfavorable and rarely encountered by in the justice system, albeit empirically interesting, the applications are quite limited.

A second possibility is that the reliable performance observed in *Experiments 1* and *5* was simply a fluke. This notion is difficult to justify due to the robustness of the findings in those experiments. The combined probability of these two results is less than .00005. Yet, the three other experiments do provide evidence that, on the whole, performance is not reliable. If performance is not stable—under any circumstances—there are significant legal and theoretical implications. For example, the notion that law enforcement officials can be hired or promoted based on their competence at lie detection, with the assumption that ability will remain constant, would be completely incorrect. More importantly, a lack of reliability might affect the value of training. There is evidence that lie detection experience does not always improve performance (e.g., DePaulo & Pfeifer, 1986; Leach et al., 2004) and attempts to train individuals have led to few improvements in accuracy (Kassin & Fong, 1999; Köhnken, 1987). These findings may be explained by the lack of a stable ability to detect deception. Thus, claims that training programs dramatically improve lie detection performance (e.g., Inbau, Reid, Buckley, & Jayne, 2001) could be unwarranted because they assume that decision-making, in the context of lie detection, is stable. Interestingly, if anything, training programs might only improve the reliability of performance by encouraging consistent cue use.

In addition, researchers must be conscious of the potential repercussions of unreliable performance. The majority of research has relied upon single-session observations (e.g., Ekman et al., 1999), possibly with the implicit assumption that lie detection performance is stable (i.e., that if these same individuals were tested at a later date, accuracy would be comparable). However, the present experiments suggest otherwise. If performance is not reliable, should previous findings based upon this assumption be negated? Although it would be unwise to suggest that all single-observation research is flawed, it might be sensible to interpret those findings with caution.

Another concern raised by this research is the significant variability in accuracy that was elicited by different types of lie detection stimuli (see Table 1). Generally, the present experiments support previous observations that most individuals have difficulty detecting deception (e.g., DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991). In keeping with other research (e.g., Lewis et al., 1989; Vrij et al., 2006), the detection of children's and adults' deception was equally poor. However, this does not mean that all of the deception scenarios produced the same patterns of performance. For example, although researchers may argue that naturalistic and experimentally-manipulated deception are comparable because they both result in chance-level performance (e.g., Leach et al., 2004), the present experiments suggest otherwise. In *Experiment 4*, individuals were more accurate when viewing experimentally-manipulated lie-telling and naturalistic truth-telling. Perhaps people who were instructed to lie or tell the truth were "overselling" (i.e., their motivation to convince viewers led their lies to be more obvious and their truth-telling to appear more sincere). This finding suggests that, when people are instructed to act in a certain way, the end results are not necessarily comparable to naturalistic behaviors. Yet, the majority of researchers base their conclusions on experimentally-manipulated deception (e.g., Ekman et al., 1999; Feldman et al., 1979). Finally, the type of interview (closed-ended vs. open-ended) also led to different levels of accuracy. As in previous research, individuals more accurately identified truth-telling in narratives (Vrij & Baxter, 1999). In a departure from previous findings, observers in the present experiment were equally accurate when viewing truthful and deceptive yes–no responses (whereas other researchers found higher accuracy for detecting lies under these conditions). Regardless, there is independent empirical support for differences in the classifications of narratives and closed-ended responses. In sum, the stimuli produced different findings and raise questions about what researchers are, in fact, studying.

A related issue is the manner in which deceivers are included in experiments. Some researchers specifically select individuals whose deception is readily detectable (e.g., Ekman & O'Sullivan, 1991). In the present experiments, all available stimuli were included to provide a representative range of lie-tellers and truth-tellers. This approach is more inclusive, but it might raise concerns

about the exact distribution of good/poor lie-tellers and truth-tellers. For descriptive purposes, binomial theorem analyses were conducted to examine whether targets were classified below, above, or at chance levels: in general, the stimuli were correctly identified at chance levels. Despite the underreporting of this type of information in studies of lie detection, it might provide an important clue to performance. Above chance performance may reflect the degree to which the stimuli are easily classified as much as, or more than, the detection abilities of the participants. In our experiments, the stimuli were difficult to detect and performance was near chance—whether analyzed in terms of the nature of the stimuli or the lie detectors. Close examination of how the inclusion of particular targets affects lie detection could be beneficial in future research.

The present research raises two major questions: (1) Is performance reliable? (2) Why is performance not more reliable (or reliable across more situations)? Considerable progress was made in addressing the first issue. Although the boundary conditions remain unclear, it seems likely that performance is unreliable under most circumstances, but may be reliable within some very narrow constraints. However, the second question remains unanswered. If performance is reliable, it may be due to cue use. That is, individuals may employ the same lie detection strategies over time. If appropriate cues are used, then accuracy should be above chance; poor cues should lead to below-chance or random performance. It is known that laypersons and law enforcement officials do have beliefs about the behaviors that are indicative of deception, even though they may not be reliable (Akehurst et al., 1996; Vrij, Edward, & Bull, 2001; Vrij & Semin, 1996). It might seem unlikely that these strategies would change drastically over a week (the time between testing sessions), suggesting that performance relying on specific cues should remain stable. The lack of reliable performance observed in the majority of the present experiments does not necessarily indicate that cue use was not stable. Individuals may have employed *irrelevant* cues consistently, producing the chance-level accuracy and random variations in performance that has been widely observed. Also, it is possible that individuals wished to employ certain cues consistently, but were unable to do so because they were not featured in the stimulus set. One avenue of research is to discover the cues that are being used (un)systematically by successful and unsuccessful lie detectors. If successful cues to intuitive lie detection can be clearly identified, the possibility exists that others could be trained to become better lie detectors (O'Sullivan, 2005).

The extent to which intuitive lie detection is a stable characteristic of the person is still unclear. The fact that the majority of researchers have worked under the assumption that performance is reliable does not mean that an empirical demonstration was not required. In fact, the present experiments suggest that this intuition has been largely incorrect. Future research is urgently needed to determine the circumstances—if any—under which lie detection performance is reliable.

Acknowledgements This research was supported by grants to the authors (Leach, Lindsay, Bala, & Lee) from the Social Sciences and Humanities Research Council of Canada.

REFERENCES

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behavior. *Applied Cognitive Psychology*, *10*, 461–471. doi:10.1002/(SICI)1099-0720(199612)10:6<461::AID-ACP413>3.0.CO;2-2.

Anderson, D. E., DePaulo, B. M., & Ansfield, M. E. (2002). The development of deception detection skill: A longitudinal study of same-sex friends. *Personality and Social Psychology Bulletin*, *28*, 536–545. doi:10.1177/0146167202287010.

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 Behavior*, *23*, 67–89. doi:10.1023/A:1021387326192.

Bala, N., Ramakrishnan, K., Lindsay, R. C. L., & Lee, K. (2005). Judicial assessment of the credibility of child witnesses. *Alberta Law Review, 42*, 995–1017.

Bond, G. D. (in press). Deception detection expertise. *Law and Human Behaviour*.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214–234. doi:10.1207/s15327957pspr1003_2.

Bond, C. F., Jr., & DePaulo, (in press). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*.

Bond, C. F., Jr., & Uysal, A. (2007). On lie detection "Wizards". *Law and Human Behavior, 31*, 109–115. doi:10.1007/s10979-006-9016-1.

Bruck, M., Ceci, S. J., & Hembrooke, H. (1998). Reliability and credibility of young children's reports: From research to policy and practice. *American Psychologist*, *53*, 136–151. doi:10.1037/0003-066X.53.2.136.

Canada Customs and Revenue Agency. (1998). *Student customs officer training program handbook – air*. Rigaud, QC: Author.

Ceci, S. J., & Bruck, M. (1995). *Jeopardy in the courtroom: A scientific analysis of children's testimony*. Washington, DC: American Psychological Association.

Davis, M., Markus, K. A., & Walters, S. B. (2006). Judging the credibility of criminal suspect statements: Does mode of presentation matter? *Journal of Nonverbal Behavior, 30*, 181–198. doi:10.1007/s10919-006-0016-0.

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effects in the communication of deception. In J. Yuille (Ed.), *Credibility assessment*. Belgium: Kluwer Academic Publishers.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74–118. doi:10.1037/0033-2909.129.1.74.

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology*, *16*, 249–267. doi:10.1111/j.1559-1816.1986.tb01138.x.

DePaulo, B. M., & Tang, J. (1994). Social anxiety and social judgment: The example of detecting deception. *Journal of*

Research in Personality, 28, 142–153. doi:10.1006/jrpe.1994.1012.

Edelstein, R. S., Luten, T., Ekman, P., & Goodman, G. S. (2006). Detecting lies in children and adults. *Law and Human Behavior, 30*, 1–10. doi:10.1007/s10979-006-9031-2.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 9*, 913–920. doi:10.1037/0003-066X.46.9.913.

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

Etcoff, N. L., Ekman, P., Magee, J. J., & Frank, M. G. (2000). Lie detection and language comprehension. *Nature, 405*, 139.

Feldman, R. S., Jerkins, L, & Popoola, O. (1979). Detection of deception in adults and children via facial expressions. *Child Development*, *50*, 350–355. doi:10.2307/1129409.

Feldman, R. S., & White, J. B. (1980). Detecting deception in children. *Journal of Communication*, *30*, 121–128. doi:10.1111/j.1460-2466.1980.tb01974.x.

Frank, F. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stakes lies. *Journal of Personality and Social Psychology, 72*, 1429–1439. doi:10.1037/0022-3514.72.6.1429.

Fritzley, V. H., & Lee, K. (2003). Do children always say yes to yes-no questions? A metadevelopmental study of the affirmation bias. *Child Development, 74*, 1297–1313. doi:10.1111/1467-8624.00608.

Garrido, E., Masip, J., & Herrero, C. (2004). Police officer's credibility judgments: Accuracy and estimated ability. *International Journal of Psychology*, *39*, 254–275. doi:10.1080/00207590344000411.

Granhag, P. A., & Strömwall, L. A. (2001). Deception detection based on repeated interrogations. *Legal and Criminological Psychology*, *6*, 85–101. doi:10.1348/135532501168217.

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

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. doi:10.1023/A:1022330011811.

Kassin, S. M., Leo, R. A., Meissner, C. A., Richman, K. D., Colwell, L. H., Leach, A.-M., & LaFon, D. (2007). Police interviewing and interrogation: A self-report survey of police practices and beliefs. *Law and Human Behavior, 31*, 381–400. doi:10.1007/s10979–006-9073-5.

Kraut, R. E., & Poe, D. (1980). Behavioral roots of person perception: The deception judgments of customs inspectors and laypersons. *Journal of Personality and Social Psychology*, *39*, 784–798. doi:10.1037/0022-3514.39.5.784.

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

Lane, J. D., & DePaulo, B. M. (1999). Completing Coyne's cycle: Dysphorics' ability to detect deception. *Journal of Research in Personality, 33*, 311–329. doi:10.1006/jrpe.1999.2253.

Leach, A.-M., Talwar, V., Lee, K., Bala, N. C., & Lindsay, R. C. L. (2004). "Intuitive" lie detection of children's deception by law enforcement officials and university students. *Law and Human Behavior, 28*, 661–685. doi:10.1007/s10979-004-0793-0.

Lewis, M., Stanger, C., & Sullivan, M. W. (1989). Deception in 3-year-olds. *Developmental Psychology, 25*, 439–443. doi:10.1037/0012-1649.25.3.439.

Mann, S., & Vrij, A. (2006). Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. *Psychology, Crime, & Law, 12*, 307–319.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology, 89*, 137–149. doi:10.1037/0021-9010.89.1.137.

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. doi:10.1023/A:1020278620751.

O'Sullivan, M. (2005). Emotional intelligence and deception detection: Why most people can't "read" others, but a few can. In R. E. Riggio & R. S. Feldman (Eds.), *Applications of nonverbal communication* (pp. 215–253). Mahwah, NJ, US: Lawrence Erlbaum Associates.

O'Sullivan, M. (2007). Unicorns or Tiger Woods: Are lie detection experts myths or rarities? A response to On lie detection "Wizards" by Bond and Ulysal. *Law and Human Behavior, 31*, 117–123. doi:10.1007/s10979-006-9058-4.

O'Sullivan, M., & Ekman, P. (2004). The wizards of deception detection. In P. A. Granhag & L. A. Strömwall (Eds.), *Deception detection in forensic contexts* (pp. 269–286). Cambridge: Cambridge University Press.

Porter, S., Campbell, M. A., Stapleton, J., & Birt, A. R. (2002). The influence of judge, target, and stimulus characteristics on the accuracy of detecting deceit. *Canadian Journal of Behavioural Science, 34*, 172–185.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. London: Sage.

Saarni, C., & Salisch, M. V. (1993). The socialization of emotional dissemblance. In M. Lewis & C. Saarni (Eds.), *Lying and deception in everyday life* (pp. 106–125). New York: The Guilford Press.

Sporer, S. (1997). The less travelled road to truth: Verbal cues in deception detection accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11*, 373–397.

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development, 26*, 436–444.

Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2006). Adults' judgments of child witness credibility and veracity. *Law and Human Behavior, 30*, 561–570.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 years. *Psychology, Public Policy, and Law, 11*, 3–41.

Vrij, A., Akehurst, L., Brown, L., & Mann, S. (2006). Detecting lies in young children, adolescents, and adults. *Applies Cognitive Psychology, 20*, 1225–1237.

Vrij, A., & Baxter, M. (1999). Accuracy and confidence in detecting truths and lies in elaborations and denials: Truth bias, lie bias and individual differences. *Expert Evidence, 7*, 25–36.

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

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

Vrij, A., & Graham, S. (1997). Individual differences between liars and the ability to detect lies. *Expert Evidence, 5*, 144–148.

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., & Mann, S. (2004). Detecting deception: The benefit of looking at a combination of behavioral, auditory and speech content related cues in a systematic manner. *Group Decision and Negotiation, 13*, 61–79.

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.

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

Wells, G. L., & Windshitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin, 25*, 1115–1125.

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