THE RELIABILITY OF "INTUITIVE" LIE DETECTION PERFORMANCE

by

AMY-MAY LEACH

A thesis submitted to the Department of Psychology

in conformity with the requirements

for the degree of Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

September 2006

Your file Votre référence
ISBN: 978-0-494-18526-1
Our file Notre référence
ISBN: 978-0-494-18526-1

Abstract

This program of research 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 *(Experiment 1)*, participants' performance was highly reliable. However, rating adults who provided truthful or fabricated accounts did not produce a significant test-retest correlation (*Experiment 2*). This lack of reliability was not due to the types of deceivers (i.e., children vs. adults) or interviews (i.e., closed-ended questions vs. extended accounts) (*Experiment 3*). Finally, the type of deceptive scenario (naturalistic vs. experimentally-manipulated) could not account for differences in reliability (*Experiment 4*). These findings are discussed in theoretical and legal contexts.

Acknowledgements

Many people have gotten me to this point. I would like to thank Dr. Rod Lindsay for everything that he has done for me. He is more than a supervisor; he is a mentor whom I wish to emulate. I always learned something from my conversations with him and I will miss our "little chats." Also, I would like to thank Drs. Lee Fabrigar and Ron Holden, who were my surrogate supervisors during Rod's sabbatical, for having gone above and beyond their duties as committee members. In addition, I am appreciative of the efforts of the members of the Child Witness Project (specifically, Drs. Lee and Talwar and Professor Nicholas Bala) and everyone who participated in my studies; without them, this research would not have been possible. Finally, I would like to thank the many others (in particular, Chris Degen, Roland Dupras, Rick Eves, Steve Ferguson, Monica Hurt, Jane Leach, Marilyn Lindsay, and Jill Moore) who worked behind the scenes to keep my research, and life as a graduate student, running smoothly.

On a personal note, I would like to thank the people in my life for their encouragement. I am grateful to my friends, who have forced me to celebrate all of the little steps along the way. My family's support was also invaluable. They always knew that I could do it, even when they didn't know what "it" was. In addition, as has become the lab tradition, I would like to thank Rod and Marilyn for keeping a starving student well-fed over the years.

Table of Contents

Abstract.. i

Acknowledgements.. ii

Table of Contents.. iii

List of Tables.. v

List of Figures... vi

Chapter 1: General Introduction.. 1

Chapter 2: Experiment 1

Introduction... 16

Method... 17

Results.. 19

Discussion... 24

Chapter 3: Experiment 2

Introduction... 26

Method... 26

Results.. 28

Discussion... 33

Chapter 4: Experiment 3

Introduction... 36

Method... 36

Results.. 38

Discussion... 46

Chapter 5: Experiment 4

Introduction........ 48

Method........ 48

Results........ 51

Discussion........ 57

Chapter 6: Consideration of Stimuli........ 58

Chapter 7: General Discussion........ 60

References........ 70

Appendix A: Prototypical Approach vs. Propositional Approach........ 79

Appendix B: Timed Anagram Test........ 82

List of Tables

Table 1. Distribution of Participants' Accuracy during Session One and Session Two in Experiment 1 23

Table 2. Distribution of Participants' Accuracy during Session One and Session Two in Experiment 2 32

Table 3. Distribution of Participants' Accuracy during Session One and Session Two in Experiment 3 43

Table 4. Distribution of Participants' Accuracy during Session One and Session Two in Experiment 4 56

Table 5. Proportions of Targets Identified at Various Levels of Accuracy in Experiments 1-4 59

List of Figures

Figure 1. Average percentage of correct identifications as a function of session and type of statement (Experiment 2).. 30

Figure 2. Average percentage of correct identifications as a function of session and type of statement (Experiment 3).. 41

Figure 3. Average confidence as a function of participant sex, target age and type of interview (Experiment 3).. 45

Figure 4. Average percentage of correct identifications as a function of session, type of statement, and scenario (Experiment 4) 53

Chapter 1

General Introduction

Individuals, such as police and social workers, often are asked to determine the veracity of people's statements based on "intuitive" lie detection; that is, they must rely upon experience and common sense rather than explicit training or technology (e.g., the polygraph machine). Is it true that some (or all) people have the ability to detect lies consistently? To date, the reliability of observers' performance has not been examined. Specifically, it is unknown whether measured lie detection performance is stable or due to a chance occurrence (e.g., a series of lucky or unlucky guesses). The present studies examined the reliability of individuals' lie detection accuracy with a test-retest paradigm.

Although this program of research focuses primarily on the reliability of lie detection performance, it might prove useful to situate it within the field of deception detection as a whole. To that end, a working definition of lying has been included. In addition, an overview of the approaches and findings related to the cues to deception and the detection of deception by special populations will follow. Both of these topics are relevant to the study of deception, but are not a major focus due to the rigorous training and/or sophisticated equipment that they require. Instead, the present research will examine "intuitive" lie detection, which is practiced by laypersons on a daily basis (e.g., DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). Thus, none of the aforementioned areas will be explicitly addressed in the present experiments; however, their review provides a sense of the direction of deception detection research and substantiates the need for an examination of the reliability of lie detection performance.

What is a lie?

The definition of lying has been the subject of philosophical debate for several centuries and will not be resolved here. One perspective has been offered by the propositional approach. According to this perspective, lying occurs when (a) the speaker makes a solemn assertion to the listener, (b) the speaker does not believe that the statement is true, and (c) the speaker intends for the listener to believe that the statement is true (Chisholm & Feehen, 1977). Thus, lying is an all-or-nothing event, as a statement can only be considered a lie if all of the conditions are met. More recently, the prototypical approach has emerged to challenge this strict conceptualization of lying. The prototypical perspective posits that three factors *may* be used to classify a statement as a lie: (a) the statement is factually false; (b) the speaker believes that the statement is false; and (c) the speaker intends to deceive the listener (Coleman & Kay, 1981). If at least one of these criteria is met, the statement is a lie. As the number of elements that are observed increases, the statement approaches prototypical lying. In fact, the statement is deemed a prototypical lie when all three elements are present. Examples of the different types of lies that can be produced using the prototypical and propositional approaches are provided in Appendix A. The prototypical approach is more inclusive and more closely reflects how adults and children conceptualize deception (e.g., Coleman & Kay, 1981; Strichartz & Burton, 1990). Although some distinctions can be made between these two approaches, there is significant overlap between both perspectives. The types of deception examined in the following experiments meet the criteria of the propositional and prototypical approaches, making an explicit distinction between the two unnecessary.

Cues to deception

There is a substantial literature that indicates that there are physiological cues to deception. The polygraph, commonly referred to as the 'lie detector machine', operates on the assumption that people experience guilt, or fear of detection, when they lie and that that guilt or fear produces changes in autonomic arousal. By monitoring how an individual's responsiveness varies with the type of questions asked, examiners are thought to be able to determine when an individual is lying. There are reports that the test can be beaten when suspects employ physical (e.g., biting the tongue) or mental (e.g., counting backward) countermeasures to falsify patterns of arousal (Honts, Raskin, & Kircher, 1994). More disturbingly, although the polygraph can be more accurate than the level of chance, the number of false positives (innocent suspects who are deemed guilty) has been deemed unacceptably high (Ben-Shakhar & Elaad, 2003; Kircher, Horowitz, & Raskin, 1988; MacLaren, 2001). As a result, members of the scientific community question the polygraph machine's effectiveness as a lie detection tool (Iacono & Lykken, 1997) and polygraph results are not admissible in Canadian or, in certain contexts, American courts (*R v. Béland*, 1987; *United States v. Scheffer*, 1998).

Some tools, such as Criteria-Based Content Analysis (CBCA) and Reality Monitoring (RM), focus on verbal – rather than physiological – cues to deception. CBCA is based on the premise that truthful statements differ from fabrications in terms of their content (e.g., the types of details provided), as assessed by 19 criteria (for a full review, see Ruby & Brigham, 1997). Individuals who have actually experienced an event are expected to recount it differently than those who have not. RM researchers suggest that memories for experienced and imagined events differ in terms of perceptual processes (e.g., visual details) and cognitive operations (e.g., inclusion of thoughts) (e.g., Sporer, 1997). Vrij and Mann's (2004) systematic analysis of CBCA and RM cues indicate that both tools can be used to detect deception well above the level of chance. In addition, a recent review has shown that accuracy rates with RM and CBCA are comparable (Masip, Sporer, Garrido, & Herrero, 2005). Although these results are encouraging, the training requirements, combined with the time-consuming transcription and scoring procedures, render CBCA and RM inaccessible to most people.

Some researchers maintain that non-verbal behaviours, which are more difficult to control than speech, are best at revealing deception (e.g., Ekman, O'Sullivan, Friesen, & Scherer, 1991). The observation of micro-expressions (i.e., split-second observations of behaviour) is associated with successful lie detection (Ekman & O'Sullivan, 1991; Frank & Ekman, 1997). In addition, truth-tellers and lie-tellers can be correctly classified, on the basis of their nonverbal behaviour, at accuracy rates approaching 80% (Vrij, Edward, Roberts, & Bull, 2000). Other researchers report that individuals are better at detecting deception when non-verbal cues are removed, particularly when they cannot see the deceiver and must rely solely upon the person's speech (e.g., Zuckerman, Koestner, & Colella, 1985). A recent meta-analysis of lie detection studies has shown that there are several differences between lie-tellers and truth-tellers, with distinctions being made in terms of both vocal (e.g., pitch) and nonverbal (e.g., pupil dilation) behaviours (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003). However, for the most part, individuals do not have access to the training and equipment necessary to perceive these differences.

Deception detection: Special Populations

Other research has focused on whether groups can detect lies intuitively. Although some evidence implies that laypersons are good lie detectors (e.g., DePaulo & Rosenthal, 1979), other researchers suggest that people, in general, cannot detect deception (e.g., Ekman & Friesen, 1974; Ekman & O'Sullivan, 1991). Given these mixed findings, untrained observers' ability to detect lies is questionable (for reviews, see Ambady & Rosenthal, 1992; DePaulo, Zuckerman, & Rosenthal, 1980). In turn, researchers have posited that law enforcement officials' extensive training and experience with lie detection would lead them to be more accurate. A few groups (Secret Service agents, CIA agents, sheriffs, and forensic clinical psychologists) have been able to perform above the level of chance (Ekman & O'Sullivan, 1991; Ekman, O'Sullivan, & Frank, 1999). However, the majority of the groups that have been examined (such as police officers, customs officials, and FBI agents) are as inaccurate as untrained observers (e.g., Ekman & O'Sullivan, 1991; Garrido, Masip, & Herrero, 2004; Kraut & Poe, 1980), with some groups (e.g., federal parole officers) performing worse than chance (Porter, Woodworth, & Birt, 2000). In addition, despite their poor performance, these groups tend to express unwarranted confidence in their decisions (e.g., DePaulo & Pfeifer, 1986; Leach, Talwar, Lee, Bala, Lindsay, 2004). Yet, there is no relationship between accuracy and confidence (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997). Thus, not only are individuals unable to detect deception, they do not recognize that fact.

Deception detection: Individual differences

The frequent finding of near-chance performance might suggest that people lack the ability to detect lies and that occasional above-chance performance is simply due to random variation in the success of guessing. Of course, this is a poor explanation for results demonstrating significant group differences (Ekman & O'Sullivan, 1991; Ekman, et al., 1999). If Secret Service agents are superior to control groups, it is difficult to attribute this to luck. How can researchers explain the fact that some groups are able to detect lies, whereas the general population appears to function at (or near) chance levels? It is possible that variation in lie detection ability is randomly distributed in the population and that certain professions attract individuals who have the ability to detect lies. It is also possible that untrained observers have no significant ability to detect lies and the professions that perform better than chance do so as a result of the training or experience provided by the occupation. It is important to know which of these explanations (if either) is true. The first pattern suggests that good intuitive lie detectors can be selected. The second pattern suggests that appropriate training can result in good intuitive lie detectors. A third possibility is that some people, but not all, can be trained, such that both selection and training are required to produce proficient lie detectors.

Researchers' focus on the performance of special populations may have obscured individual differences in lie detection. It is reasonable to assume that lie detection performance lies on a continuum. For example, Vrij and Graham (1997) found that the accuracy of untrained individuals ranged from 20% to 70% (with police officers' accuracy ranging from 20% to 90%). However, there are few explanations for why some individuals are more accurate than others. Sex differences have been closely examined, but they have produced mixed findings. Women tend to be better at deciphering speakers' nonverbal behaviours, but they are generally less suspicious than men (DePaulo, Epstein,

& Wyer, 1993). In turn, the majority of researchers have failed to find sex-related differences in lie detection ability (for a full review, see Zuckerman, DePaulo, & Rosenthal, 1981). Many other attributes (e.g., self-monitoring, shyness, extraversion, the "Big 5") are not linked to performance (Porter, Campbell, Stapleton, & Birt, 2002; Vrij & Baxter, 1999; Zuckerman et al., 1981). Although certain characteristics (e.g., heightened social anxiety, dysphoria, aphasia, left-handedness) affect lie detection performance (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 the wide variations in accuracy within the general population.

Reliability of lie detection

There is a widely held assumption that some people have the ability to detect deception, whereas others do not (e.g., Ekman et al., 1999). Although there has not been an explicit conceptualization of lie detection as a skill or personality trait, lie detection is considered to be an underlying, inherent part of the person. However, researchers have not addressed the most fundamental requirement of a valid interpersonal difference: test-retest reliability (or, alternate forms as different stimuli should be used to test the stability of performance over time). If there is no evidence of reliability, then the notion that people have the ability to detect lies is incorrect and fluctuations in data can be attributed to random guessing. Correspondingly, the near-chance overall performance in most studies is consistent with the possibility that high and low performers were simply lucky and unlucky guessers. The relationship between validity and reliability is important because it affects the interpretation of previous research. For example, if underlying lie detection ability did not exist, it would be unclear what researchers were measuring and their findings may replicate only because performance is consistently at or near chance levels. Thus, establishing the reliability of performance is an important component in lie detection research.

Usually, individuals' accuracy has been assessed within one session (e.g., DePaulo & Pfeifer, 1986; Ekman et al., 1999). In a typical experiment, an individual is asked to view a series of video clips. Each video clip features a speaker who is communicating either truthful or deceptive information. After each clip, the observers must decide whether the speaker (i.e., the person featured in the video) was lying or telling the truth. Although observers may make upwards of 40 lie detection decisions (e.g., Leach et al., 2004; Vrij, Evans, Akehurst, & Mann, 2004), the majority of studies feature fewer than 20 targets (e.g., Ekman & O'Sullivan, 1991; Lane & DePaulo, 1999; Vrij, Edward, & Bull, 2001). At the end of the session, the individual's accuracy is calculated. Given this procedure, a person is deemed a good (or poor) lie detector based on their discrimination between a limited number of lie-tellers and truth-tellers on a single occasion. Thus, it is difficult to assess whether lie detection ability is a trait or characteristic of individuals from previous research because performance could simply be the result of a series of (lucky or unlucky) guesses.

Indirectly, a few researchers have examined the reliability of lie detection performance. Granhag and Strömwall (2001) had observers and interviewers detect deception over the course of repeated interrogations. Observers' performance increased over time (one week). Yet, these researchers did not test for the consistency of performance and, in turn, offer no insight into whether accuracy was reliable. Another project assessed deception detection within same-sex friendships at two points – 5 months apart (Anderson, DePaulo, & Ansfield, 2002; Anderson, DePaulo, Ansfield, Tickle, & Green 1999). There were no changes in the types of cues used, nor the average accuracy of observers. However, the researchers involved did not explicitly examine the reliability of individual performance. Although global performance at a group level did not change over time, there is no direct evidence that individuals' performance was stable (or unstable). In addition, the studies only examined friends and, thus, the results might not generalize to other contexts (e.g., the deception of strangers). Ekman and O'Sullivan's (2004) Wizards Project provides additional evidence about the reliability of deception detection within a small segment of the population. After assessing hundreds of individuals, these researchers have found fourteen people who have achieved high levels of lie detection accuracy across three types of deceptive scenarios (i.e., featuring lies about emotions, opinions, and crimes). Thus, the "wizards'" performance is stable across lie detection tasks. Yet, Ekman and O'Sullivan indicate that accuracy rates vary with the type of deception. It is unknown whether these fluctuations suggest that performance is not completely reliable or that some tasks are less successful at tapping into underlying ability. Also, Ekman and O'Sullivan focused on the performance of highly proficient individuals. Apparently, there have been no direct tests of the reliability of lie detection within the general population.

Experimentally-manipulated vs. naturalistic deception

Also, the majority of lie detection studies involve the artificial production of deception (henceforth referred to as "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. Topics of deception include misrepresenting life events (e.g., Anderson et al., 2002; Porter et al., 2002), opinions (e.g., Ekman et al., 1999; Feldman, Jenkins, & Popoola, 1979), and 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, breaking and entering) 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 behaviour.

Do current procedures produce ecologically valid examples of 'lying'? The majority of respondents in the scientific community were hesitant to accord weight to polygraph findings that were obtained using experimentally-manipulated deception (Iacono & Lykken, 1997). Ekman (1996; 1997) argues that an integral part of deception is that the listener is *unaware* of the deceiver's intent; thus, "an actor is not a liar." There might be an assumption that the object of the speaker's deception is the lie detector (i.e., the participant who decides whether individuals are lying or telling the truth at a later date). We posit that the actual focus of the deceiver's account is the experimenter. Indeed, often the lie-teller is taped and does not know when, or to whom, the lie will be shown. If this is true, and both the deceiver and experimenter are operating under an expectation that the lie-teller is producing a stimulus for a future experiment, then it could be said that the majority of lie detection tasks involve acting rather than lying. Also, the deceiver's awareness of the purpose of the study ought to lessen any guilt or unease associated with particular behaviours (e.g., stealing and lying). In fact, they should perceive the situation as an explicit endorsement of lying within the experimental context. Finally, some studies have included only poor lie-tellers, or those whose deception was readily observable (e.g., Ekman & O'Sullivan, 1991; Ekman et al., 1999). Selecting the stimuli in this manner may decrease external validity because, in real life, there is a full range of deceptive ability. Indeed, one might expect unsuccessful lie tellers to lie less often than successful lie tellers because they are likely rewarded less often and punished more often for lying.

It is not fully 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, because there are only a handful of studies that feature this type of deception. In one study (Vrij & Mann, 2001a), police officers attempted to discriminate between truthful and deceptive fragments of a convicted murderer's police interview. Individuals were able to detect deception above the level of chance. However, these findings must be treated with caution because the study featured only one 'suspect' (i.e., the murderer), that suspect spoke a foreign language that could not be understood by the lie detectors, and a limited number of lie detection decisions were made (i.e., six fragments were rated in all). Vrij and Mann (2001b) attempted to resolve some of these issues by having police officers view videotapes of individuals who had made televised appeals to the public to find their missing relatives or their relatives' murderers. All of those who made these appeals were later found to have murdered their relatives themselves. The officers could detect deception at the level of chance. But, as in the previous study, the appeals were made in a foreign language and only lies were detected (i.e., fillers were rated, but not included in analyses because there was no way to ensure that these individuals were really telling the truth). In another study, researchers

(Mann, Vrij, & Bull, 2004) examined the detection of suspects' lie-telling and truth-telling about a variety of actual events (e.g., theft, murder). Police officers' accuracy was approximately 65%, which is higher than expected by chance. This research suggests that the detection of naturalistic deception differs from experimentally-manipulated deception. At the very least, discriminating between truth-telling and lie-telling appears to be easier during the former. How this affects the reliability of lie detection performance is unknown.

Children's deception

Although the majority of researchers have focused on adults' deception, 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). After children have been provided with distorted details of an event, clinicians and researchers have difficulty correctly identifying the accuracy of children's claims (Ceci & Huffman, 1997). One problem associated with this suggestibility research is the possibility that children have come to believe that the fabricated events actually happened. The few studies that have examined children's ability to knowingly and intentionally lie have produced inconsistent results. One set of findings indicated that deception improves with age (Feldman et al., 1979; Feldman & White, 1980). That is, young children's deception was easier to detect than that of older individuals (e.g., college students). However, this work focused on inconsequential, experimentally-manipulated deception (i.e., individuals were told to lie or tell the truth about their opinions). More recent research by Lewis, Stanger, and Sullivan (1989) has examined the deception of children after an actual transgression (peeking at a forbidden toy). In that study, undergraduate students could not discriminate between 3-year-old lie-tellers and truth-tellers. Replications and extensions of this research have shown that undergraduate students, police officers, customs officers, and even the children's own parents also have difficulty differentiating between 3- to 12-year-old lie-tellers and truth-tellers (Leach et al., 2004; Talwar & Lee, 2002). In addition, cues to deceit are similar across children and adults (Vrij, Akehurst, Soukara, & Bull, 2004). Although the weight of the evidence suggests that children's and adults' deception is comparable, it is not clear whether the stability of lie detection performance would be affected by the age of the deceiver.

The present experiments

The purpose of these experiments was to directly assess the reliability of lie detection performance. First, individuals' accuracy was measured across two separate sessions, approximately one week apart. A test-retest paradigm was employed to reveal whether lie detection performance was due to consistent responses or simply random guessing.

After addressing the first issue, the generalizability of the result was examined across several relevant domains. The reliability of lie detection accuracy was assessed using different types of scenarios. Specifically, both naturalistic and experimentally-manipulated deception were examined. Naturalistic deception scenarios gave individuals the opportunity to commit a transgression (i.e., peek at a toy or cheat on a test), but no particular instructions were given; their actions were volitional. 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). Thus, individuals' lie detection performance was examined in a variety of contexts.

Additionally, two different populations, children and adults, were examined. The mixed findings present in previous research suggest that there might be differences between adults' and children's deception. Even if both groups' deception is equally difficult to detect, it is still possible that the age of the deceiver could affect the reliability of observers' lie detection performance. To explore this possibility, children's and adults' deception was examined.

Finally, a large number of deceivers and truth-tellers were included, regardless of their ability to deceive. That is, targets were not selected because they were poor or proficient deceivers; all eligible lie-tellers and truth-tellers were included in the studies. This approach ensured that the examples of deception represented a full range of believability, as it varies (across good and poor lie-tellers *and* truth-tellers) in the real world.

This program of research consisted of four experiments. In each experiment, undergraduate students were shown video footage of individuals discussing events. Participants were asked to differentiate between the lie-tellers and truth-tellers during two separate sessions. In *Experiment 1,* undergraduates viewed naturalistic deception, in which children lied or told the truth about an actual transgression (i.e., peeking at a forbidden toy). *Experiment 2* featured experimentally-manipulated footage of adults who were instructed to lie or tell the truth about an event (i.e., they discussed an event that either had or had not actually occurred). Again, in *Experiment 3*, children and adults were instructed to lie or tell the truth when recounting events and answering closed-ended questions about their narratives. Participants were shown either children or adults providing either full accounts or short (yes-no) responses. Finally, in *Experiment 4*, participants detected the deception of undergraduate students who had been given the opportunity to commit a transgression (cheating on a test). Half of the targets were given explicit instruction about their actions (producing experimentally-manipulated deception), whereas the other individuals acted volitionally (producing naturalistic deception).

Given previous, indirect, research about the stability of lie detection accuracy, performance was expected to be reliable. That is, it was hypothesized that there would be a statistically significant, positive correlation between accuracy over both sessions. Overall, accuracy was expected to be poor (i.e., at the level of chance). However, naturalistic deception was expected to be easier to detect than experimentally-manipulated deception. Although a relationship between confidence and accuracy was not hypothesized, participants' levels of certainty in their decisions were expected to be reliable. Differences were not expected between the detection of children's and adults' deception.

Chapter 2

Experiment 1

This study examined the reliability of the detection of children's naturalistic deception. Children were interviewed, in a neutral context, about having committed a transgression (peeking at a forbidden toy while the experimenter was out of the room). They were not instructed to provide a particular response when they were questioned about the matter (i.e., they could lie or tell the truth about their actions of their own volition). Short video clips of the children's responses were shown to undergraduate students over the course of two sessions, held approximately one week apart. Given previous research with the same stimuli (Leach et al., 2004), it was posited that observers would have difficulty discriminating between truth-tellers and lie-tellers (i.e., performance would be at chance levels). However, children's deception may have been difficult to detect because their behaviour 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 behaviours in various contexts (Saarni & Von Salisch, 1993). For example, children's eyes might wander during a conversation because they are unaware of the nonverbal behaviour (eye contact) that is most suitable in that situation. One initial consequence is that these children might be labelled lie-tellers because adults often list averted gaze as a marker of deception (Akehurst, Köhnken, Vrij, & Bull, 1996; Vrij & Semin, 1996). However, after viewing several children exhibiting these patterns, observers might become confused and uncertain about their decisions. That is, they could recognize that it is unlikely that all of the children are lie-tellers and vacillate between attributing averted gaze to lying or more innocuous explanations (e.g., that the child is shy). It was unknown how this would affect guessing and the stability of performance.

Method

Participants

A total of 81 undergraduate students participated in the study in exchange for course credit; however, 15 individuals were dropped from analyses because they did not attend both testing sessions. Preliminary analyses failed to reveal any differences, in first session performance, between individuals who did and did not return – this was also true for subsequent studies. Overall, 66 students (50 women and 16 men, *M* age = 20.23 years, *SD* = 4.81) completed the study.

Materials

Video clips. Video footage of a temptation resistance paradigm was 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 seconds, *SD* = 6.66) produced from the recorded exchanges (featuring all three questions about whether the child had peeked). 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 to determine whether each child was lying or telling the truth using a forced-choice paradigm. In addition, participants were asked to indicate their level of confidence in their decisions as a percentage (from 0% to 100% confident).

Procedure

Participants were tested individually, or in small groups, in a quiet room. A female experimenter randomly assigned participants to one of two groups (which determined which of the two videotapes they viewed first). In each session, the entire procedure took approximately 45 minutes to complete. Prior to playing the videotape, the experimenter explained the task, noting that the clips were randomly presented and that about half of the children were lying/telling the truth. This information was also provided in all subsequent experiments. The videotape was shown with pauses in between the clips to allow participants time to record their responses. Approximately one week later, participants returned for the second session. The procedure was identical to that of *Session 1*, except that they were shown the previously unseen videotape (which featured different lie- and truth-tellers).

Results

Preliminary analyses failed to reveal any significant effects of participant sex. 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 ($M = .93$, $SD = .14$), suggesting that they were paying attention to the video clips. Because eliminating participants who misidentified one or more confessors ($n = 23$) did not change any of the findings, all data were included in the analyses. There were not enough confession trials, nor any false confessions, to justify a meaningful lie detection analysis. As a result, 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", whereas each incorrect decision was given a "0". These scores were averaged across children 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 (Rosenthal, 1991). These approaches were employed in all subsequent experiments.

Lie Detection Reliability

A correlational analysis was performed to determine whether participants' performance was stable over time. There was a significant correlation between accuracy in the first and second sessions, $r(65) = 0.64, p < .001$.

One possibility is that this finding might have been due to the 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 the binomial theorem analysis described below (to determine individual performance differing from chance). A correlational analysis revealed that there was a significant correlation between accuracy in the two sessions, $r(60) = 0.43, p < .001$.

Although sex differences were not observed when comparing overall accuracy, it is still possible that reliability could be affected. There was a significant correlation between accuracy in the first and second sessions for women, $r(49) = 0.66, p < .001$, and men, $r(15) = 0.54, p < .05$.

Lie Detection Accuracy

The range of scores was examined for descriptive purposes. In session one, accuracy ranged from 17% to 71% correct. This range was expanded somewhat in the second session (14% to 86% correct).

Additional analyses were conducted to determine whether participants' scores improved across sessions. A Statement type (Lie vs. Truth) x Session (Session 1 vs. Session 2) repeated-measures ANOVA was performed on participants' accuracy scores. Observers performed significantly better in the second session ($M = .53$, $SD = .12$) than in the first ($M = .48$, $SD = .11$), $F(1, 65) = 21.85, p < .001, r = .50$. There was no significant interaction between the variables.

One-sample t-tests were used to compare the overall accuracy of each session to the level of chance (i.e., 0.50). In session one, the overall accuracy of individuals was at chance, $t(65) = -1.77, p = .08$, whereas individuals performed significantly above chance in session two, $t(65) = 2.31, p < .05$.

Signal Detection Analyses

Accuracy reflects two separate dimensions of the participants' decision-making process: 1) discrimination between truth- and lie-tellers (usually referred to as d'), and 2) bias (i.e., the tendency to favour a particular response, such as categorizing children as lying, usually referred to as ß). To extract this information, additional analyses were conducted using signal detection theory.

Discrimination. A paired-samples t-test was performed on the d' values. Observers were better able to differentiate between lie-tellers and truth-tellers in the second session ($M = .18$, $SD = .67$) than in the first ($M = -.12$, $SD = .58$), $t(65) = 4.65, p < .001, r = .50$.

One-sample t-tests were performed to see whether observers could actually differentiate between the children during each session. In this analysis, the level of sensitivity was compared to 0 (no ability to differentiate between lie-tellers and truth-tellers). In session two, observers could reliably discriminate between the truth-tellers and lie-tellers, $t(65) = 2.24, p < .05$; in session one, they could not, $t(65) = -1.74, p = .09$.

Response Bias (ß). Another aspect of performance is response bias (ß). A paired-samples t-test was conducted on observers' tendency to indicate that children were lie-tellers (or truth-tellers) during each session. There were no significant differences between bias in session one ($M = 1.02$, $SD = 0.18$) and session two ($M = 1.06$, $SD = 0.30$), $t(65) = -0.98, p = .33, r = .12$.

Using t-tests, each ß was compared to 1 (no bias). Observers were not biased in session one, $t(65) = 1.10, p = .28$, or session two, $t(65) = 1.53, p = .13$. However, it is possible that the base rate information that was provided prior to this lie detection task (and in all subsequent experiments) could have influenced participants' biases.

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%. Using binomial theorem, it is possible to calculate the minimum number of correct or incorrect decisions required to indicate that the performance of the individual participant was significantly different ($p < .05$) from chance. Performance was significantly above chance if 23 or more of the participants' decisions were correct and significantly below chance if 12 or less of their decisions were correct. This analysis allows for a sense of the distribution of reliability (i.e., whether it was participants whose accuracy was below, above, or at chance who were most likely to display stable performance). The performance of the majority of participants ($n = 47$) was not significantly different from chance during both sessions. Very few participants performed consistently above ($n = 2$) or below chance ($n = 3$) (see Table 1).

Confidence

A Statement type x Session repeated-measures ANOVA was performed on participants' confidence ratings. Observers were significantly less confident in their decisions in the second session ($M = .62$, $SD = .13$) than in the first ($M = .64$, $SD = .12$),

Table 1

Distribution of Participants' Accuracy during Session One and Session Two in Experiment 1.

	Session One		
	Below chance	At chance	Above chance
Session Two			
Below chance	3	1	0
At chance	5	47	1
Above chance	0	7	2

$F(1, 65) = 4.86, p < .05, r = .27$. In addition, individuals were more confident when rating lie-telling ($M = .65, SD = .12$) than truth-telling ($M = .62, SD = .14$), $F(1, 65) = 18.80, p < .001, r = .47$. Additional analyses revealed that there was a significant correlation between average confidence in sessions one and two, $r(65) = 0.86, p < .001$. However, there was no significant relationship between confidence and accuracy in session one, $r(65) = .02, p = .90$, or session two, $r(65) = .04, p = .75$.

Discussion

Overall, participants could not accurately classify truth-tellers and lie-tellers. However, their accuracy improved across sessions (although, only to a level trivially above chance performance). In addition, there was no evidence of a response bias. More importantly, participants' performance and confidence were stable over time, such that a surprisingly high correlation existed between accuracy during the first and second sessions. This finding was surprising because observers' consistency was not a product of accuracy. In fact only 2 of 65 participants (3%), performed significantly better than chance during both sessions. Despite increases in lie detection accuracy, participants were less confident in their decisions over multiple sessions. There was no relationship between accuracy and confidence.

My primary interest in investigating this issue (i.e., the reliability of lie detection performance) is in its application to legal contexts. 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 the majority of law enforcement officials' lie detection decisions are based upon 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 when rating 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. Thus, results may not generalize to the detection of lengthy, open-ended accounts of events. Intuitively, lengthy accounts should offer more information (e.g., contradictions) upon which lie detectors can base their decisions. In turn, 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). *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.

Chapter 3

Experiment 2

Untrained observers' ability to detect children's deception appears to be reliable even though it is near chance levels. *Experiment 1* examined deception related to an actual transgression; it is unknown whether lie detection performance will generalize to other deceptive contexts (i.e., discussing a contrived event). In *Experiment 2*, the reliability of observers' detection of adults' deceptive narratives was tested using experimentally-manipulated deception. Despite differences in the types of lie-tellers and truth-tellers and the types of interviews noted earlier, the pattern of results may be the same (i.e., performance may be reliable). This pattern should occur if the reliability of lie detection performance is indicative of the ability of the observer (i.e., a stable characteristic of the person) rather than the stimuli.

Method

Participants

A total of 57 undergraduate students participated in the study in exchange for course credit; however, nine participants were 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 footage was compiled 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 story for one individual within the pair was the false story 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 seconds, *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 of the targets 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 explicitly asked whether each adult was lying or telling the truth about a particular event (e.g., car accident). In addition, participants were asked to indicate their level of confidence in their decisions as a percentage (from 0% to 100% confident).

Procedure

Participants were tested individually, at a computer terminal, in a quiet room. An experimenter randomly assigned participants to view one of the four subsets. The computer program carried out all other parts of the procedure. Participants read the instructions (on the computer screen) and provided demographic information (e.g., age). The video clips were shown with pauses in between to allow participants to make their decisions. Approximately one week later, participants returned for the second session.

Participants followed the same procedure, except that they saw a different subset of video clips than in the first session.

Results

Lie Detection Reliability

As in *Experiment 1*, a correlational analysis was performed to determine whether participants' performance was stable over time. There was no significant relationship between accuracy in sessions one and two, $r(47) = -0.02, p = .91$.

Unlike *Experiment* 1, there were no reliable outliers in this experiment (please see below for a more detailed description).

Again, the reliability of performance was examined in terms of sex differences. There was no significant correlation between accuracy in the first and second sessions for women, $r(31) = 0.03, p = .85$, or men, $r(15) = -0.34, p = .20$.

Lie Detection Accuracy

The range in accuracy appeared similar in session one (29% to 75% correct) and session two (29% to 67% correct).

Additional analyses were conducted to determine whether participants' scores improved across sessions. A Participant sex x Statement type (Lie vs. Truth) x Session (Session 1 vs. Session 2) mixed-factors ANOVA (with the last two variables as the repeated measures) was performed on participants' accuracy scores. There was a significant main effect of participant sex, such that women ($M = .53, SD = .07$) were more accurate than men ($M = .47, SD = .05$), $F(1, 46) = 7.53, p < .01, r = .38$. In addition, participants were significantly better able to identify truth-telling ($M = .61, SD = .11$) than lie-telling ($M = .40, SD = .12$), $F(1, 46) = 74.38, p < .001, r = .79$. However, this was qualified by a significant interaction between statement type and session $F(1, 46) = 8.72$, $p < .01$, $r = .40$. Specifically, participants were better at detecting truth-tellers in the second session when compared to the first session; detection of lie-tellers was equal across sessions (Figure 1). There were no other significant main effects or interactions.

One-sample t-tests compared accuracy to chance. Accuracy was at chance in session one, $t\ (47) = 0.12$, $p = .90$, and session two, $t\ (47) = 0.74$, $p = .46$. In addition, men performed significantly below chance, $t(15) = -2.32$, $p < .05$; women's accuracy was at chance, $t(31) = 1.97$, $p = .06$. Participants correctly identified truth-tellers at above chance levels, $t(47) = 7.17$, $p < .001$; however, they performed worse than expected by chance when rating lie-tellers, $t(47) = -6.14$, $p < .001$

Signal Detection Analyses

Discrimination. A Participant sex x Session mixed-factors ANOVA was performed on the d' values (i.e., discrimination). Women ($M = .14$, $SD = .39$) were significantly better able to differentiate between lie-tellers and truth-tellers than men ($M = -.16$, $SD = .31$), $F(1, 46) = 5.42$, $p < .05$, $r = .32$. There were no other significant main effects or interactions.

One-sample t-tests were performed to see whether observers could actually differentiate between the adult truth-tellers and lie-tellers during each session. Neither women, $t(31) = 1.88$, $p = .07$, nor men, $t(16) = -2.06$, $p = .06$, could discriminate between lie-tellers and truth-tellers. Overall, participants could not differentiate between the groups in session one, $t(47) = 0.20$, $p = .84$, or two, $t(47) = 0.86$, $p = .39$.

Response Bias (ß). A Participant sex x Session mixed-factors ANOVA was performed on observers' tendency to indicate that adults were lie-tellers (or truth-tellers)

Figure 1.

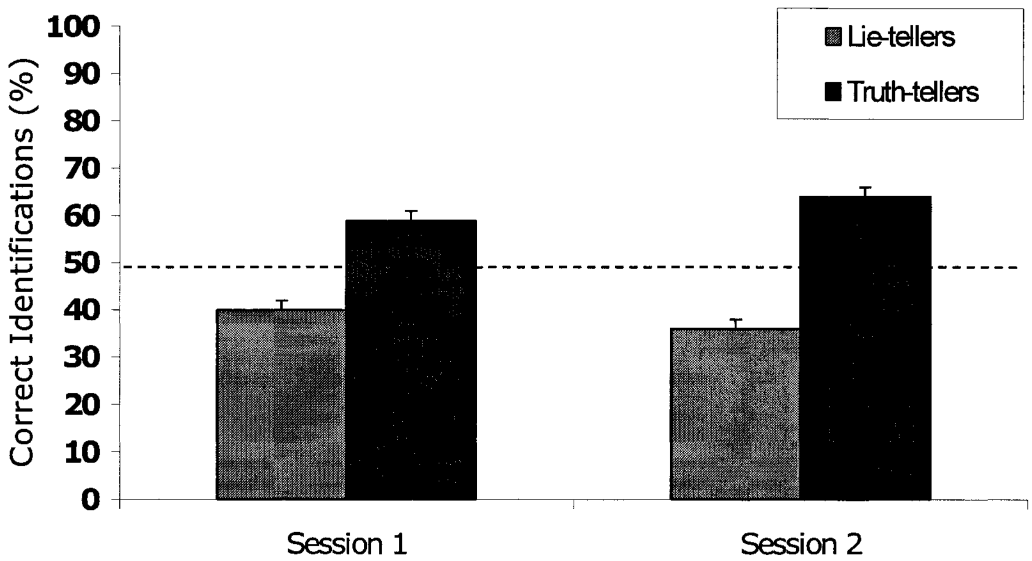

Figure 1. Average percentage of correct identifications as a function of session and type of statement.

during each session. Overall, there were no significant main effects or interactions between the variables.

Using t-tests, each ß was compared to 1 (no bias). Observers were not biased in session one, $t(47) = 1.33, p = .19$, or session two, $t(47) = 1.84, p = .07$.

Binomial theorem analysis

Again, for descriptive purposes, each individual's performance was compared to chance using a binomial theorem analysis. Performance was significantly above chance if 17 or more of the participants' decisions were correct and significantly below chance if 7 or less of the participants' decisions were correct. The majority of participants ($n = 43$) performed at chance (i.e., scored between 8 and 16 during both sessions) (see Table 2).

Confidence

A Participant sex x Statement type x Session mixed-factors ANOVA was performed participants' confidence ratings. Unlike in *Experiment* 1, there were no changes in confidence from the first session ($M = .67, SD = .10$) to the second ($M = .66, SD = .12$), $F(1, 46) = 0.31, p = .58$. In fact, there were no significant main effects or interactions.

Additional analyses revealed that there was a significant correlation between average confidence in sessions one and two, $r(47) = 0.82, p < .001$. There was no significant relationship between confidence and accuracy in session one, $r(47) = .16, p = .29$. The relationship was significant, but negative, in session two, $r(47) = -.38, p < .01$.

Table 2

Distribution of Participants' Accuracy during Session One and Session Two in Experiment 2.

	Session One		
	Below chance	At chance	Above chance
Session Two			
Below chance	0	2	0
At chance	2	43	1
Above chance	0	0	0

Discussion

Overall, participants had difficulty classifying truth-tellers and lie-tellers. However, women were more accurate than men. Although participants were better able to correctly identify truth-tellers than lie-tellers (particularly in the second session), signal detection analyses failed to reveal significant response biases. More importantly, lie detection performance was not reliable. This was surprising given the strong consistency in performance observed in the first experiment and descriptive data in the present experiment (offered by binomial theorem analyses) that suggested that the majority of participants performed at chance during both sessions (as in *Experiment 1)*. Participants' confidence was stable over time and there were no changes in confidence across sessions. Finally, although there was no relationship between accuracy and confidence in the first session, there was significant negative correlation between the two variables in the second session.

These results differed dramatically from those of *Experiment 1*. First, there was a relationship between accuracy and confidence. Specifically, in the second session, participants who were less accurate were more confident (and vice versa). This finding underscores individuals' inability to correctly represent their performance. It is particularly disturbing because law enforcement officials' intuitions about veracity have major consequences. For example, a prominent police training manual (Inbau, Reid, Buckley, & Jayne, 2001) recommends that officers use their assessments of suspects' veracity to determine, during pre-interrogation interviews, whether to proceed to interrogation or set the suspect free. Thus, if an officer is unjustifiably confident in her ability to detect deception, suspects could be at risk for guilt-presumptive interrogations, which could lead to false confessions (e.g., Kassin, Goldstein, & Savitsky, 2003). Although it is important to be cognizant of the risks associated with inaccurate perceptions of deception, it must be emphasized that the relationship between accuracy and confidence was only observed during one session of one experiment. In fact, the majority of researchers have failed to find a link between accuracy and confidence (DePaulo et al., 1997). The possibility of random error should be considered.

The most important finding was that lie detection performance was not reliable. There are several reasons why this might have been the case. First, the findings of either one of the two experiments could have been a fluke. That is, random error could account for performance being reliable (*Experiment* 1) or unreliable (*Experiment* 2). Second, the deception stimuli varied across the two experiments. In *Experiment 1* observers rated children and in *Experiment 2* observers rated adults. The majority of researchers have focused on the deception of either children or adults (e.g., Ekman et al., 1999; Leach et al., 2004). In fact, 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 (Feldman et al., 1979; Feldman & White, 1980), whereas the other suggested that there are no developmental differences (Leach et al., 2004; Lewis et al., 1989; Talwar & Lee, 2002). The lack of replication between the first experiment (featuring children) and the second experiment (featuring adults) suggests that individuals may not be as reliable when detecting both age groups' deception. However, these 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. A review of the literature (Ambady & Rosenthal, 1992) has shown that the length of the interview presented does not affect accuracy. However, this does not mean that the type of interview does not matter. Vrij and Baxter (1999) maintain that observers assess elaborations and denials differently due to the information afforded by each type of interview. Thus, 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 these issues.

Chapter 4

Experiment 3

Two issues were examined in *Experiment 3*. First, observers were asked to detect either adults' or children's deception. Given the mixed findings related to children's deceptive abilities, it was unclear how the reliability of observers' performance would be affected. However, researchers who had used identical stimuli to that present in *Experiment 1* found children's deception to be comparable to adults' (Leach et al., 2004; Talwar & Lee, 2002). Thus, there was no reason to expect that the age of the deceiver would affect accuracy and, in turn, reliability. A second comparison in this experiment involved two types of interviews (i.e., yes-no questions vs. open-ended accounts). Researchers suggest that open-ended accounts provide more material upon which to make a decision (Vrij & Baxter, 1999). However, this could negatively affect reliability. With so much information to process, individuals could have more difficulty sticking with a particular strategy. In turn, their performance would not appear stable over time. Thus, it was expected that performance would be reliable when observers rated yes-no responses, but not narratives.

Method

Participants

A total of 230 undergraduate students participated in this study in exchange for course credit. However, 33 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. There were four experimental conditions. The "adult story" condition contained the same video footage as in *Experiment 2*. At the end of the original discussion (about a true/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.

The final two conditions featured 4- to 7–year-old children telling stories about real or fabricated events. 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?"). Footage of the children discussing the events was compiled for the "child story" condition. As with the adults, children were matched such that one child's true story was another child's false story. 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 were used in the "child question" condition.

There were 96 clips produced for the "adult story" (M length = 98.13 seconds, SD = 34.30) and "adult question" (M length = 9.54 seconds, 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 story" (*M* length = 85.17 seconds, SD = 34.36) and "child question" (*M* length = 15.90 seconds, 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. Also, participants were asked to indicate their level of confidence in their decisions as a percentage (from 0% to 100% confident).

Procedure

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

Results

Lie Detection Reliability

A correlational 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 Story," $r(48) = 0.11$, $p = .47$, "Child Question," $r(48) = -0.08$, $p = .59$, "Adult Story," $r(48) = -0.02$, $p = .90$, and "Adult Question," $r(49) = -0.00$, $p = .98$, clips.

Again, the reliability of performance was examined in terms of sex differences. There was no significant correlation between accuracy in the first and second sessions for women, $r(163) = -0.00$, $p = .93$, or men, $r(32) = 0.02$, $p = .91$.

Lie Detection Accuracy

The range in accuracy appeared similar in session one (0% to 79% correct) and session two (8% to 79% correct).

A Participant Sex x Target age (Children vs. Adults) x Interview type (Question vs. Story) x Statement type (Lie vs. Truth) x Session (Session 1 vs. Session 2) mixed-factors ANOVA (with the last two variables as the repeated measures) was performed on participant's accuracy scores. Participants were significantly better able to identify truth-telling ($M = .47$, $SD = .15$) than lie-telling ($M = .40$, $SD = .15$), $F(1, 189) = 5.67$, $p < .05$, $r = .17$. In addition, answers to yes-no questions ($M = .44$, $SD = .11$) were more accurately identified than full narratives ($M = .41$, $SD = .11$), $F(1, 189) = 4.06$, $p < .05$, $r = .15$. However, this was qualified by several significant interactions. There was a significant interaction between statement type and session, $F(1, 189) = 4.59$, $p < .05$, $r = .15$. Specifically, participants were worse at detecting lie-tellers in the second session when compared to the first; detection of truth-tellers did not differ across sessions. In addition, there was a significant interaction between statement type and participant sex, such that men performed equally well when detecting lie-tellers and truth-tellers; women were much poorer at detecting lies than truths, $F(1, 189) = 4.38$, $p < .05$, $r = .15$. Finally, there was a significant interaction between the type of statement and the type of interview, $F(1, 189) = 5.60$, $p < .05$, $r = .17$. Participants were equally able to identify deceptive and truthful responses to yes-no questions; when viewing full narratives, participants were less accurate at detecting lie-telling than truth-telling (Figure 2). There were no other significant main effects or interactions.

One-sample t-tests compared accuracy to chance. Accuracy was significantly below chance in session one, $t\,(196) = -6.58, p < .001$, and session two, $t\,(196) = -7.13, p < .001$. Overall, participants' performance was significantly below chance when viewing "Child Story," $t\,(48) = -6.15, p < .001$, "Child Question," $t\,(48) = -4.21, p < .001$, "Adult Story," $t\,(48) = -5.52, p < .001$, and "Adult Question," $t\,(49) = -3.67, p < .01$ clips. Participants' tendency to identify truth-tellers at a level below chance was significant, $t(196) = -1.95, p = .05$; they correctly identified lie-tellers below the level of chance, $t(196) = -9.45, p < .001$. Both women, $t(163) = -8.47, p < .001$, and men, $t(32) = -4.76, p < .001$, performed below the level of chance.

Signal Detection Analyses

Discrimination. A Participant sex x Target age x Interview type x Session mixed-factors ANOVA was performed on the d' values (i.e., discrimination). Unlike in *Experiment 2*, women ($M = -.38$, $SD = .65$) were no better able to differentiate between lie-tellers and truth-tellers than men ($M = -.36$, $SD = .52$), $F(1, 189) = 0.08, p = .78$. There were no significant main effects or interactions.

One-sample t-tests were performed to see whether observers could actually differentiate between the adults during each session. Women, $t(164) = -7.63, p < .001$, and men, $t(32) = -3.93, p < .001$, could discriminate between lie-tellers and truth-tellers. However, as indicated by the negative d' values they tended to mislabel the targets. Overall, participants could differentiate between the groups in session one, $t(196) = -5.98$, $p < .001$, and two, $t(196) = -6.84, p < .001$.

Figure 2.

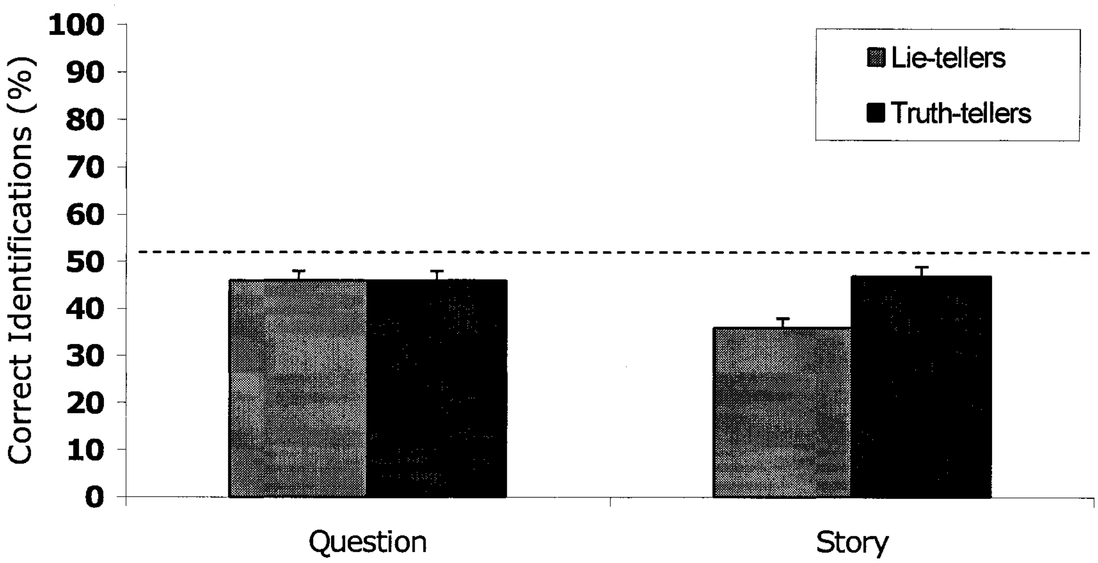

Figure 2. Average percentage of correct identifications as a function of type of interview and statement.

Response Bias (β). A Participant sex x Target age x Interview type x Session mixed-factors ANOVA was performed observers' tendency to indicate that adults were lie-tellers (or truth-tellers) during each session. Overall, there were no significant main effects or interactions. The interaction between session and participant sex approached significance, $F(1, 189) = 6.96, p = .06$. In addition, the interaction between session, target age, and interview type was marginally significant, $F(1, 189) = 3.60, p = .06$.

Using t-tests, each ß was compared to 1 (no bias). In session one, observers tended to indicate that targets were telling the truth ($M = 1.22, SD = 1.45$), $t(196) = 2.18$, $p < .05$. This truth bias also occurred in session two ($M = 1.24, SD = 1.43$), $t(196) = 2.35$, $p < .05$.

Binomial theorem analysis

Again, for descriptive purposes, each individual's performance was compared to chance using a binomial theorem analysis. Performance was significantly above chance if 17 or more of the participants' decisions were correct and significantly below chance is 7 or fewer decisions were correct. The majority of participants ($n = 115$) performed at chance (i.e., scoring between 8/24 and 16/24) during both sessions (see Table 3).

Confidence

A Participant sex x Target age x Interview type x Statement type x Session mixed-factors ANOVA was performed participants' confidence ratings. Women ($M = .67, SD = .11$) expressed more confidence in their decisions than men ($M = .63, SD = .09$), $F(1, 189) = 3.90, p = .05$. This was qualified by a higher-order interaction between participant sex, target age, and interview type, $F(1, 189) = 8.29, p < .01$. When women rated children's deception, they were more confident when viewing yes-no responses

Table 3

Distribution of Participants' Accuracy during Sessions One and Two in Experiment 3.

	Session One		
	Below chance	At chance	Above chance
		Adult Question Condition	
Session Two			
Below chance	2	6	0
At chance	10	23	4
Above chance	2	3	0
		Adult Story Condition	
Session Two			
Below chance	3	7	0
At chance	5	30	3
Above chance	0	1	0
		Child Question Condition	
Session Two			
Below chance	3	8	0
At chance	1	32	3
Above chance	1	1	0
		Child Story Condition	
Session Two			
Below chance	2	8	1
At chance	6	30	1
Above chance	0	1	0

than narratives; however, when rating adults, women were similarly confident when viewing yes-no responses and narratives. The opposite was true for men (i.e., men were equally confident when rating children's yes-no responses and narratives, but more confident when rating adult's yes-no responses than narratives) (Figure 3). There were no other significant main effects or interactions.

Additional analyses revealed that there was a significant correlation between average confidence in sessions one and two, $r(196) = 0.82, p < .001$. Specifically, correlation analyses revealed a significant relationship between confidence in sessions one and two when participants viewed "Child Story," $r(48) = 0.72, p < .001$, "Child Question," $r(48) = 0.84, p < .001$, "Adult Story," $r(48) = 0.83, p < .001$, and "Adult Question," $r(49) = 0.84, p < .001$, clips. In addition, there was no significant relationship between confidence and accuracy in session one, $r(196) = .10, p = .17$, or session two, $r(196) = .05, p = .51$.

Figure 3.

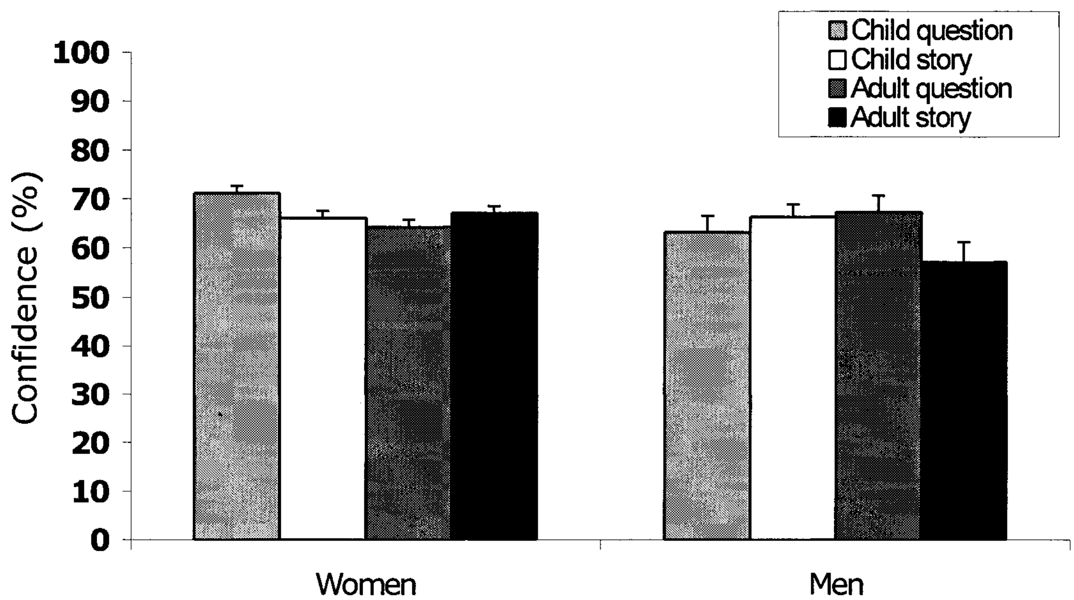

Figure 3. Average confidence as a function of participant sex, target age and type of interview.

Discussion

Again, participants had difficulty classifying truth-tellers and lie-tellers. Overall, performance was below chance. In particular, the detection of lie-telling was poorer for women, for those who viewed narratives, and in the second session (vs. the first). A significant truth bias may account for the low levels of lie detection. In addition, signal detection analyses revealed that individuals were able to discriminate between lie-tellers and truth-tellers, but they tended to mislabel them (i.e., indicate that truth-tellers were lying and vice versa). Regardless of the age of the deceiver or the type of interview, lie detection performance was not reliable. Although confidence was stable over time, there was no relationship between accuracy and confidence in any of the sessions. In keeping with their differences in accuracy, men and women differed in their confidence in their decisions.

Signal detection analyses provided a revealing perspective on participants' poor performance. The presence of a truth bias (i.e., individuals' increased likelihood of indicating that targets were telling the truth) offers the most intuitive explanation for the unexpectedly low number of correct decisions: if participants indicated that lie-tellers were telling the truth, accuracy would suffer. However, it could be argued, that a large truth bias would lead to chance performance because accurate identification of truth-tellers would be high (and, perhaps, counter low lie detection accuracy). Instead, an examination of individuals' ability to discriminate between lie-tellers and truth-tellers may account for low overall accuracy. Specifically, participants could sense that there was a difference between the two groups, but that did not translate into correct decisions due to labelling errors (e.g., a lie-teller was called a truth-teller). Previous research

(Leach et al., 2004) suggests that police officers also share this pattern of response. It may be that individuals were employing an effective set of cues inappropriately. For example, changes in pupil dilation can be associated with deception (DePaulo et al., 2003), but participants may have thought that increased dilation was indicative of truth-telling. Thus, a useful cue may have been identified and utilized, albeit unsuccessfully.

Although this experiment was designed to account for the differences in reliability between *Experiments 1* and *2*, it did not. Neither varying the age of the deceiver, nor the type of interview led to changes in the reliability of performance (i.e., lie detection was not stable under any of these conditions). It is possible that the observed reliability in *Experiment 1* was simply due to random error. Two experiments, with sizeable sample sizes, have suggested that performance is not reliable. In addition, the present experiment featured children's yes-no responses (i.e., stimuli that were quite similar to *Experiment 1*) and the reliability effect could not be replicated. However, there exists a plausible explanation for differences between *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 actions and responses were completely volitional). All 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 more accurate than the detection of experimentally-manipulated deception (e.g., Mann et al., 2004). It is reasonable to expect that there might be corresponding effects on reliability. *Experiment 4* was designed to explore this issue.

Chapter 5

Experiment 4

This experiment explicitly examined the differences between naturalistic and experimentally-manipulated deception. Previous experiments have focused on either the former or latter type of deception (e.g., Ekman et al., 1991; Mann et al., 2004). Given that the two contexts have never been directly compared, it is unknown how they might differ. It is expected that naturalistic deception might be easier to detect because it more closely approximates what individuals encounter daily. In addition, naturalistic lie-tellers might have more difficulty concealing their deception due to motivational constraints. The lies of individuals who are highly motivated to deceive are easier to detect (e.g., DePaulo & Kirkendol, 1989). Thus, an individual who thinks that he/she will be caught cheating on a test might be more motivated to escape detection (and lie) than someone who is given explicit permission to cheat and lie about it. Beyond differences in lie detection accuracy associated with different types of deception, effects on the reliability of performance are the primary focus. Due to the reasons listed above (i.e., realism and the motivational impairment effect) and the findings of Experiments *1 – 3,* lie detection performance is expected to be reliable for naturalistic deception, but not experimentally-manipulated deception.

Method

Participants

A total of 42 undergraduate students participated in this study. However, 7 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 footage was obtained using a temptation resistance paradigm. Adults' upper bodies and faces 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 anagrams ranged in difficulty from common words (e.g., sugar) to those that were more obscure (e.g., jocose) (see Appendix B for a copy of the anagram test). The experimenter provided one of two incentives for high performance. Some individuals were told that there was a monetary reward (i.e., $50) for the highest score; others were informed that if they scored over 15/20, they could leave early and that most of their peers had achieved this standard. Individuals were told that they had ten minutes to complete the exam and that they would be left alone in the room during that time. As the experimenter left the room, she instructed participants to continually refer to the clock and to take a pencil from the box on the table (which was placed upon 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?"

To produce naturalistic and experimentally-manipulated scenarios, 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). To produce experimentally-manipulated scenarios, the same procedure was used. However, prior to the female experimenter entering the room, 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.

In all, there were 56 video clips used in *Experiment 4*: 28 involved naturalistic scenarios (M length = 7.44 seconds, SD = 1.97) and 28 featured experimentally-manipulated scenarios (M length = 8.91 seconds, 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. Also, participants were asked to indicate their level of confidence in their decisions as a percentage (from 0% to 100% confident).

Procedure

The procedure was identical to the previous experiments, except that the study was run in a classroom setting.

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 this 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

A correlational 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, $r(34) = -0.17$, $p = .34$, scenarios.

Again, the reliability of performance was examined in terms of sex differences. There was no significant correlation between accuracy in the first and second sessions for women, $r(27) = 0.08$, $p = .69$, or men, $r(6) = -0.64$, $p = .12$.

Lie Detection Accuracy

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

A Participant Sex x Scenario (Naturalistic vs. Experimentally-manipulated) x Statement type (Lie vs. Truth) x Session (Session 1 vs. Session 2) mixed-factors ANOVA (with the last two variables as the repeated measures) was performed on participant's accuracy scores. A main effect of statement type approached significance, $F(1, 33) = 3.53$ $p = .07$, as did the interaction between participant sex and statement type, $F(1, 33) = 3.18$, $p = .08$. There was a significant interaction between the type of statement and the type of deception, $F(1, 33) = 5.32$ $p < .05$, $r = .37$. Specifically, participants were equally accurate when rating naturalistic and experimentally-manipulated truth-telling; they were less accurate when rating naturalistic deception than experimentally-manipulated deception. However, this was qualified by a higher-order interaction between the session, statement, and scenario, $F(1, 33) = 19.76$, $p < .001$, $r = .61$ (Figure 4). During the first session, participants were more accurate when detecting experimentally-manipulated (vs. naturalistic) deception; however, they were better able to detect naturalistic (vs. experimentally-manipulated) truth-telling. In the second session, there were no differences between naturalistic and experimentally-manipulated lie-telling or truth-telling. There were no other significant main effects or interactions.

One-sample t-tests compared accuracy to chance. Accuracy was at chance in session one, t (34) = 0.51, $p = .62$, and session two, t (34) = 1.03, $p = .31$. Participants correctly identified naturalistic scenarios at chance levels, $t(34) = -0.26$, $p = .80$. However, ratings of experimentally-manipulated scenarios were more accurate than expected by chance, $t(34) = 2.03$, $p = .05$.

Figure 4.

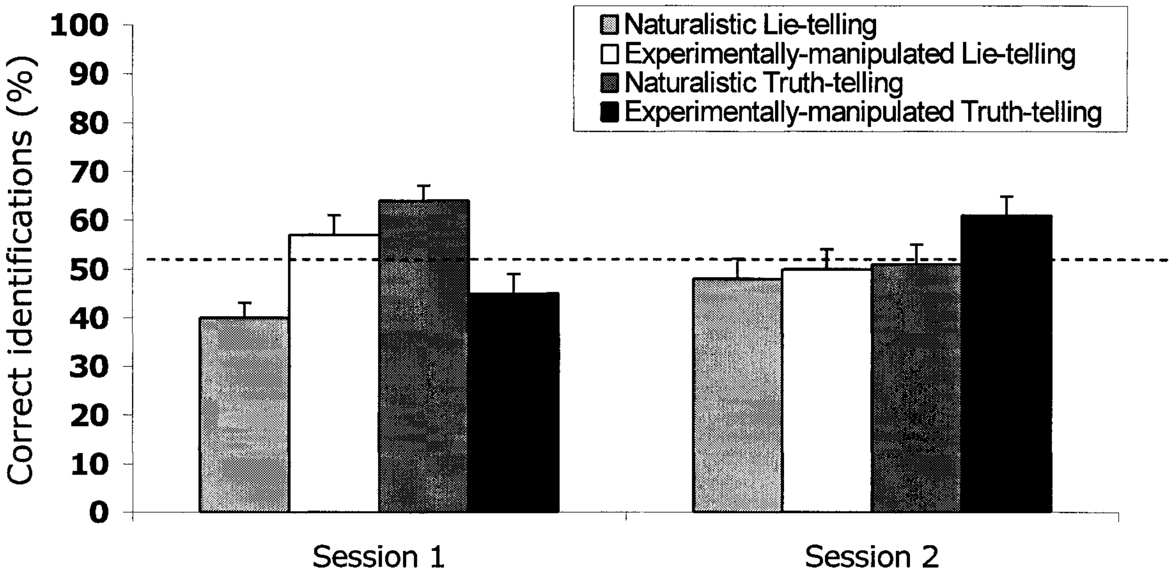

Figure 4. Average percentage of correct identifications as a function of session, type of statement, and scenario.

Signal Detection Analyses

Discrimination. A Participant Sex x Scenario x Session mixed-factors ANOVA was performed on the d' values (i.e., discrimination). There were no significant main effects or interactions.

One-sample t-tests were performed to see whether observers could actually differentiate between lie-tellers and truth-tellers during each session. Overall, participants could not differentiate between the groups in session one, $t(34) = 0.67$, $p = .44$, or two, $t(34) = 1.03$, $p = .31$. In addition, discrimination was poor for naturalistic, $t(34) = -0.42$, $p = .68$, and experimentally-manipulated scenarios, $t(34) = 1.50$, $p = .14$.

Response Bias (ß). A Participant Sex x Scenario x Session mixed-factors ANOVA was performed on observers' tendency to indicate that adults were lie-tellers (or truth-tellers) during each session. Overall, there were no significant main effects or interactions.

Using t-tests, each ß was compared to 1 (no bias). In both session one, $t(34) = 0.35$, $p = .73$, and two, $t(34) = -0.48$, $p = .63$, there was no evidence of bias. Participants' responses were not biased when they viewed naturalistic, $t(34) = -0.18$, $p = .86$, or experimentally-manipulated, $t(34) = -1.40$, $p = .17$, scenarios.

Binomial theorem analysis

Again, for descriptive purposes, each individual's performance was compared to chance using a binomial theorem analysis. For each type of scenario, performance was significantly above chance if 11 or more of the participants' decisions were correct and significantly below chance if 3 or less of the participants' decisions were correct. The majority of participants ($n = 115$) performed at chance (i.e., scored between 4 and 10 out of 14) during both sessions (see Table 4).

Confidence

A Participant Sex x Scenario x Statement type x Session mixed-factors ANOVA was performed on participants' confidence ratings. Participants were more confident in session one ($M = .61$, $SD = .12$) than session two ($M = .58$, $SD = .15$), $F(1, 33) = 7.15$, $p = .42$. There were no other significant main effects or interactions.

Additional analyses revealed that there was a significant correlation between average confidence in sessions one and two, $r(34) = 0.91$, $p < .001$. Specifically, correlation analyses revealed a significant relationship between confidence in sessions one and two when participants viewed naturalistic, $r(34) = 0.79$, $p < .001$, and experimentally-manipulated, $r(34) = 0.89$, $p < .001$, scenarios. In addition, there was no significant relationship between confidence and accuracy in session one, $r(34) = -.07$, $p = .70$, or session two, $r(34) = .15$, $p = .40$.

Table 4

Distribution of Participants' Accuracy during Session One and Session Two in Experiment 4.

	Session One		
	Below chance	At chance	Above chance
	Naturalistic Scenarios		
Session Two			
Below chance	0	1	0
At chance	0	32	0
Above chance	0	2	0
	Experimentally-manipulated Scenarios		
Session Two			
Below chance	0	0	0
At chance	0	33	1
Above chance	0	1	0

Discussion

As observed in all previous experiments, participants had difficulty identifying truth-tellers and lie-tellers (i.e., performance was at chance). Overall, individuals had more difficulty detecting naturalistic lie-telling than experimentally-manipulated lie-telling. However, this finding did not emerge when using a more stringent analysis (i.e., signal detection theory). This analysis also failed to reveal any bias. Regardless of the type of scenario (naturalistic or experimentally-manipulated), lie detection performance was not reliable. Although confidence decreased across sessions, average confidence was stable. Again, there was no relationship between accuracy and confidence in any of the sessions.

Chapter 6

Consideration of the Stimuli

Although the proficiency of targets' lie production was not a primary focus of this research, it is a related issue. Some researchers specifically select targets 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 on the average proportion of correct identifications associated with each target. Binomial theorem analyses allowed for a stringent test of whether targets were classified below, above, or at chance levels: for the most part, the stimuli were correctly identified at chance levels (see Table 5). Despite the underreporting of this type of information in studies of lie detection, it might provide an important clue to performance. Thus, a closer examination of how the inclusion of particular targets affects lie detection could be beneficial in future research.

Table 5

Proportions of Targets Identified at Various Levels of Accuracy in Experiments 1-4.

	Experiment			
	1	2	3	4
Correct Identifications (as lie-tellers or truth-tellers)				
0-10 %	*.09*	*.00*	*.01*	*.00*
11-20 %	*.07*	*.04*	*.05*	*.11*
21-30 %	*.06*	*.10*	*.10*	*.05*
31-40 %	.11	.18	.17	.13
41-50 %	.16	.24	.18	.18
51-60 %	.11	.13	.19	.20
61-70 %	.13	.11	.12	.14
71-80 %	*.17*	*.15*	*.09*	*.14*
81-90 %	*.10*	*.03*	*.07*	*.05*
91-100 %	*.00*	*.02*	*.02*	*.00*

Note. All italicized proportions represent levels of accuracy that were significantly different from chance.

Chapter 7

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). Unexpectedly, lie detection performance did not appear to be reliable in the majority of experiments. Accuracy was stable when individuals classified children denying having committed a transgression (*Experiment 1*). However, this finding did not generalize (i.e., 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* suggested that the type of scenario (i.e., experimentally-manipulated or naturalistic deception) could 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. If this is true, the boundaries of reliability are still unknown. *Experiment 1* suggests that children's yes-no responses to questions about resisting temptation (by not peeking at a toy) elicit stable performance. *Experiments 2-4* provide a wider number of situations that do not. Future research may wish to address this issue and explore other circumstances under which performance is, or is not, reliable. In addition, the implications of this explanation must be considered. The most judicially relevant variables (specifically, adults' and children's narratives) produced unreliable performance. Certainly, it is difficult to elicit complete accounts from children (e.g., Ceci & Bruck, 1995) and there might be a few instances in which officials would have to differentiate between children's yes-no responses. However, officials are much more likely to be interested in information about others' transgressions than those of the children. Thus, if performance is only reliable under specific conditions that are rarely encountered by the majority of law enforcement officials, although empirically interesting, the applications are quite limited.

A second possibility is that the reliable performance observed in *Experiment 1* was simply a fluke. There is overwhelming evidence, provided by all of the other experiments, that performance is not reliable. Even when similar stimuli were used, whether it be children's closed-ended experimentally-manipulated responses or adults' yes-no responses after an actual transgression, the reliability effect could not be replicated. If performance is not reliable – under any circumstances – there are significant legal and theoretical implications. For example, the notion that law enforcement officials can be chosen based on their initial competence, with the assumption that ability will remain constant, would be completely incorrect. If individual differences are not reliable, it may not be sensible for organizations (e.g., police forces, the FBI, and the CIA) to develop tests to identify the degree to which their personnel have this ability. More importantly, a lack of reliability might severely affect training. There is evidence that experience with lie detection, including training, 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, training programs' claims that they dramatically improve lie detection performance (e.g., Inbau et al., 2001) could be unrealistic because they rely upon the notion that decision-making, in the context of lie-detection, is stable. 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 for researchers to interpret those findings with caution.

Another concern raised by this research is the significant variability that was elicited by different types of lie detection paradigms and stimuli. Generally, the present experiments support previous observations that most individuals have difficulty detecting deception (e.g., DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991; Etcoff et al., 2000). This finding is also observed in the clinical judgment literature (e.g., Ambady & Rosenthal, 1992). In keeping with other research (e.g., Lewis et al., 1989; Leach et al., 2004), 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 clips featuring naturalistic deception are comparable to those that are experimentally-manipulated because they both result in chance-level performance (e.g., Leach et al., 2004), the present experiments suggest otherwise. Namely, in *Experiment 4*, individuals were more accurate when viewing experimentally-manipulated lie-telling and naturalistic truth-telling. It could be argued that 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 behaviours. Yet, the majority of researchers (e.g., Ekman et al., 1999; Feldman, 1979) base their conclusions on experimentally-manipulated deception. Finally, the type of interview (closed-ended vs. open-ended) also led to different levels of accuracy. As in previous research (Vrij & Baxter, 1999), individuals more accurately identified truth-telling than lie-telling in narratives. In a departure from previous findings, observers in the present experiment were equally accurate when viewing truthful and deceptive yes-no responses (whereas the other researchers found higher accuracy for detecting deception 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. Although many different aspects of performance were affected, there were three main areas of concern: (1) sex differences; (2) truth biases; (3) the relationship between accuracy and confidence.

The mixed findings regarding sex differences should be noted. There are a few cases (e.g., Forrest & Feldman, 2000; Leach et al., 2004) in which women have outperformed men on lie detection tasks. As noted earlier, women are better able to identify what the speaker intends to convey (Rosenthal and DePaulo, 1979). However, some researchers (e.g., Vrij & Mann, 2001) have argued that this would compromise the ability to detect deception, as the lie-teller intends to communicate a falsehood. Women's lower levels of suspicion (DePaulo et al., 1993) would also be expected to decrease their lie detection accuracy. Although decreased lie detection accuracy was observed in women during *Experiment* 3, women were more accurate than men in *Experiment 2*. Given the absence of sex differences in the majority of studies to date (Zuckerman et al., 1981) - including the other experiments and analyses in this program of research - it is possible that these sex effects were due to random error. At the very least, the present experiments illustrate the conflicting results associated with examinations of sex differences in lie detection performance.

In addition, the experiments offer conflicting evidence of a truth bias. That is, in *Experiments 2* and *3,* individuals were better able to detect truth-telling than lie-telling. Researchers have reported cognitive and behavioural truth biases when assessing friends' deception (Stiff, Kim, & Ramesh, 1992). Other researchers have also reported strangers' tendency to label adults (Zuckerman et al., 1981) and children (Leach et al., 2004) as truth-tellers. Moreover, Vrij and Baxter (1999) report a truth bias for elaborations (i.e., longer narratives) and a lie bias for denials. The present findings are somewhat consistent with their conclusions, as individuals more accurately identified truth-tellers when they rated full accounts (*Experiment 2*) *and* yes-no responses (*Experiment 3*). However, perplexingly, signal detection analyses failed to reveal concrete evidence of a truth bias. Thus, the two approaches used to analyse deception detection (i.e., overall accuracy vs. signal detection analysis) offer contradictory results. Findings related to a truth bias should be treated with caution.

Examining the relationship between accuracy and confidence offers another perspective on performance. Consistent with previous research (DePaulo et al., 1997), the majority of the present experiments failed to reveal an overt relationship between correct classifications of lie-telling and truth-telling and self-reported confidence. Yet, in several experiments, there was a slight decrease in confidence across sessions. These findings suggest that individuals were attuned to the difficulty of the task (and, possibly, their poor performance). Also, other research has shown that individuals who are asked to make intuitive lie detection decisions, without the aid of context or evidence, can become less confident over time (Leach, Douara, & Lee, 2003). Thus, there may be a link between confidence and participation in lie detection tasks where veridical feedback is available. In fact, these findings suggest a method to alleviate the overconfidence observed in many front-line workers. Previous research has shown that, despite comparable performance, law enforcement officials tend to be more confident than untrained observers (DePaulo & Pfeifer, 1986; Frank & Ekman, 1997; Leach et al., 2004). This overconfidence may pose problems in the justice system if officers use their level of confidence to gauge a suspect's guilt or innocence. In turn, one implication of these findings is that law enforcement officials may benefit from exposure to controlled lie detection stimuli (as opposed to work-related interactions) and veridical feedback about the accuracy of their decisions. Finally, the stability of confidence across sessions suggests that people, while not well calibrated to their accuracy level, do show habitual behaviour regarding expressed confidence.

Although this program of research was one of the first to explicitly examine the reliability of lie detection performance, further empirical research is needed. There may be some concern about the type of deception employed. The majority of the lies were about trivial events (e.g., peeking at a toy, breaking a limb). Given that researchers have suggested that highly motivated lie-tellers are less effective (DePaulo & Kirkendol, 1989), viewing insignificant deception could affect lie detectors' decision-making. It could be argued that being accused of cheating on a test (*Experiment* 4) would have been a stressful and meaningful event for undergraduate students. However, these students were fully aware that they were participating in a psychology study and they may have considered the ramifications associated with copying anagrams to be relatively minor (as compared to being accused of academic dishonesty). Anecdotal evidence suggests that the children who participated in the present experiments were extremely concerned about the consequences associated with deception (e.g., some children refused to lie when asked to tell made up stories even though it was endorsed by their parents). It could be argued that deception was considered "high-stakes" for this group. Increasing the stakes associated with deception should increase lie detection rates by making deception more transparent. This may affect the reliability of performance because cues to deception that were not available in the current research could be observable in such instances. Thus, a pressing matter is whether the types of deception are representative of those seen in the justice system. The scenario that might have been the most legally applicable involved individuals who were coached to lie or tell the truth about an event. Even then, it does not fully capture the circumstances that are present in legal contexts (e.g., when an individual professes innocence to a murder). Thus, the generalizability of the present findings may be limited and reliability must be examined under more judicially relevant conditions.

Another potential limitation was the nature of the lie detection task. In these experiments, observers simply watched another person interview the targets. Law enforcement officials are unlikely to be constrained in this way (i.e., they can ask their own questions and follow up on any inconsistencies). Intuitively, increased task involvement should improve performance. However, active participants are actually less accurate than observers (e.g., Buller, Strzyzewski, & Hunsaker, 1991; Dunbar, Ramirez, & Burgoon, 2003). Thus, the present research may have actually overestimated real-world lie detection accuracy. It is unknown how reliability could be affected by increased task involvement, but it provides an interesting avenue for future research.

Similarly, the types of lie detectors who were recruited for this research could have affected the findings. Specifically, the majority of the participants performed at, or near, the level of chance. Perhaps it is unrealistic to expect these unskilled individuals to exhibit stable performance. Instead, research might focus on individuals who are proven, proficient lie detectors. Ekman and O'Sullivan (2004) have identified "wizards" who are extremely accurate at detecting deception. However, limitations with their approach should be addressed with additional research. For one, individuals who were initially recruited were self-selected (e.g., they volunteered that they were superior lie detectors). It is unknown whether self-assessments of lie detection performance are predictive of ability. In addition, participants were recruited primarily from workshops on lie detection. Given this focus on a very narrow sample, there is little sense of the distribution of proficient (and reliable) lie detection performance within the general population (although, the present experiments indicate that it is very low). It would prove interesting to study 'wizards' to see how and why their performance differs from that of most other individuals.

More generally, the present research raises two major questions: (1) Is performance reliable? (2) Why is performance (un)reliable? Considerable inroads have been made on the first issue. Although the boundary conditions remain unclear, it seems likely that performance is unreliable under most circumstances. However, the second question remains unanswered. If performance is found to be reliable, it may be based on cue use. That is, individuals may employ the same lie detection strategies over time. If appropriate cues are used, then performance 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 behaviours that are indicative of deception (Akehurst et al., 1996; Vrij et al., 2001; Vrij & Semin, 1996). It might seem unlikely that these strategies would change drastically over a week (the time between testing sessions), suggesting that it would not be surprising that performance relying on specific cues would remain stable. However, the lack of reliable performance observed in the majority of the present experiments does not necessarily indicate that cue use was not stable. An alternative explanation is that individuals employed *irrelevant* cues consistently, producing the chance-level accuracy (and random variations in performance) that has been widely observed. It is also 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 intuitive lie detectors via several different approaches. The majority of researchers ask participants to identify relevant lie detection strategies from a list (e.g., Vrij & Semin, 1996) or after a lie detection task (e.g., Anderson et al., 1999; Forrest et al., 2004). Although both of these approaches appear reasonable, it is unknown whether such reports are reliable indices of actual behaviour. Respondents may erroneously indicate that they employed a particular strategy for many reasons (e.g., implicit causal theories, lack of insight) other than accurate self-report (e.g., Nisbett & Wilson, 1977). Thus, a combination of stimuli selection (i.e., choosing targets who are known to exhibit particular cues), self-report, and behavioural validation (e.g., tracking respondents' eye gaze) may be useful. If successful and detrimental 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).

It is still unclear whether intuitive lie detection is a stable characteristic of the person. The fact that the majority of researchers have worked under this assumption for some time 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.

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.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111,* 236-274.

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.

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.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology, 88,* 131-151.

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.

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.

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.

Ceci, S. J., & Huffman, M. L. C. (1997). How suggestible are preschool children? Cognitive and social factors. *Journal of the Academy of Child and Adolescent Psychiatry, 36,* 948-958.

Chisholm, R. M., & Feehan, T. D. (1977). The intent to deceive. *The Journal of Philosophy, 75,* 143-159.

Coleman, L.,& Kay, P. (1981). Prototype semantics: The English word lie. *Language, 75,* 26-44.

DePaulo, B. M. (1994). Spotting lies: Can humans learn to do better? *Psychological Science, 3,* 83-86.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *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, NY, US: The Guilford Press.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. W., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979-995.

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effects in the communication of deception. In J. Yuille (Ed.), *Credibility assessment* (pp. 51-70). 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.

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

DePaulo, B. M., & Tang, J. (1994). Social anxiety and social judgment: The example of detecting deception. *Journal of Research in Personality, 28,* 142-153

DePaulo, B. M., Zuckerman, M., & Rosenthal, R. (1980). Humans as lie detectors. *Journal of Communication, 30,* 129-139.

Dunbar, N. E., Ramirez, A., Jr., & Burgoon, J. K. (2003). The effects of participation on the ability to judge deceit. *Communication Reports, 16*, 23-33.

Ekman, P. (1996). Why don't we catch liars? *Social Research, 63,* 801-817.

Ekman, P. (1997). Lying and deception. In N. L. Stein, P. A. Ornstein, B. Tversky, & C. Brainerd (Eds.), *Memory of everyday and emotional events* (pp. 333-347). Mahwah, New Jersey: Lawrence Erlbaum Associates.

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, 9,* 913-920.

Ekman, P., & O' Sullivan, M. (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.

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

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

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

Feldman, R. S., & White, J. B. (1980). Detecting deception in children. *Journal of Communication, 30,* 121-128.

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.

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

Granhag, P. A., & Strömwall, L. A. (2001). Deception detection based on repeated interrogations. *Legal and Criminological Psychology, 6,* 85-101.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology, 79,* 252-259.

Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie-detector test: Two surveys of scientific opinion. *Journal of Applied Psychology, 82,* 426-433

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.

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioural confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, 27, 187-203.

Kircher, J. C, Horowitz, S. W, & Raskin, D. C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law & Human Behavior, 12,* 79-90.

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.

Köhnken, G. (1987). Training police officers to detect deceptive eyewitness statements: does it work*? Social Behaviour, 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.

Leach, A-M., Douara, D., Lee, K. (July, 2003). *Learning to detect children's deception.* Paper presented at the American Psychology-Law Society International, Interdisciplinary Conference, Edinburgh, Scotland.

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.

Lewis, M., Stanger, C., & Sullivan, M.W. (1989). Deception in 3-year-olds. *Developmental Psychology, 25,* 439-443.

MacLaren, V. V. (2001). A qualitative review of the Guilty Knowledge Test. *Journal of Applied Psychology, 86*, 674-683.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: a review of the empirical evidence. *Psychology, Crime & Law, 11,* 99-122.

Nisbett, R. E., & Wilson, T.D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

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.

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.

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.

R. v. Béland, 2 S.C.R. 398 (1987).

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

Rosenthal, R., & DePaulo, B. M. (1979). Sex differences in eavesdropping on nonverbal cues. *Journal of Personality and Social Psychology, 37,* 273-285.

Ruby, C. L., & Brigham, J. C. (1997). The usefulness of the Criteria-Based Content Analysis technique in distinguishing between truthful and fabricated allegations. *Psychology, Public Policy & Law, 3,* 705-737.

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.

Stiff, J.B., Kin, H. J., & Ramesh, C. N. (1992). Truth biases and aroused suspicion in relational deception. *Communication Research, 19,* 326-345.

Strichartz, A. F., & Burton, R. V. (1990). Lies and truth: A study of the development of the concept. *Child Development, 61,* 211-220.

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.

United States v. Scheffer, 188 S. Ct. 1261 (1998).

Vrij, A., Akhurst, L., Soukara, S., & Bull, R. (2004). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research, 30,* 8-41.

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., Evans, H., Akehurst, L., & Mann, S. (2004). Rapid judgements in assessing verbal and nonverbal cues: Their potential for deception researchers and lie detection. *Applied Cognitive Psychology, 18,* 283-296.

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

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. (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., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology, 89,* 137-149.

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

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.

Zuckerman, M., Koestner, R., & Colella, M.J. (1985). Learning to detect deception from three communication channels. *Journal of Nonverbal Behavior, 9,* 188-194.

Appendix A

Prototypical Approach vs. Propositional Approach

As mentioned earlier, the propositional approach posits that lying only occurs when the speaker makes a solemn assertion to the listener *and* the speaker believes that the statement is false *and* the speaker intends to deceive the listener. According to the prototypical approach, lying is observed when one or more of the following criteria are met: (a) the statement is factually false; (b) the speaker believes that the statement is false; and (c) the speaker intends to deceive the listener. Both the propositional and prototypical approaches agree that a statement that lacks all elements (*Example 8*) is not a lie. The two perspectives also indicate that statements that are believed to be false and are intended to deceive (*Examples 1 and 5*) are lies[1]. However, only the prototypical approach suggests that other scenarios (*Examples 2, 3, 4, 6, and 7*) may depict lying. The eight different logical possibilities are illustrated below.

	False		**Not False**	
	Believed False	**Not Believed False**	**Believed False**	**Not Believed False**
Intended to Deceive	*Example 1*	*Example 2*	*Example 5*	*Example 6*
Not Intended to Deceive	*Example 3*	*Example 4*	*Example 7*	*Example 8*

Example 1: False - Believed to be False - Intended to Deceive

Jennie eats a bag of chips. Chris asks Jennie, "Did you eat the bag of chips?" Jennie says, "No."

[1] This is true only if the speaker has made a solemn assertion to the listener. Without that element, none of these examples would be lies according to the propositional approach.

Example 2: False – Not Believed to be False – Intended to Deceive

Jennifer thinks that she must pass the bank on the way to a restaurant; however, the bank has moved. Duane does not approve of her eating out. As she is about to leave, to go to the restaurant, Duane asks her where she is going. She says, "I am going by the bank."

Example 3: False – Believed to be False – Not Intended to Deceive

Vicky and Phil go on a blind date together. They both have a terrible time. It is clear that neither wishes to see the other again. At the end of the evening, Vicky says to Phil, "I'll call you," even though she has no intention of doing so. She isn't really trying to convince him that she'll call, but it is what she always says at the end of a date, even if she doesn't expect him to believe it.

Example 4: False – Not Believed to be False – Not Intended to Deceive

Caroline and Isabelle are sitting in their office. There are no windows in the office. Caroline asks Isabelle if it is raining outside. It was raining the last time that Isabelle looked outside, and meteorologists had forecast a full day of rain, so she replies, "Yes." However, it has since cleared up and it is no longer raining.

Example 5: Not False – Believed to be False - Intended to Deceive.

Rupinder doesn't feel like taking out the garbage today – even though he knows that is the regular 'garbage day.' Meg asks, "Is garbage being picked up today?" Rupinder says, "No." To Rupinder's surprise, the garbage is not picked up because the workers are on strike.

Example 6: Not False – Not Believed to be False – Intended to Deceive

Annette asks Iannick if he likes her cowboy hat. Iannick really dislikes the hat. He says to Annette, "Cowboy hats are really in style right now." In fact, currently, cowboy hats are considered fashionable.

Example 7: Not False – Believed to be False – Not Intended to Deceive

Karuna has a wedding to attend on Thursday. Her boss knows that she has to attend this wedding and says, "I can't justify giving you the day off. You are going to have to come up with a good excuse if you miss a day of work." The day of the wedding, Karuna calls her boss and says, "I can't come in to work because I have to go to the dentist." Ironically, Karuna doesn't get to go to the wedding because the slight toothache that she had that morning turns out to be a chipped tooth that must be repaired immediately by her dentist.

Example 8: Not False – Not Believed to be False – Not Intended to Deceive

Jennie eats a bag of chips. Chris asks Jennie, "Did you eat the bag of chips?" Jennie says, "Yes."

Appendix B

TAT ©

Timed Anagram Test

©

VERSION B

IMPORTANT

When completing the Timed Anagram Test ©, you are free to skip questions that you might have difficulty answering and to come back to them later during the time provided. You may also change the answer that you recorded on the answer sheet by erasing it completely and writing in your desired answer. **Access to external aids is strictly forbidden.**

Your score will be determined by the number of questions of which you chose the correct answer. In addition, please note that you will be penalized for all unanswered questions and incorrect responses.

Work as rapidly as you can without being careless. Although you may revise your answers at any time, you must stop writing immediately once the administrator states, "please stop work."

TAT – VERSION B

Directions: Each word jumble can be unscrambled to form a common word. There is a correct answer for each jumble. You will have ten minutes in which to complete the test.

1. UGARS: _____(SUGAR)_____
2. AGMNO: _____(MANGO/AMONG)_____
3. IBNNGE: _____(BENIGN)_____
4. FKSFI: _____(SKIFF)_____
5. EEDNU: _____(ENDUE)_____
6. RPUEYN: _____(PENURY)_____
7. BCEHA: _____(BEACH)_____
8. OEDML: _____(MODEL)_____
9. KLRCE: _____(CLERK)_____
10. UFLAT: _____(FAULT)_____
11. OLBDUE: _____(DOUBLE)_____
12. NITGA: _____(GIANT)_____
13. AACETV: _____(CAVEAT/VACATE)_____
14. NCEGOT: _____(COGENT)_____
15. ELEVD: _____(DELVE)_____
16. MLYIOH: _____(HOMILY)_____
17. MLUTTU: _____(TUMULT)_____
18. HMUOT: _____(MOUTH)_____
19. EORTT: _____(TORTE/OTTER)_____
20. ESJCOO: _____(JOCOSE)_____