REPORT DOCUMENTATION PAGE

Form Approved OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	20070425	FINAL REPORT 20010715 - 20061031

4. TITLE AND SUBTITLE
Detecting Deception in the Military Infosphere:
Improving and Integrating Human Detection Capabilities with Automated Tools

5. FUNDING NUMBERS
F49620-01-1-0394

6. AUTHOR(S)
Judee K. Burgoon
Jay F. Nunamaker, Jr.
Joey George
David Biros

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for the Management of Information
Eller College of Management
University of Arizona
Tucson, AZ 85721

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AF Office of Scientific Research
875 N. Randolph Street, Room 3112
Arlington, VA 22203
Dr Robert Herklotz/NM

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-SR-AR-TR-07-0193

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This 5-year project conducted principally by University of Arizona, Florida State University, Michigan State University, and Air Force Institute of Technology, reports results of (1) theoretical development on a model of interactive deception, (2) laboratory and field testing to identify reliable indicators of deceit and variables that moderate those effects, (3) identification of cognitive biases that adversely impact human deception detection, (4) development of a prototype suite of tools—*Agent99*—for automatically identifying linguistic, vocal, and visual indicators of deceit, (5) development of a curriculum and computer-based system for delivering training in deception detection, and (6) field testing of the prototypes and training. Several tools have been implemented, and lessons learned provide important guidance for future development of deception detection tools and procedures.

14. SUBJECT TERMS
Deception Detection, Critical Infrastructure, Electronic Communication

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Executive Summary

Problem Statement

Throughout history, effective information and communication systems have been key enablers of successful combat and peacetime missions. Yet the sheer volume, complexity, and speed of information transmission and communication, and our lack of knowledge of unintended as well as intended capabilities of newly adopted tools and systems, pose accompanying risks to information assurance. Risks include biased and erroneous intelligence, inability to fuse data and ideas into operational concepts, inadequate assessment of alternative interpretations, and faulty, even catastrophic, decision-making if implementation is not accompanied by deeper understanding, training, and tools to mitigate these risks.

The focus of the research to be reported herein is deception and its detection. The complexity of the task cannot be overstated. Extensive social science research has confirmed that humans are very adroit at dissembling yet very poor at detecting it. Thus, whenever humans are involved as information sources, conduits, or recipients, the risk of undetected deception and false alarms is unacceptably high, and may become magnified when messages and information are derived from IT artifacts such as computers and networks.

Although it is seductively appealing to try to replace human detectors with completely automated tools, it is unrealistic and infeasible to expect that artificial intelligence solutions can compensate fully for errors in human judgment. And, humans cannot be removed from the full data and information fusion chain. At best, then, computer-based tools should augment more finely honed human detection strategies and skills. The research reported herein was intended to address the need for improved information assurance by bringing together a multidisciplinary and multi-institutional research team to develop a theoretical model informed by state-of-the-art knowledge, to identify reliable indicators of deception through controlled laboratory experiments and field observations, to incorporate that knowledge into computer-assisted tools to detect deception, to identify factors influencing accurate detection by humans, and to develop training programs to overcome detection biases.

Objectives

This report presents the results of a five-year research project funded by the U. S. Air Force Office of Scientific Research under the Department of Defense's University Research Initiative. The six specific objectives of the project were to:

- Create an integrated model of human deception and detection to guide improved deception detection capabilities by humans and development of tools to augment human judgment.
- Verify reliable linguistic, vocalic, and kinesic-proxemic indicators of deceit present in face-to-face and electronically transmitted communication; determine variables that moderating these effects.
- Identify cognitive biases in human information-processing that result in failed deception detection and false alarms.

- Develop and test a prototype of an automated system, *Agent99,* to "flag" potentially deceptive messages and trigger more penetrating investigation.
- Develop a training program to improve the probability of accurate human deception detection and reduce the probability of false positives.
- Test the combined training procedures and automated system for their ability to improve detection accuracy and judgment processes.

Theoretical Development

Several theories of deception are reviewed. An integrated model is presented that combines *interpersonal deception theory* and *channel expansion theory*. The model views deception as a dynamic, interactive, adaptive and strategic activity and extends principles of deception to electronic forms of communication. The model calls into question the generalizability of previous research findings collected under noninteractive or minimally interactive contexts and guides all the experiments and field studies that were conducted as well as the tool development and training program intended to improve detection abilities.

Identification of Reliable Indicators

In all, 27 experiments entailing 3380 subjects were conducted to address the objectives of identifying reliable indicators of deceit, identifying moderators of those relationships, examining the influence of cognitive biases on detection accuracy, and testing computer-based training tools that were developed as part of the project.

The analysis of reliable indicators produced a host of indicators that successfully discriminated between truth and deception at a much higher rate than the current estimate for human detection of 54% accuracy overall. For linguistic indicators, a total of 33 different features differentiated truth from deception. Classification models from the laboratory experiments achieved detection accuracy rates as high as 88% for deceivers and 91% for truthtellers. Classification models from the field studies yielded detection accuracy rates of 90% for both deceivers and truthtellers. All of these analyses were conducted with the automated tools that were developed as part of this project, demonstrating the proof of concept for automating linguistic analyses.

Nevertheless, there were inconsistencies in cue emergence and general directions of classes of cues across studies. This variability implies that there are a number of moderating factors that govern what language is in use. Many of the patterns are at odds with previous findings collected under less interactive circumstances. They highlight the critical need for more testing and careful determination of the factors that define a particular situation (e.g., planned or spontaneous discourse, formal or informal interaction, narrative about events versus opinions or feeling states, high or low jeopardy for deceit being detected). With more planning time possible, deceivers could conjure up more details to appear more believable, although the extra information could be superfluous rather than useful. The fact that quantity cues are unreliable can explain the low accuracy in human judgment of deception because humans tend to rely heavily on these convenient (but unreliable) quantity cues.

Analyses of vocalic indicators demonstrated that up to 34 different vocalic cues differentiated truth from deception at accuracies of up to 100% for truth and 100% for deception; however, caution is warranted due to the small sample sizes for the experiments that were conducted.

Nevertheless, these results are very encouraging that vocal features can be reliable indicators of deceit. Moreover, such indicators are often less controlled by communicators and therefore may be useful telltale indicators in a variety of circumstances. Results also demonstrate the viability of automating their detection.

Analyses of kinesic indicators revealed at least 33 different features that effectively distinguished truth from deception. The best model accurately predicted 94% of the truthful cases and 100% of the deceptive cases. Again, small sample sizes warrant caution in interpretation and replication. Even in tests with limited sample size, however, automated extraction and analysis of nonverbal features performs better than typical human judgment. Further, these results demonstrate that automatically extracting nonverbal features for the purpose of deception detection may be feasible.

The measurement of linguistic, vocalic and kinesic features permitted the most comprehensive examination to date of the utility of fusing multiple indicators into single models. A variety of approaches was taken to conducting fusion oriented research and demonstrated high classification accuracy with combinations of objectively-measured and subjectively-measured features.

Several moderator variables were also tested for their influence on deception displays and detection accuracy. Motivation, task complexity, modality of communication, group size, suspicion, and familiarity among group members all affected displays and/or detection accuracy. Viewing deception as an interactive and adaptive activity necessarily requires taking these moderators into account. Each deserves continued research attention.

Identification of Cognitive Biases

A third objective of the research was to examine cognitive biases that influenced human detection ability. Fourteen different cognitive heuristics and biases were identified that could undermine deception detection accuracy. Four that were examined experimentally were truth bias, visual bias, demeanor bias, and expectancy violations bias. Results indicated that all four biases influenced judgments and pointed to reliance on audio rather than video-based communication as producing more accurate judgments.

Prototype Development

A fourth objective was to develop a prototype for deception detection. As part of this endeavor, we developed a suite of tools that we named *Agent99*. To analyze linguistic features automatically, we developed *Agent99 Parser and Client,* which were built upon two open-source tools, General Architecture for Text Extraction (GATE) and WEKA, a platform that implements machine learning algorithms and statistical classification. A separate *Analyzer* was built to facilitate recording and exporting of manually coded linguistic features generated by trained human coders. To conduct kinesic (body movement) and proxemic (spatial) analysis, C-BAS (C sharp Behavioral Annotation System), a tool for video-based behavioral observation, was developed and implemented for human-annotated vocal and kinesic behavior. Another video-based tool, *A99 AutoID Behavioral Analysis System,* is the set of components and processes for automatic extraction and identification of behavior from video. Possible interfaces for field-usable displays as another *BAS* component were also prototyped.

Finally, a computer-based *Trainer* was designed and implemented. The Agent99 Trainer that was built on our previous web-based multimedia training system called Learning-by-Asking. It is based on a three-layer client/server architecture, which includes client, application and database layers. Its user modules include Watch Lecture, Lecture Transcript, View Examples, Ask Questions (Natural Language Search), Navigable Outline, and Pop-Up Quizzes. Further details of tool development, including system architecture, interface design and user requirements are reported.

Testing of Prototypes and Training Tools

The fifth and sixth objectives were to test the tools and test the training curriculum. The training curriculum was developed in a format similar to that used at USAF training installations. Three lectures with PowerPoint were prepared and videotaped in three topics: deception detection generally, cues used to detect deception, and heuristics for decision making that are susceptible to deception. Following extensive pilot testing with university students and Air Force personnel, the *A99 Trainer* was field-tested twice at a USAF training location and also tested at FSU. Experiments examined the value of including the computer-based interface, nonlinear navigation, availability of additional illustrative examples, search capabilities, and intermittent pop-up quizzes.

Results revealed that the curriculum itself improved knowledge of deception and the ability to apply that knowledge in judgment tests. *A99 Trainer* also improved learning relative to straight lecture, and the fully featured version of the trainer produced the greatest increments in knowledge and judgmental accuracy. Usability tests were also conducted and confirmed that the system was helpful, easy to use, interesting, well-synchronized, allowed good learner control, and provided useful illustrations.

Major Lessons Learned

Several important lessons learned are recapped in the report. They are:

1. Computer tools can assist users in detection.
2. Biases exist.
3. No single cue is sufficient for detection.
4. Context must be considered.
5. Culture must be considered.
6. Ground truth is difficult to obtain.
7. More data is better.
8. The multi-disciplinary approach is valuable.
9. Research methods should be theory-driven.
10. Both laboratory and field testing are necessary.

Transitions

Several products resulted from this project. Software tools that have been implemented elsewhere from the *Agent99 Suite* are the *Trainer* and *C-BAS. StrikeCom* has been used as both a research tool and training tool for groups planning network-centric warfare. An interface has also been developed for delivering automated results from an intelligent agent to end users. Its impact on usage and judgments is being tested with students and professionals.

Finally, a searchable repository was developed to house the data, video files, and 116 publications that emanated from this project. Publications are listed in Appendix A. Additional conference papers are also available.

Table of Contents

I. Acknowledgement

This five-year project would not have been possible with the extensive collaboration and involvement of a large number of faculty members, research scientists, graduate students and undergraduate students. We wish to gratefully acknowledge all the contributions of the individuals who served on the various research teams at the University of Arizona, Michigan State University, Florida State University, Air Force Institute of Technology, Baylor University, University of Portsmouth, and Oklahoma State University. All individuals who participated in the conduct of some aspect of this research or were consultants to it are listed below. The size of the list is indicative of the magnitude of labor and expertise that were invested in this project.

We also are grateful to the department heads and their superb staff members for providing excellent administrative support throughout the conduct of this research. In particular, Dr. Charles Atkin, Chair of the Department of Communication, and Ann Wooten, Administrative Assistant, were instrumental in facilitating the collaboration with MSU.

Finally, the extensive progress we made during this effort is in no small measure due to the encouragement, support, and flexibility of our program manager, Dr. Robert Herklotz. Not only was he instrumental in facilitating collaborations with the Air Force Research Laboratory, Rome, NY, and the Center for Biomedical Imaging and Modeling, Rutgers University (PI: Dr. Dimitris Metaxas, but he was also very receptive to expansions and revisions of the research effort, such as incorporating the extraordinarily challenging but critically important data collection at the U.S.-Mexico Border. His broad vision and understanding of the importance of this program of research remained a strong motivation for our efforts to justify his confidence in us and has provided us with a firm foundation for continued research in this direction. We are deeply appreciative of his unwavering support and guidance.

Multidisciplinary Team Members and Consultants

Dr. Mark Adkins
Dr. Lisa Allspach Miczo
Dr. Walid Afifi
Bjorn Bengtsson
Dr. Dan Benincasa
Lt. Col. (ret) David Biros
Dr. Pete Blair
Dr. Joe Bonito
Kathy Broneck
Dr. David Buller
Dr. Judee Burgoon
Dr. John Carlson
Capt. Frank Caruso
Dr. JinWei Cao
Dr. Fang Chen
Rudy Cole
Steven Coulon
Dr. Janna Crews
Ltc (Ret.) Kathy Debolt
Dr. Amit Deokar
Tom Dibert
Chris Diller
Lt Monica Dziubinski
Dr. Norah Dunbar
Dr. Amy Ebesu Hubbard
Dr. Kory Floyd
Capt. Christopher L. Ford
Ronald Friend
Christie Fuller
Dr. Joey George
Dr. Gabe Giordano
John J. Grieco

Lauren Hamel
Dr. Dave Harris
Capt. Michael Hass
Joel Helquist
Clifford Hopkins
Mark C. Huggins
Sean Humpherys
Matthew Jensen
Dr. Kara Kam
Capt Monte Knode
Dr. John Kruse
Lt Mark Lankowski
Dr. Aaron D. Lawson
Dr. Jim Lee
Dr. Beth LePoire
Dr. Timothy Levine
Dr. Ming Lin
Dr. Shan Lu
Dr. Kent Marett
Brian R. McBride
Dr. Dimitris Metaxas
Tom Meservy
Dr. Nathan Miczo
Kevin Moffitt
Emily Moyer
Dr. Jay Nunamaker
John J. O'Reilly
Ben Ousley
John G. Parker
Tim Peters
James Pidd
Tiantian Qin
Dr. Artemio Ramirez
Roy Ratley
Deborah M. Rodriguez
Dr. Neal Rowe
Capt. Jachin Sakamoto
Cynthia C. Sharpe
Dr. Allen Sears
Matthew Stoner
Dr. Gary Strong
Dr. Renee Strom
Dr. Gabriel Tsechpenakis
Dr. Patti Tilley
Dr. Douglas Twitchell
Dr. Aldert Vrij
Dr. Joe Walther
Melody Weist
Dr. Stanley J. Wenndt
Dr. Cindy White
Bob Younger
Dr. DongSong Zhang
Dr. Lina Zhou

II. Statement of the Problem

Throughout history, effective information and communication systems have been key enablers of successful combat and peacetime missions. Nowhere is this fundamental principle more evident than in the emerging Command, Control, Communications and Computer Systems Infosphere (C4I) that underpins the joint battlespace. The explosive emergence of new communication and information technologies portends profound changes in the conduct of military operations in the 21st century, with unprecedented capacities for the rapid, real-time, global exchange of messages and complex information needed for battlefield success. Indeed, the envisioned transformation of the joint forces into full spectrum dominance in the 21st century depends on successful achievement of information superiority, especially in the face of asymmetric warfare. Yet the sheer volume, complexity, and speed of information transmission and communication, and our lack of knowledge of unintended as well as intended capabilities of newly adopted tools and systems, pose accompanying risks to information assurance. Risks include biased and erroneous intelligence, inability to fuse data and ideas into operational concepts, inadequate assessment of alternative interpretations, and faulty, even catastrophic, decision-making if implementation is not accompanied by deeper understanding, training, and tools to mitigate these risks.

Joint Vision 2020 (http://www.dtic.mil/jv2020/) underscored the importance of information technology (IT) to the war-fighter in the coming years. Information and IT are key enablers toward achieving the goal of "decision superiority." However, while superior IT offers many advantages, it also creates vulnerabilities that our adversaries can exploit. For example, Biros, Zmud and George (2002) demonstrated how personnel specialists could be spoofed into making erroneous decisions when the data in their Personnel Concept III (PC-III) system was manipulated. Participants in this study not only made inaccurate decisions, they also failed to identify obvious errors in the data presented to them. Similar results were obtained at Wright-Patterson Air Force Base Division of Information Technology (AFIT) in studies dealing with Airborne Warning and Control System simulations. Study participants were easily spoofed into believing friendly aircraft were foes and adversaries were friendly aircraft. These studies demonstrate how easily military personnel can be spoofed by deceptive information. As well, personnel may fail to question information that is not necessarily deceptive but invalid or erroneous nonetheless.

Broadly defined, *deception* entails messages and information knowingly transmitted to create a false conclusion (Buller, Burgoon, White, & Ebesu, 1994). Deception comes in many guises. It includes not just lies and fabrications but also evasions, equivocations, exaggerations, misdirections, deflections, and concealments. In fact, the latter forms of deceit are far more common than outright lies (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Turner, 1975). Field informants may omit critical details about suspicious activities. Disinformation campaigns may use the magician's trick of misdirecting attention to bogus operations and away from real ones. Adversaries may leak information that exaggerates or downplays the state of their weapons arsenals and make public speeches that conceal their true intentions. Intelligence analysts may equivocate about the thoroughness of their analysis. If all of these forms of diverging from "the truth, the whole truth, and nothing but the truth" are included under the umbrella of deception, it becomes apparent that deception may compromise all stages of data and information fusion in

which humans play a role, whether it be initial data-gathering by forward operating controllers; use of human-computer interfaces (HCI) to mine and integrate data; situation, threat, and impact assessment by information assurance specialists and intelligence analysts; or process refinement and formulation of action plans by the joint forces commander or other senior decision-makers. And, the higher the degree of inference-making, the more opportunities for omitted, exaggerated, ambiguous, and fabricated information to become fused with valid information, making resultant knowledge and decision-making erroneous.

The complexity of the task of detecting deceit cannot be overstated whenever humans are involved as information sources, conduits, or recipients. Extensive social science research has confirmed that humans are very adroit at dissembling yet very poor at detecting it. The consistent and notoriously low estimates of human accuracy in detecting detection (Bond & DePaulo, 2006; Feeley & deTurck, 1995; Miller & Burgoon, 1982; Zuckerman & Driver, 1985) point to human decision-makers as the likely weakest link in any C4I system. Human deception detection is hampered by several factors: the lack of a reliable, stable, and uniform set of indicators of deceit; information-processing biases that lead humans to regard incoming communications as truthful; tendencies to rely on nondiagnostic indicators when deceit is suspected; and tendencies for heightened suspicion to backfire, leading to "false positives" (i.e., judging truthful and valid information as deceptive (Anolli & Ciceri, 1997; Biros, George, & Zmud, 2002; Buller, Burgoon, Buslig, & Roiger, 1996; Buller, Burgoon, White, & Ebesu, 1994; Feeley & deTruck, 1995; Fiedler, 1993; Levine et al., 2000; Vrij, 1994, 1999; Vrij, 2000; Vrij, Akehurst, & Morris, 1997). Even highly trained law enforcement and military personnel have often shown little better than chance accuracy in detecting deception (Burgoon, Buller, Ebesu, & Rockwell, 1994; Ekman, Friesen, O'Sullivan, & Scherer, 1980).

These already serious difficulties are likely to become *magnified* when messages and information are derived from IT artifacts such as computers and networks. C4I technologies such as email, wireless voice communication, teleconferencing, and computer agents that aggregate data from unauthenticated sources may exacerbate detection challenges, not only because operators, analysts, and decision-makers may be unaware of how deceit can be perpetrated in the new infosphere and but also because new technologies introduce additional cognitive biases, such as placing undue trust on information delivered via computers or mass media (George & Carlson, 1999; Nass, 1993; Nass, Fogg, & Moon, 1996; Nass, Steuer, & Tauber, 1994; Nass & Reeves, 1991). Too, the accelerated pace of information exchange, especially under the physically and cognitively taxing conditions that characterize wartime and combat operations other than war, may heighten reliance on heuristic processing (use of mental shortcuts) that divert attention from diagnostic information to invalid indicators, thereby further eroding detection accuracy. Reliance on visual interfaces, for example, ironically can make detection worse rather than better (DePaulo, Stone, & Lassiter, 1985; Zuckerman, DePaulo, & Rosenthal, 1981). And opportunities for deceivers to plan, rehearse, and edit their messages prior to transmission may place recipients at a further disadvantage (Greene, O'Hair, & Yen, 1985; Zuckerman & Driver, 1985).

III. Project Objectives and Approach

Clearly, deception and its detection pose a significant threat to information superiority. The first line of defense in information assurance thus must begin with hardening the C4I against deception entering the knowledge engine at the data capture stage and secondarily having strategies and tools to flag, probe, and counter deceptive information that evades initial detection. Whereas much attention in military information security has focused on intrusion detection systems (IDS) such as the Automated Security Incident Measurement tool, intrusion detection systems can be overcome by low and slow attacks (also on-going AFIT research) and are very labor intensive for network administrators who are already overworked. Furthermore, IDS do not prevent intrusion into military networks by other means such as social engineering. A "red team" attack at AFIT clearly demonstrated this threat. Base employees (military, civilian, and contractor) were sent an email message by someone identified as a systems administrator who needed the email message receivers' login ID and password to accomplish some system upgrade. Many of the recipients replied to that message by giving the requested information, revealing how easy it would be for an adversary to spoof an IP address and conduct a similar operation. This incident underscores the oft-repeated conclusion that humans are the weakest link in the infosphere. Coping with human fallibility thus remains a major problem to tackle.

Although it is seductively appealing to try to replace human detectors with completely automated tools, it is unrealistic and infeasible to expect that artificial intelligence solutions can compensate fully for errors in human judgment. Past instruments (e.g., voice stress analyzers and the polygraph) have had varying and sometimes unimpressive success rates. Moreover, even if a dependable set of indicators could be verified, automated systems could not replace the extraordinary (if underutilized) human capacity to recognize, integrate, and interpret subtle and highly variable behavioral anomalies. And, humans cannot be removed from the full data and information fusion chain. At best, then, computer-based tools should augment more finely honed human detection strategies and skills.

How to integrate human detection with automated tools requires investigating deception and its detection under conditions like those facing today's joint forces. Yet the voluminous research on deception conducted to date is not very informative. Virtually none of it has been conducted utilizing the kinds of computer-mediated systems and human-computer interfaces undergirding the joint battlespace C4I. Further, prior research has typically entailed fairly sterile, static, and inconsequential tasks (e.g., students telling short, innocuous lies recorded for later judging by human "detectors") that bear little resemblance to the tasks faced by military personnel responsible for information assurance. Thus, research must better approximate the kinds of dynamic, complex, and sometimes taxing conditions that characterize military operations in the 21st century.

The research reported herein was intended to address these concerns by bringing together a multidisciplinary and multi-institutional research team to develop a theoretical model informed by state-of-the-art knowledge, to identify reliable indicators of deception through controlled laboratory experiments and field observations, to incorporate that knowledge into computer-assisted tools to detect deception, to identify factors influencing accurate detection by humans, and to develop training programs to overcome detection biases.

The team consisted of researchers from the disciplines of communication, human development, and management information systems at the University of Arizona; criminal justice and communication at Michigan State University; information systems at Florida State University; information technology and warfare at Air Force Institute of Technology; and psychology at University of Portsmouth, UK. The research objectives were as follows:

- Create an integrated model of human deception and detection to guide creation of human and automated tools for improving detection capabilities.
- Verify reliable indicators of deceit in content-based, linguistic, and nonverbal signals present in electronically transmitted information and conditions moderating those signal profiles.
- Identify cognitive biases in human information-processing and reasoning about uncertainty that result in failed deception detection and false alarms.
- Develop and test a prototype of an automated system, *Agent99*, to "flag" potentially deceptive messages and trigger more penetrating investigation.
- Develop a training program to improve the probability of accurate human deception detection and reduce the probability of false positives.
- Test the combined training procedures and automated system for their ability to improve detection accuracy and judgment processes.

This report summarizes the accomplishments on each of these objectives. It concludes with a summary of lessons learned and transitions out of the project. Publications from the research are listed in Appendix A.

IV. Theories and Models of Deception

The scientific examination of human deception has a long history. Over a century of research and theorizing has seen physiognomic, physiological and psychological models all advanced as the best approach to tell if someone is lying. Yet accurate rates have remained poor. Given our desire to develop unobtrusive, cost-effective, scalable, and field-worthy tools, we have taken a behavioral approach, searching for the most diagnostic indicators of deceit and extending the research domain into the arena of electronic communication. Four theories and models have framed our research: *interpersonal deception theory* (IDT), and *channel expansion theory* (CET), *expectancy violations theory* (EVT) and *signal detection theory* (SDT).

A. Interpersonal Deception Theory

Interpersonal deception theory (IDT) arose out of the conviction that understanding of deception is best realized when grounded in the interpersonal interactions that give deceit its sustenance. Human deception is a common daily occurrence that is part and parcel of every relationship: "even the most publicized of deceits is comprised of endless interpersonal encounters in which lies, exaggerations, misrepresentations and the like are created and perpetuated" (Burgoon & Buller, 2004).

IDT (Buller & Burgoon, 1996; Burgoon & Buller, 2004) can be contrasted to more psychological explanations for deceptive communication in emphasizing the strategic and dynamic nature of deception displays and the mutual influence between sender and receiver that occurs in interpersonal encounters (Burgoon, Buller, White, Afifi, & Buslig, 1999; White & Burgoon, 2001). Although initially applied to face-to-face deception, IDT's principles and findings apply as well to mediated forms of communication, such as email, voice communication, and videoconferencing, and to two-person or multi-person communication. The original version of IDT (Buller & Burgoon, 1996) articulated assumptions about deception and about interpersonal communication upon which the theory was founded. It then advanced a number of empirically testable statements and presented the results of numerous experimental tests in face-to-face contexts. We combined this theory with CET to better account for deceptive behavior and its detection when transmitted via electronic media.

Three decades of research on communicator credibility, nonverbal and verbal message features, violations of expectations, and influence processes were important tributaries to IDT (Buller, 1987; Buller & Burgoon, 1986; Burgoon, Buller, Guerrero, & Afifi, 1996; Burgoon & Hoobler, 2002; Burgoon & Doran, 1983; Miller & Burgoon, 1982). Subsequent to its publication, IDT-generated hypotheses were put to test in at least 15 experiments and field studies that provided substantial validation of IDT. This body of work, detailed more fully below, was the central foundation for the current research program.

1. Assumptions of IDT

When people communicate, all parties are both senders and receivers of messages. In fact, it is a misnomer in interpersonal interaction to separate senders from receivers, except in an abstract sense (which we do henceforth). In normal conversations, senders are simultaneously producing their own nonverbal and verbal messages while observing feedback and other overt reactions such as emotional displays from listeners. Likewise, listeners are not passive message recipients. While listening, they provide verbal and nonverbal feedback, manage their outward demeanor, and formulate their own turn at talk. All parties to deceptive episodes are likewise concerned with such multiple goals as preserving good interpersonal relationships, masking inappropriate emotions, keeping conversations running smoothly, and appearing credible. In achieving these multiple conversational functions, they must manage a host of verbal and nonverbal behaviors. Thus, conversations are dynamic, multifunctional, multidimensional, and multimodal events in which participants must perform numerous communication tasks in real time.

Such juggling requires considerable skill to accomplish effectively. Communicators must respond to a host of cognitive and behavioral factors that influence deliberate communication acts and produce some unintended behaviors. Although conducting social interaction is arguably a cognitively demanding activity, it appears that people are generally good at it because much of normal conversation is fairly routinized. Too, social interaction is made easier by the fact that people have learned culturally prescribed rules and expectations. Among the most relevant expectations are that people will be truthful, that they will display a moderate degree of involvement, and that they will match and reciprocate one another's verbal and nonverbal behavior in conversation. Violations of these expectations are assumed to elicit suspicion.

As regards deceptive messages and their detection, IDT assumes that deception entails three classes of strategic, or deliberate, activity--information, behavior, and image management. The term "management" implies that deception is a motivated behavior, undertaken for a purpose. Usually that purpose is one that benefits the sender, although senders frequently claim that they deceive to benefit the receiver or a third party to the conversation. Information management refers to efforts to control the contents of a message and usually concerns verbal features of the message. Behavior management refers to efforts to control accompanying nonverbal behaviors that might be telltale signs that one is deceiving. It derives from the assumption that verbal and nonverbal messages are constructed as a unified whole and that nonverbal behaviors are often intended to augment and extend the meanings conveyed by verbal content. Image management refers to more general efforts to maintain credibility and to protect one's face, even if caught. It derives from the assumption that individuals are motivated to protect their self and public image.

These three classes of strategic activity work hand in hand to create an overall believable message and demeanor. By way of example, a detainee suspected of arms smuggling might tell a border agent, "I did not know that the weapons were in the false bottom of the truck" (information management) while crossing his arms to avoid nervous gestures or body movements (behavior management) and maintaining eye contact to appear honest (image management).

The assumption that senders' verbal and nonverbal behavior reflects planning, rehearsal, editing, and other conscious or semi-conscious efforts to pull off deceit does not mean that deceivers are always successful at doing so. IDT also assumes that deceivers engage in nonstrategic actions–classes of behavior that may be involuntary and uncontrolled. Nonstrategic activity may result in poor, unnatural, or embarrassing communication performances. A case in point is blushing when a person gives a nontruthful answer to a pointed inquiry. The complexity of deceptive messages, and the knowledge that deception violates conversational rules and social prescriptions against deceit, can alter the mental state of senders. It can increase the cognitive effort needed to formulate this multifaceted conversational behavior. It also may increase arousal and provoke negative affect. All of these processes may result in inadvertent signals that something is not quite normal about a person's communication, although IDT does not assume that such nonstrategic signals are necessarily or universally present.

Finally, in IDT, the actions of recipients of deceit must be taken into account. Receivers' perceptions of deceit and their suspicion (a belief held without sufficient evidence or proof to warrant certainty that a communicator may be deceptive) are factors that influence their own behavior, the credibility they attribute to senders, and the accuracy of their detection of deceit. (See (Buller & Burgoon, 1996, and Burgoon & Buller, 2004, for fuller explanations of these assumptions.)

2. IDT Propositions

With these assumptions about interpersonal communication and deception as a backdrop, we formulated a theoretical model of deception containing 18 propositions summarized in Table 1 (following page), which is reproduced from Burgoon and Buller (2004). These are also explicated in more detail in Buller and Burgoon (1996) and Burgoon, Buller, Guerrero and Afifi (1996).

B. Channel Expansion Theory

CET (Carlson, George, Burgoon, Adkins & White, 2004; George & Carlson, 1999; Tung, Lam, & Tsang, 1997) was developed to draw attention, beyond features of actors, features of transmission channels, features of messages, and the information exchange process itself. CET argues that the information bandwidth of an interface is not fixed and is not based solely on the objective characteristics of the medium. Rather, as participants develop experience with each other, the channel, the message topic, and the communication context, they will perceive the channel as being better able to handle rich, equivocal, socio-emotional messages.

While CET encompasses all communications media, it is especially apropos for computer-mediated environments such as e-mail and chat, both known to "filter out" certain information cues (e.g., tone of voice), which may make the latter stages of the data fusion process more difficult. This argues for taking a longer, longitudinal view on how deceivers and detectors adapt to information technology–on the sender side, to use leaner media to their advantage in evading detection; on the receiver side, to acquire greater acuity in detecting deception. Our integrated theoretical model, published and elaborated in (George & Carlson, 1999) merges features of IDT with CET (see Figure 1). Testable hypotheses are derivable from the relationships depicted, in combination with the assumptions and propositions of IDT.

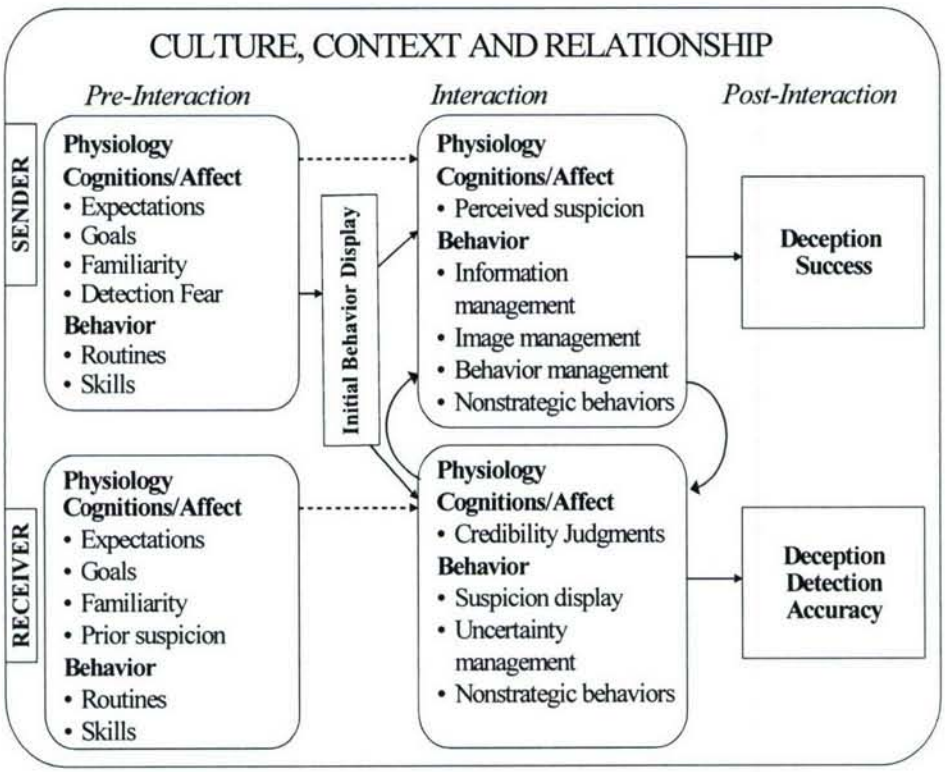


Figure 1. Model merging IDT and CET.

Table 1. Propositions in Interpersonal Deception Theory.

1. Sender and receiver cognitions and behaviors vary systematically as deceptive communication contexts vary in (a) access to social cues, (b) immediacy, (c) relational engagement, (d) conversational demands, and (e) spontaneity.
2. During deceptive interchanges, sender and receiver cognitions and behaviors vary systematically as relationships vary in (a) relational familiarity (including information and behavioral familiarity) and (b) relational valence.
3. Compared with truth tellers, deceivers (a) engage in greater strategic activity designed to manage information, behavior, and image and (b) display more nonstrategic arousal cues, negative and dampened affect, noninvolvement, and performance decrements.
4. Context interactivity moderates initial deception displays such that deception in increasingly interactive contexts results in (a) greater strategic activity (information, behavior, and image management) and (b) reduced nonstrategic activity (arousal, negative or dampened affect, and performance decrements) over time relative to noninteractive contexts.
5. Sender and receiver initial expectations for honesty are positively related to degree of context interactivity and positivity of relationship between sender and receiver.
6. Deceivers' initial detection apprehension and associated strategic activity are inversely related to expectations for honesty (which are themselves a function of context interactivity and relationships positivity).
7. Goals and motivations moderate strategic and nonstrategic behavior displays such that (a) senders deceiving for self-gain exhibit more strategic activity and nonstrategic leakage than senders deceiving for other benefits and (b) receivers' initial behavior patterns are a function of (1) their priorities among instrumental, relational and identity objectives and (2) their initial intent to uncover deceit.
8. As receivers' informational, behavioral, and relational familiarity increases, deceivers not only (a) experience more detection apprehension and (b) exhibit more strategic information, behavior, and image management but also (c) more nonstrategic leakage behavior.
9. Skilled senders better convey a truthful demeanor by engaging in more strategic behavior and less nonstrategic leakage than unskilled ones.
10. Initial and ongoing receiver judgments of sender credibility are positively related to (a) receiver truth biases, (b) context interactivity, (c) and sender encoding skills; they are inversely related to (d) deviations of sender communication from expected patterns.
11. Initial and ongoing receiver detection accuracy are inversely related to (a) receiver truth biases, (b) context interactivity, and (c) sender encoding skills; they are positively related to (d) informational and behavioral familiarity, (e) receiver decoding skills, and (f) deviations of sender communication from expected patterns.
12. Receiver suspicion is manifested through a combination of strategic and nonstrategic behavior.
13. Senders perceive suspicion when it is present. (a) Deviations from expected receiver behavior increase perceptions of suspicion. (b) Receiver behavior signaling disbelief, uncertainty, or the need for additional information increase sender perceptions of suspicion.
14. Suspicion (perceived or actual) increases senders' (a) strategic and (b) nonstrategic behavior.
15. Deception and suspicion displays change over time.
16. Reciprocity is the predominant interaction adaptation pattern between senders and receivers during interpersonal deception.
17. Receiver detection accuracy, bias, and judgments of sender credibility following an interaction are a function of (a) terminal receiver cognitions (suspicion, truth biases), (b) receiver decoding skill, and (c) terminal sender behavioral displays.
18. Senders' perceived deception success is a function of (a) terminal sender cognitions (perceived suspicion) and (b) terminal receiver behavioral displays.

A further innovation of the merged model is the analysis of medium characteristics that must be taken into account as influences on deceptive encoding and decoding. Drawing upon numerous analyses of media characteristics (e.g., (George & Carlson, 1999; Tung, Lam, & Tsang, 1997) we have concluded that the following are especially germane for deception research: synchronicity, symbol variety, cue multiplicity, tailorability, reprocessability, and rehearsability (ability to plan, edit, or mentally rehearse one's messages before transmission).

C. Expectancy Violations Theory

Expectancy violations theory (EVT) was originally developed by J. Burgoon and colleagues to predict and explain the consequences of deviating from expected or normal nonverbal behavior during communication. (Burgoon, & Burgoon, 2001; Burgoon, Buller, Dillman, & Walther, 1995a; Burgoon & Hale, 1988; Burgoon & Le Poire, 1993; Burgoon & Walther, 1990) explains and predicts the consequences of It differentiates between behavioral confirmations (behavior that matches expectations) and behavioral violations (behavior that deviates noticeably from expectations) and identifies factors that result in confirmations or violations being positive or negative. The model was subsequently expanded to apply to verbal behavior and to a wider array of nonverbal behaviors and patterns than originally envisioned. Many of the behaviors identified as potentially reliable indicators of deceit qualify as negative violations because they deviate from normal conversational patterns and provoke suspicion. EVT and IDT together predict that people attune to these violations, even if only subconsciously. Thus, recognition of violations becomes a key principle for identifying suspicious behavior and alerting humans to same.

D. Signal Detection Theory

Signal detection theory (SDT) is a model for identifying whether judgments match ground truth (Swets, 1986, 2000a, 2000b). It did not originate in the field of deception but has become the standard for determining acceptable levels of accuracy in detecting deception and/or truth. It provides the classification model for distinguishing hits (judging actual deception as deceptive or truthful messages as truthful), misses (judging deception as truthful), and false alarms (judging truthful behavior as deceptive). It is also used to develop receiver operating curves (ROC) to examine trade-offs between false positives and false negatives. As well, its calculations identify the degree and nature of bias in judgments.

An updated model of our approach based on these three theories is shown in Figure 2 in which we distinguish between deviations from general norms, which would be applicable to making judgments about unknown others, and deviations from personal norms, which are applicable to making judgments about single individuals for whom a personal history of behavior is available.

In this model, we envision deception that is multimodal, with numerous cues that are candidates for analysis. Linguistic cues include features like word selection, phrasing, and sentence structure. Content/theme cues are taken from the meaning of the sender's words. Meta-content cues are derived from features that are related to content—e.g., number of details—but can be calculated without contextual information. Kinesic cues concern what in the popular vernacular is known as body language and specifically relates to physical movement. Proxemic cues concern the distancing and spacing patterns between people. Chronemic cues concern a person's use of time as a message. For example, a person might establish dominance by arriving late to a

meeting. Vocalic cues are features of the voice other than the words and are often referred to as prosody or paralanguage.

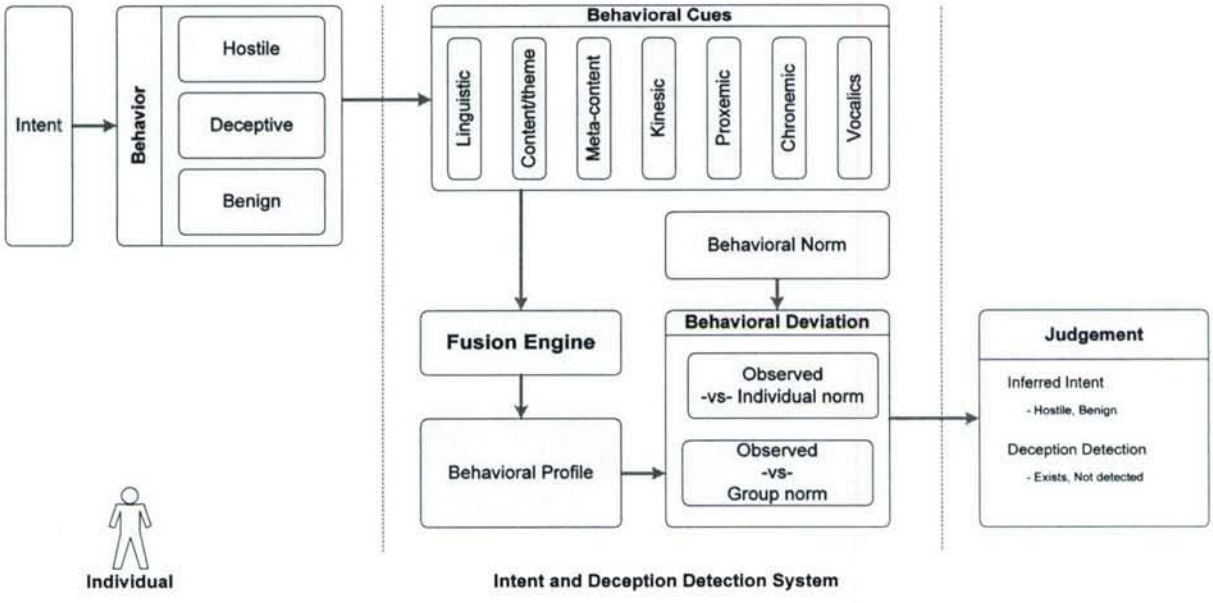


Figure 2. Model Employing Expectancy Violations As Signal Threshold.

From these cues, any deviation between observed behavior and past individual or general norms is noted. The deviations from multiple types of cues and from multiple communications channels are then fed into a fusion engine which weighs the importance of each indicator and compounds the most salient ones into a judgment of deception or truth. Our research, described next, has centered on identifying which indicators are most useful, separately or in combination, for identifying deception and which can be automated.

V. Methods for Identifying Reliable Indicators

Identification of reliable indicators proceeded on multiple fronts. In selecting verbal, vocalic, and kinesic cues that might effectively discriminate truth from deception, we focused on those indicators that were amenable to automated or computer-assisted analysis. Before describing the experimental and field work that tested for indicators, we describe tool development inasmuch as the tools were instrumental in conducting the analysis of the verbal and nonverbal indicators.

A. Development of Tools for Automated Analysis

The objective of this stream of the research project was to develop tools that could automatically identify, extract, and analyze verbal cues, vocal cues, and kinesic-proxemic cues.

1. Verbal Analysis of Deceptive Cues

The first decision point was to identify which kinds of verbal cues were the most amenable to automation and that might be reliable indicators. Verbal cues include the syntax, semantic structures, and vocabulary related to text-based comments, messages and reports or transcripts of

recorded face-to-face and audio communication. Although studies on verbal cues for deception detection have existed for more than four decades, not until recently have researchers considered looking for deceptive cues via an automated deception detection system based on natural language processing (Burgoon, Blair, Qin, & Nunamaker, 2003; Burgoon & Qin, 2006; Zhou, Burgoon, & Twitchell, 2003). Research suggests that we can learn a great deal about peoples' underlying thoughts, emotions, and motives by counting and categorizing the words they use to communicate (Newman, Pennebaker, Berry, & Richards, 2003). For example, previous summaries and meta-analyses (e.g., DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003) suggested that deceivers are less forthcoming and tend to give briefer responses than truth-tellers. By disclosing less information, they decrease the chances of being detected. Deceivers' messages also were thought to lack vivid and specific details because they do not have corresponding experiences that give rise to such details (Vrij, Edward, Roberts, & Bull, 2000).

a) Verbal Cues to be Extracted

Based on our synthesis of previous research and theorizing, we developed a taxonomy of classes of indicators to be investigated (Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Table 2 lists the classes of cues, specific indicators, and their definitions.

In order to automatically extract verbal cues from text, natural language processing (NLP) techniques were applied. NLP analyzes language by sub-sentential, sentential and discourse processes. The sub-sentential process can be further defined as phonological analysis, morphological analysis, syntactic parsing, and semantic analysis. The morphological analysis determines the part-of-speech in the sentence; the syntactic parsing decides the structure of a sentence following syntactic grammar. Current forms of semantic analysis may produce many ambiguities and so were excluded from the current efforts.

Following a thorough evaluation of the pros and cons of several proprietary, commercial, off-the-shelf and open source products, and comparing a proprietary tool (iSkim) to the Grammatik tool available in WordPerfect (see Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003), we opted for using an open-source shallow parser that could be readily modified to add new features. During shallow parsing, parts of speech are identified and cues can be calculated from the constituents. We selected the General Architecture for Text Extraction (GATE) (Bontcheva, Cuningham & Tablan, 2002) as the base program for analyzing written text and transcriptions of oral communication. Additional algorithms were written for complex measures such as emotiveness and readability. To measure affective states, we added a separate plug-in, a look-up dictionary developed by Whissell and colleagues (Whissell, 1986, 2001). The Whissell dictionary has more than 8,000 words with scaled values for affect-related indicators of activation, pleasantness, and imagery. Extremes in these affective states were measured as 1 or 2 standard deviations from the mean.

Table 2. Proposed Verbal Indicators and Definitions.

Quantity

1. Word: a written character or combination of characters representing a spoken word.
2. Verb: a word that characteristically is the grammatical center of a predicate and expresses an act, occurrence, or mode of being.
3. Sentence: a word, clause, or phrase or a group of clauses or phrases forming a syntactic unit which expresses an assertion, a question, a command, a wish, an exclamation, or the performance of an action, which usually begins with a capital letter and concludes with appropriate end punctuation.

Complexity

4. Average sentence length: (total # of words) divided by (total # of sentences)
5. Average word length: (total # of characters) divided by (total # of words)
6. Pausality: (total # of punctuation marks) divided by (total # of sentences)

Uncertainty

7. Modal verb: an auxiliary verb that is characteristically used with a verb of predication and expresses a modal modification.
8. Modifier: describes word or make the meaning of the word more specific. There are two parts of speech that are modifiers--adjectives and adverbs.

Verbal non-immediacy

9. Passive voice: the form of a verb used when the subject is being acted upon rather than doing something.
10. References: sum of self references (singular first personal pronoun), you-references, group references (first personal plural pronoun) and other reference (third personal pronoun).

Diversity

11. Content word diversity: (total # of different content words) divided by (total # of content words), where content word primarily expresses lexical meaning.
12. Lexical diversity: (total # of different words) divided by (total # of words), which is the percentage of unique words in all words.
13. Redundancy: (total # of function words) divided by (total # of sentences), where a function word is a word expressing a primarily grammatical relationship.

Specificity

14. Spatial details: information about the location or spatial arrangement of people and/or subjects
15. Temporal details: information about when the event happened or explicitly describes a sequence of events
16. Spatial and temporal details: sum of spatial and temporal details
17. Sensory: sensory experiences such as sounds, smells, physical sensations and visual details

Affect

18. Affect: conscious subjective aspect of an emotion apart from bodily changes
19. Imagery: words that provide a clear mental picture
20. Pleasantness: positive or negative feelings associated with the emotional state.
21. Activation: the dynamics of emotional state

b) Extraction Methods

Figures 3a through 3d demonstrate the extracting process. First, a segment of text that has been converted to XML format is input into GATE (a). Then the features to be applied (e.g., lexical diversity, word count) are selected and a copy of the lines is constructed between the message

and the cues program (b). Next, the message is scanned and the parts of speech are tagged in the interface (c). Finally, the summary results of the textual analysis are returned (d). These data can then be exported to any data mining or statistical analysis tool. For our verbal analyses, we exported data to WEKA, an open-source platform developed at the University of Waikato for implementing machine learning algorithms.

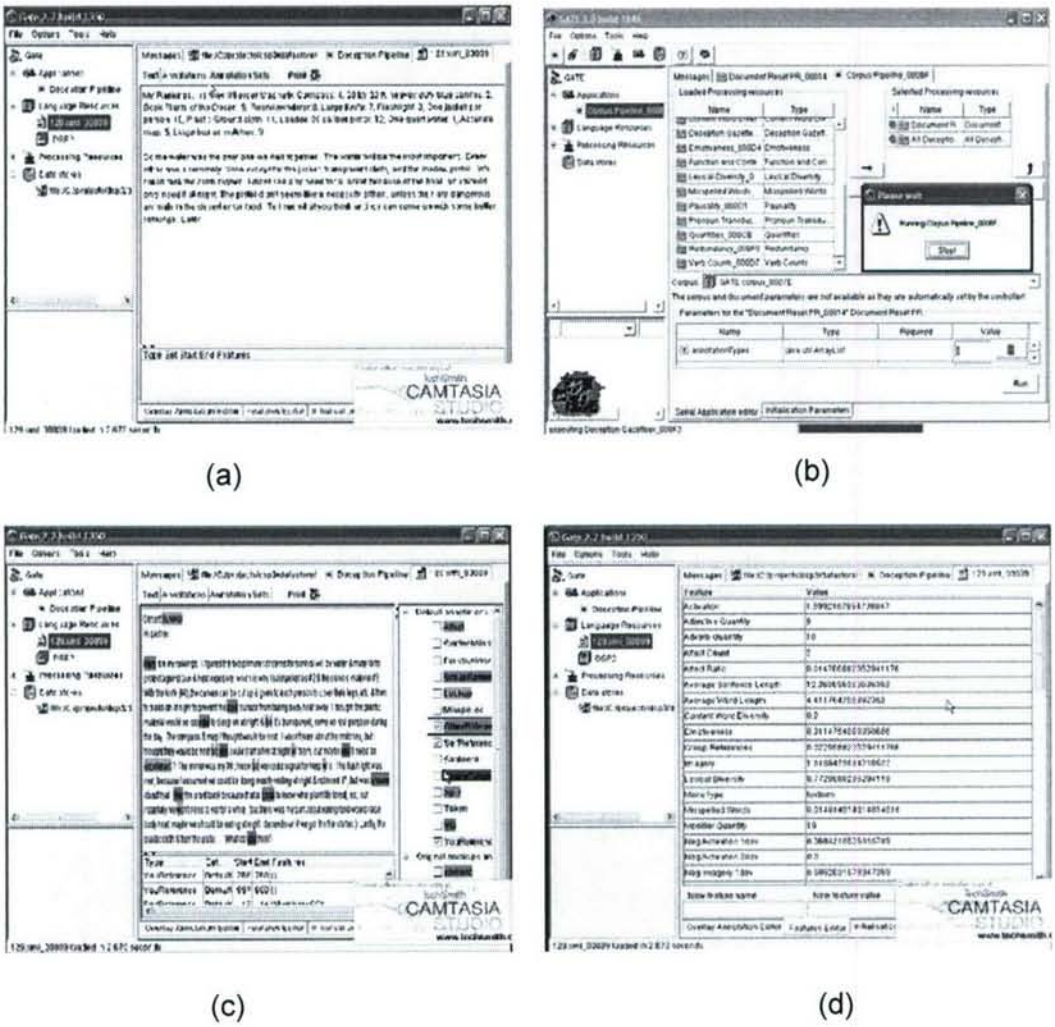


Figure 3. The Extracting Process Using GATE.

2. Classification Methods

The linguistic data derived from GATE were subjected to a variety of analyses including discriminant analysis, logistic regression, decision trees, and neural networks (see Qin, Burgoon, & Nunamaker, 2004; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Three out of four methods could differentiate between deceptive subjects and truthful subjects from the training data nearly perfectly. However, tests on the holdout data in cross-validations showed variable and sometimes substantial degradation in performance. These results highlighted the great variability in the data sets, pointing to the difficulty of predicting deception. No single method

emerged as being superior for predicting deception. All of the four methods under investigation are potentially good alternatives if models are pruned to include only significant indicators.

3. Vocalic Analysis of Deception Cues

The vocalic or paralinguistic facet of deception has long been of scientific interest. Belief that the voice is a very revealing channel is grounded partly in the integral role of the voice in expressing emotional states and arousal and partly in beliefs that the voice is less easily controlled and monitored than other communication channels, making it especially promising as the font of telltale indicators of deceit (Ekman & Friesen, 1969; Hocking & Leathers, 1980; Zuckerman, DePaulo & Rosenthal, 1981). A variety of commercial tools have been developed that are predicated on the voice being the ideal site for determining human stress levels (see Hollien & Harnsberger, 2006). Because the prosodic features of the voice are intrinsically linked to the verbal content of utterances, they may also supply important insights into cognitive states and the meanings communicators intend to express or conceal. For all these reasons, the voice is an important channel for deception detection.

a) Vocalic Cues to be Extracted

Our approach to vocal analysis was twofold. We used trained coders to identify some vocal features and used automated tools to identify others. Here we describe the automated feature extraction.

Previous research (e.g., DePaulo et al., 2003; Rockwell, Buller & Burgoon, 1997; Zuckerman & Driver, 1985) had identified utterance length, vocal tension, pitch (fundamental frequency) and pitch variation, loudness and loudness variation, intensity range, response latency, fluency, speech disturbances such as stutters and intruding sounds, vocal involvement/immediacy, vocal uncertainty, tempo change, and vocal pleasantness as features distinguishing truth from deception. Other features such as jitter or tidal respiration had been proposed but not investigated systemically. We grouped features into categories related to their etiology, specifically, as to whether they were thought to spring from arousal, from emotional stress and negative emotions, from cognitive effort to create a plausible response, from efforts to retrieve information from memory, or from intentional strategies to convey involvement, submissiveness, uncertainty (and thus lack of culpability), or pleasantness. The automatically extracted features are listed in Table 3 along with their definitions.

b) Extraction Process for Vocal Cues

Our approach for identifying vocal indicators associated with deception is similar to the approach adopted in many pattern classification systems. First, raw data is collected and segmented into meaningful units. Low-level features are then extracted from these segments. Additional higher-level features are computed for the segments and then all of the features are summarized. Finally these features are used to classify the segments. The following sections provide additional details of our approach to classify audio as deceptive or truthful. Figure 4 illustrates our approach for classifying audio as deceptive or truthful.

Table 3. Automatically Extractable, Low-Level Vocal Indicators of Deceit.

Arousal

1. Pitch: the fundamental frequency (number of cycles per second of a sound wave
2. Loudness/intensity: amount of energy expended, expressed in decibels
3. Intensity range: the minimum and maximum loudness of the voice

Negative Affect/Stress

4. Vocal tension: degree of muscle tremor in the larynx, measured by low-pass filter

Cognitive Effort

5. Speech disturbances: unfilled pauses within turns, filled pauses (ah, um, er) and other dysfluencies such as stammers, stutters, incoherent intruding sounds, and repeated sounds
6. Response latency: delay before the onset of a voiced response
7. Silences/pauses: lack of vocalization during one's speaking turn

Memory Retrieval

8. Utterance length: portion of time within a speaking turn that is voiced, divided by the number of turns
9. Tempo: (slow) rate of speaking

Strategies: Involvement

10. Tempo: (rapid) rate of speaking
11. Pitch variety: variance in pitch
12. Tempo variety: variance in tempo

Strategies: Submissiveness/Uncertainty

13. Rising intonation: vocal pattern with higher pitch at the terminal juncture (as in a question)

Strategies: Pleasantness

14. Resonance: vocalization in cavities of the vocal tract

Recorded voices in the form of digitized audio files serve as input for our approach. The audio data used in our research was 16 bit, linear Pulse-code manipulation (PCM) stereo sampled at 48000 samples per second (48 kHz). All audio files were down sampled to 8000 samples per second (8 kHz). Down sampling to 8 kHz was performed because it reduces the total number of data points that need to be analyzed. Additionally, some of the toolkits that were used to extract the low-level features require the data to be sampled at 8 kHz to match typical sampling of speech signals. All low-level features were extracted and computed on 8 kHz data.

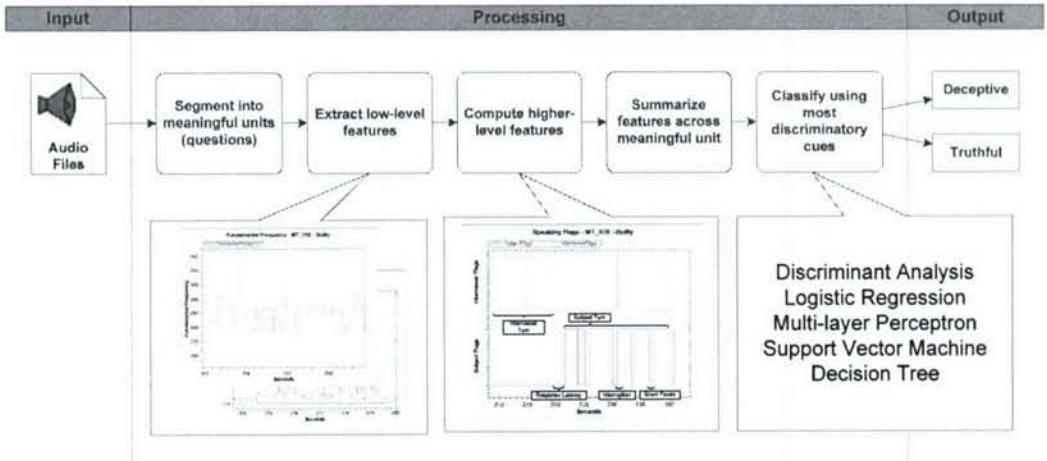


Figure 4. Overview Of Approach To Classify Truth And Deception From Audio.

c) Audio Segmentation

There are many strategies that exist to automatically segment audio data into logical units (Kemp, Schmidt, Westphal, & Waibel, 2000). However, these strategies are not error-proof. To increase our understanding of deceptive vocal cues, we minimized the amount of error introduced into our classification task by manually segmenting the input audio files into logical question-level units. In addition to the segmentation of audio files into question-level units, the audio signal for both the subject and the interviewer also needs to be identified. We did not conduct automatic identification and segmentation of speech segments spoken by each individual in a conversation, though several automated methods do exist (Adami, Kajarekar, & Hermansky, 2002). In our mock theft data set, subject and interviewer voices were recorded on separate channels. Thus, features for each individual were extracted using audio in the relevant channels.

d) Extraction of Low-Level Features

Research partners at the Air Force Research Lab's (AFRL) Audio Processing Group, located in Rome, NY, extracted the following low-level audio features using proprietary toolkits: fundamental frequency, low-pass filter output, gain/energy, response latency, and audio sample speech/silence segments for both the subject and the interviewer. All of the low-level features are provided for each subject and only the speech/silence segments are provided for the interviewer/interrogator. Many low-level features were computed on a frame-by-frame basis. These features all use a frame duration of approximately $1/30^{th}$ of a second (33 milliseconds). In other words, there are approximately 30 measurement points per second provided for these features. This frame duration was selected so that each speech frame could eventually be fused directly with features extracted from video frames. A few features—interviewer and subject speech/silence and the low-pass feature—were provided at 8000 samples per second.

For each interviewee, the *fundamental frequency* was computed over the duration of the audio channel. Those frames which had a signal-to-noise ratio (SNR) less than or equal to 9 dB were declared silence frames—fundamental frequency was not computed for silence frames. Fundamental frequency was extracted and then calculated on a frame-by-frame basis. Each frame lasts approximately $1/30^{th}$ of a second and subsequently could be directly fused with other features extracted from a video frame. Additionally, a pitch filter was used to eliminate any signals that were either too high (e.g. greater than 800 Hz) or too low (e.g. less than 40 Hz) to have been produced by human speech.

Figure 5a illustrates fundamental frequency. The subject's vocal tract filter *gain* and audio sample *energy* were also calculated on a frame-by-frame basis. The energy for each frame can be thought of as the area underneath the raw audio signal's curve. Gain can be thought of as the resulting signal strength after it has been multiplied by a constant gain factor (Nathan, 1998). Silence frames were detected if the signal-to-noise ratio was less than or equal to 9 dB. Both the gain and energy feature are reflections of the intensity of an audio signal.

Figure 5b provides an example of gain and energy. The subject's audio channel was *low-pass filtered* at 0 Hz to 30 Hz. The low-pass feature is the output of the low-pass filter, converted to a real number between -1 and 1. Typically, a low-pass filter is used in deception detection because lower frequencies in this range may tell us about the change in background noise (e.g., when someone splices audio segments together). More interestingly, frequencies in this range may also hint at tension in the voice of the subject.

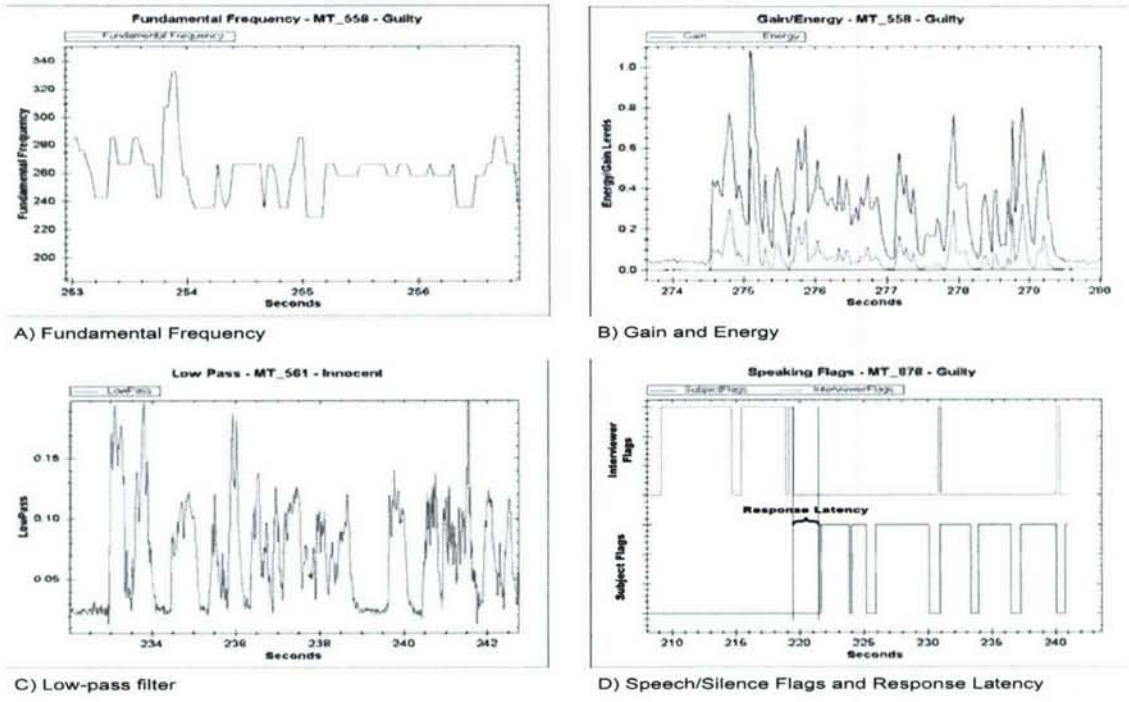


Figure 5. Illustrations of Low-Level Audio Features.

Figure 5c graphs the results of a low-pass filter on a subject's response. From observation of graphs illustrating the results of the low-pass filter, it was noticed that this feature had non-zero values when the subject was not speaking. To narrow the focus of this feature to only when the subject is talking, a new feature, adjusted low-pass, was created. This new feature only included low-pass feature data for when the subject was talking, rather than data for the whole segment. *Response latency* measures the time between when an interviewer's question ends and when the subject responds. Thus, the response latency feature captures the length of time of silence between the end of an interviewer's question and the beginning of the subject's response.
Figure 5d illustrates response latency and speaking turns. *Speech and silence* were calculated for the subject and the interviewer. If the signal-to-noise ratio for a sample exceeded 9 dB it was classified as speech, otherwise it was considered a silent frame.

e) Computation of Higher-Level Features

From the low-level features, we computed additional higher-level features that may help to distinguish deception from truth.

Table 4 provides a description of high-level features and also places these features in the taxonomy.

Table 4. Summary of Higher-Level Features for Automated Analysis.

Taxonomy Subcategory	Feature	Description

Time	Interviewer turns	The number of interviewer turns at talk within a segment
	Subject turns	The number of subject turns at talk within a segment
	Interviewer turn duration	Total amount of interviewer talk time within a segment
	Subject turn duration	Total amount of subject talk time within a segment
	Response latency (calculated)	Time between when a question ends and subject responds as calculated from the speech/silence segments
Fluency	Non-silence	A binary value indicating when either the subject or interviewer is speaking; calculated for each frame.
	Silence	A binary value indicating when neither the subject nor the interviewer is speaking; calculated for each frame.
	Interruptions (overlap)	A binary value indicating when both the subject and the interviewer are speaking; calculated for each frame
	Unfilled pause length	The duration of silence within each subject talking turn
Voice Quality	Adjusted low-pass	Low-pass filter value only included if the subject is speaking

A simple feature to capture the turn-taking in a conversation was created from the speech and silence segments. A turn begins when an individual begins talking, and is the only one talking, and ends when the other individual begins talking and is the only one talking. The number of interviewer turns and the number of subject turns are recorded, as well as the duration of each turn. Response latency was also re-calculated based on these speech segments. From the speech/silence segments a number of additional features were calculated that focus on the fluency sub-category of the taxonomy. These features include when someone is speaking (non-silence), when no one is speaking (silence), and when both the interviewer and the subject are speaking (interruptions). A feature was also created for the unfilled pause length cue. This feature is calculated by summing all silent pauses within each turn for the subject. This feature provides insight into the fluency of a subject's response.

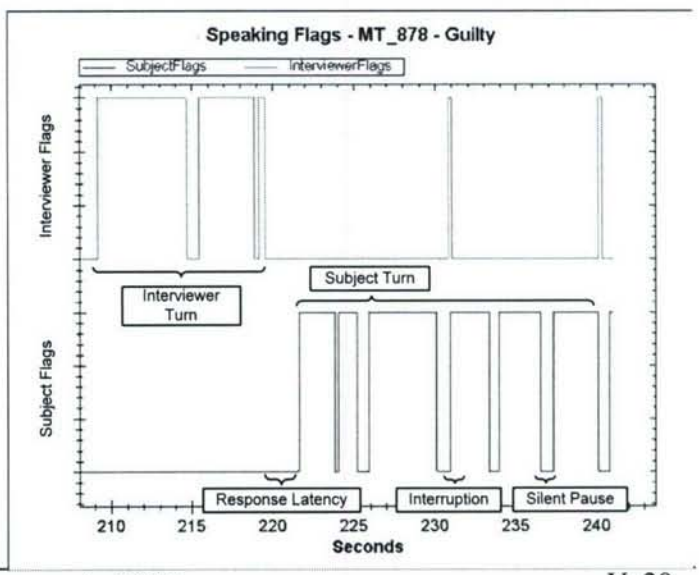


Figure 6 depicts speaker and interviewer turns, response latency, silent pauses, and interruptions.

Figure 6. Illustrations of Higher-Level Audio Features—Turns, Response Latency, Interruptions, and Pauses.

f) Summarization

Before we can use the classification methods that we have selected, the low-level and higher-level features need to be summarized for each of the questions (segments) previously identified. Simple means are calculated for the majority of features, as well as the average deviation from the mean. Average deviation is calculated as the summation of the absolute value of the individual data point minus the mean divided by the total number of measurements:

$$AvgDev = \frac{\sum |x - \bar{x}|}{n}$$ Equation 1 – Average deviation calculation

Initially, variances were calculated, however, after inspection of the variances it was noted that many were not normally distributed. The average deviation produces less-skewed distributions than variances of the same features. Additionally, min, max and range values are calculated for many of the features. Table 5 lists all summarized features that are created for each segment. Some of the features listed were not summarized because the raw feature was calculated for the entire segment (e.g., count of subject turns, first response latency of subject).

Table 5. Summarized Vocal Features.

Category	Feature	Average	Average Deviation	Minimum	Maximum	Range
Time	Interviewer speech	✔	✔	✔	✔	✔
	Subject speech	✔	✔	✔	✔	✔
	Latency (low-level)	✔	✔	✔	✔	✔
	Latency (calculated)	✔				
	Response latency first duration					
	Interviewer turns					
	Interviewer turn duration	✔				
	Subject turns					
	Subject turn duration	✔				
Intensity	Gain	✔	✔	✔	✔	✔
	Energy	✔	✔	✔	✔	✔
Frequency	Fundamental frequency	✔	✔	✔	✔	✔
Fluency	Unfilled pause length	✔				
	Overlap (interruptions)	✔	✔			
	Non-silence (someone speaking)	✔	✔			
	Silence (no-one speaking)	✔	✔			
Voice Quality	Low-pass	✔	✔	✔	✔	✔
	Adjusted Low-pass	✔	✔			

g) Classification Methods

To understand the predictive and discriminatory power of the low-level and high-level features that are extracted from the audio segments, a variety of classification methods can be used. As a form of statistical analysis, both discriminant analysis and logistic regression were utilized but primarily logistic regression in the case of audio cues, as the data were not normally distributed. Additionally, machine learning methods were applied. We utilized decision trees, multi-layer perceptrons, and support vector machines to classify cases as truthful or deceptive.

4. Kinesic Analysis

Freud believed that "He who has eyes to see and ears to hear may convince himself that no mortal can keep a secret. If his lips are silent, he chatters with his finger-tips; betrayal oozes out of him at every pore" (Freud, 1959, cited in Vrij, 2000). While deception detection may not be as simple as Freud suggested, theoretical and empirical research has shown that certain behaviors *do* differentiate deceivers from truthtellers. Kinesic analysis makes use of these behaviors to identify deception. As part of kinesic analysis the movements of one person engaged in a

recorded face-to-face interaction are examined for possible cues of deceit. The movement of the head and hands are analyzed as they move throughout the recorded segment, and features are calculated that give insight into whether or not the observed person is being deceitful.

a) Head and Hand Features to be Extracted

One might pose the question, "Why focus on the head and hands in inferring deception?" There are a number of reasons for such focus. First, there are theoretic reasons why deception may be manifest in movement of the head and hands. It is believed that gesturing is uniquely tied with the development and understanding of speech (Kelly, Kravitz, & Hopkins, 2004; McNeill, 1992). McNeill (1992), a leading scholar on human gesture, believes that through gesture people unwittingly disclose their inner thoughts and perceptions of their world. During deception, the deceiver must carefully control what is said and this control is also evident in gesturing and head movement. Zuckerman et al. suggested that deceivers must manage generalized arousal and specific affect such as guilt or fear when lying (Zuckerman, DePaulo, & Rosenthal, 1981). DePaulo and colleagues suggest a self-presentational approach to understand gesture and other nonverbal (DePaulo, Blank, Swain, & Hairfield, 1992). In this approach, DePaulo and colleagues describe people's attempts at impression management as they endeavor to maintain an air of sincerity and credibility in the eyes of others. They acknowledge that both truth-tellers and deceivers participate in monitoring self presentation; however, they believe that deceivers manage their behavior differently than truth-tellers and suggest this difference can be observed.

Another reason why gesturing and head movement may be affected by deception is the cognitive "taxation" that deception imposes. It has long been argued that deception is a difficult mental task and should impose demands on cognitive resources that result in a suppression of nonverbal behavior (Zuckerman, DePaulo, & Rosenthal, 1981). IDT similarly posits that deception is, on average, more difficult than truth-telling and that the cognitive effort needed to construct deceit will result in some performance impairment. Additionally, IDT posits that deceivers monitor their own performance and receivers' feedback to assess their deception success. If they perceive suspicion, they will attempt to adapt their communication so as to alleviate the suspicion. At the onset of the interaction, this can be a difficult task for the deceiver and may result in hampered nonverbal communication. However, the difficulty subsides as each party grows accustom to the communication style of the other party.

Overcontrol, self-presentation, and cognitive load do not solely account for all behavioral changes noted in deceivers. There are numerous moderating influences which alter the relationship between deception and observable behaviors. One thought to be particularly important is motivation (DePaulo et al., 2003; Vrij, 2000). One can imagine the difference in motivation between an experimental subject lying about something she didn't do and a guilty murderer trying to convince a jury of his innocence. A number of experiments, including one to be reported here, have manipulated or measured motivation (Burgoon, 2005; Zuckerman & Driver, 1995).

Second, there is strong empirical evidence which suggests that deceivers' head and hands move differently than truth-tellers'. Two recent meta-analyses conclude that there is a significant decrease in the amount of illustrating deceivers do in comparison to truth-tellers (DePaulo et al., 2003; Vrij, 2000). Illustrating gestures are those gestures which normally accompany speech.

They can include *iconics*, *metaphorics*, *beats*, and *cohesives* in the McNeill classification (McNeill, 1992). Illustrating gestures can represent semantic content in speech, can emphasize certain points, or can designate a relationship between ideas in speech.

While illustrating decreases significantly, it is important to note that, contrary to common opinion, self-directed gesturing such as scratching and preening were not found to differentiate between deceivers and truth-tellers (Vrij, 2000; DePaulo et al., 2003; Vrij, 2000). Deceivers were found to exhibit significantly more undirected fidgeting but not more object fidgeting, self fidgeting and facial fidgeting. Deceivers also displayed significantly more chin raises than truth-tellers but not more undifferentiated head movement. However, an interactive study by Buller, Burgoon, White and Ebesu (1994) found that deceivers showed significantly less total head movement than truth-tellers.

A final reason to focus on heads and hands is that such behavior is readily monitored, captured, and analyzed. The monitoring can take place unobtrusively and without the knowledge of the person being examined (Meservy et al., 2005). This is in contrast to many other forms of deception detection that require the use of sensors attached to the body (e.g., polygraph). Behavior monitoring has been shown to retain its accuracy even when the video frame rate falls (Meservy, Jensen, Kruse, Burgoon, & Nunamaker, 2006) (Lower frame rates are common in inexpensive security cameras). Further, behavior monitoring can be merged easily with other methods of deception detection for increased accuracy. Specifically, linguistic analysis, voice analysis, and thermal imaging methods might be used in conjunction with behavior monitoring and the combined system would still retain its unobtrusive qualities.

b) Tracking the Head and Hands

Numerous techniques exist for automatic tracking of human head and hands. Notable among these techniques are Pfinder, developed at MIT (Wren, Azarbayejani, Darrell, & Pentland, 1997) and Vector Coherence Mapping developed at Wright State University (Quek, Ma, & Bryll, 1999). The features used in this study to differentiate between truth and deception are completely independent of the tracking method and can be used with various tracking methods. For the feature set to be used, a number of measurements for each frame in a video segment must be collected. An ellipse should be formed around the head and the two hands and the center x, y position, major axis length, minor axis length, major axis angle should be collected for each of the hands and the head. The necessary measurements are shown in Figure 7.

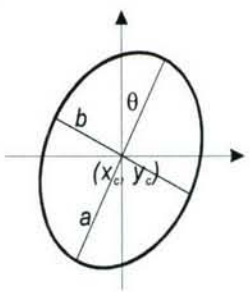


Figure 7. Necessary Measurements for Feature Use (Meservy Et Al., 2005).

Kinesic analysis utilizes a tracking method developed by Computational Biomedicine Imaging and Modeling Center (CBIM) at Rutgers University (Lu, Tsechpenakis, Metaxas, Jensen, &

Kruse, 2005). The method extracts hand and face regions using the color distribution from a digital image sequence. A three-dimensional look-up-table (3-D LUT) is prepared to set the color distribution of the face and hands. This 3-D LUT is created in advance of any tracking using skin color samples. After extracting the hand and face regions from an image sequence, the system computes elliptical "blobs" identifying candidates for the face and hands. The 3-D LUT may incorrectly identify candidate regions which are similar to skin color, however these candidates are disregarded through fine segmentation and comparing the subspaces of the face and hand candidates. Thus, the most face-like and hand-like regions in a video sequence are identified. From the blobs, the left hand, right hand and face can be tracked continuously. A complete technical description of the BAS system is beyond the scope of this study; however the interested reader is directed to Lu et al. (2005) and Meservy et al. (2005).

c) Features

The features used to differentiate between truth and deception were originally proposed by Meservy et al. (Meservy, Jensen, Kruse, Burgoon, & Nunamaker, 2005) and were tested in subsequent studies (Jensen, Meservy, Kruse, Burgoon, & Nunamaker, 2005; Meservy et al., 2005). A brief description of these features is reproduced here. Table 6 displays each of the feature names and their descriptions.

Table 6. Features Used To Discriminate Truth and Deception.

Factor	Feature	Body Part Measured	Description
Group 1	X	Head, RH, LH	x position of the blob
	Y	Head, RH, LH	y position of the blob
	Height	Head, RH, LH	length of the major axis
	Width	Head, RH, LH	length of the minor axis
	Angle	Head, RH, LH	angle of the major axis
Group 2	angle_diff	Head, RH, LH	difference in angles between previous frame and current frame
	Diff	Head, RH, LH	Euclidean distance between x, y pos. between current and previous frame
	diff_2	Head, RH, LH	diff (Euclidean distance) squared
Group 3	tri_center_x		x position of the triangle formed by connecting head and hands blobs
	tri_center_y		y position of the triangle formed by connecting head and hands blobs
	tri_center_distance	Head, RH, LH	Euclidean distance between the x, y pos. of the triangle center and the x, y pos. of the blob
	tri_center_angle	Head, RH, LH	angle of the blob from the triangle center
	tri_area		triangle area
Group 4	Q1	RH, LH	dichotomous flag indicating if the blob is in quadrant 1 in the current frame
	Q2	RH, LH	dichotomous flag indicating if the blob is in

			quadrant 2 in the current frame
	Q3	RH, LH	dichotomous flag indicating if the blob is in quadrant 3 in the current frame
	Q4	RH, LH	dichotomous flag indicating if the blob is in quadrant 4 in the current frame
	angular_mvmnt_sum	Head, RH, LH	Sum of angular movement over 5 frames
Group 5	last_mvmnt_angle	Head, RH, LH	amount of angular movement between the previous and current frame
	C	Head, RH, LH	dichotomous flag indicating if the blob has remained stationary
	R	Head, RH, LH	dichotomous flag indicating if the blob has moved right
Group 6	Ur	Head, RH, LH	dichotomous flag indicating if the blob has moved up-right
	U	Head, RH, LH	dichotomous flag indicating if blob has moved
	Ul	Head, RH, LH	dichotomous flag indicating if the blob has moved up-left
	L	Head, RH, LH	dichotomous flag indicating if the blob has moved left
	Dl	Head, RH, LH	dichotomous flag indicating if the blob has moved down-left
	D	Head, RH, LH	dichotomous flag indicating if the has moved down
	Dr	Head, RH, LH	dichotomous flag indicating if blob has moved down right

Group 1 features are the original measures taken from tracking. Group 2 variables deal with differences in angles and x, y positions between the previous and the current frame. Group 3 features are all centered on a triangle generated by connecting the two hands and the head. The x, y position of the center of the triangle is meant to approximate the location of the center of the body. The tri_area is meant to judge the openness of a person's posture. The tri_center_angle feature is calculated from a horizontal line which crosses the triangle center x, y position as shown in Figure 8a. Group 4 features are quadrants which are calculated from the head blob as shown in Figure 8b. A hand is in quadrant 1 when it is above the lowest point of the head blob. A hand is quadrant 2 when it is below the lowest point in the head blob and at least 1 head blob width left of the head center point. A hand is in quadrant 4 when it is below the lowest point in the head blob and at least 1 head blob width right of the head center point. Quadrant 3 is located between quadrants 2 and 4.

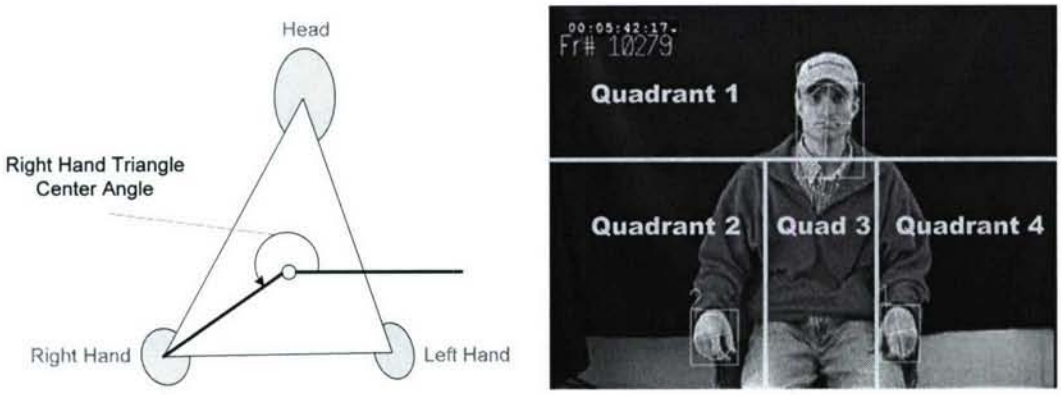


Figure 8. (a) Triangle center angle feature. (b) Quadrants (Meservy et al., 2005).

Group 5 addresses angular movement as shown in Figure 9a. It was proposed that angular movement would be important in distinguishing between illustrating gestures and self touching. The feature angular_mvmnt_sum is the total angular movement divided by the total number of frames. The feature last_mvmnt_angle is the change in angular movement from the previous frame to the current frame. For example, in Figure 9a, if the current frame is 4, the last_mvmnt_angle would be θ_2. Group 6 specifies binary directions that each blob may travel see Figure 9b). Since the number of directions each blob may travel is infinite, this group of features attempts to summarize all these possible directions into a manageable subset of directions. Each blob may only travel in one direction or may remain stationary between frames.

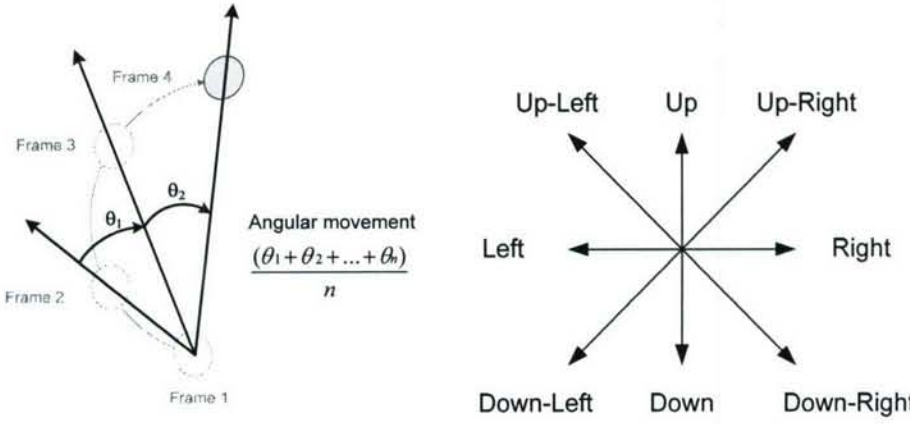


Figure 9. (a) Angular movement feature. (b) Binary direction features (Meservy et al., 2005).

Each feature described in Table 6 is recorded for each measured body part for each frame in a video segment. This level of granularity allows for detailed time series analysis. The features also permit summarization across a segment of time. In the current research, the mean and standard deviation were calculated for each feature within each time segment. This summarization tactic has been used successfully in previous classification efforts with this set of features (Jensen, Meservy, Kruse, Burgoon, & Nunamaker, 2005; Meservy, Jensen, Kruse, Burgoon, & Nunamaker, 2005; Meservy et al., 2005). The interpretation of the average (mean) is straightforward. The variance of each feature is a direct measure of how much that feature deviates from the mean. For example, if one is interested in the amount of movement of the right hand, the variance of the right x and y positions are a good indication. With the combination of

the (Feature) x (Body Part Measured) x (Summarization), there are 154 total features that are eligible to be used to discriminate truth from deception.

B. Human Behavioral Observation

Human behavioral observation was conducted by undergraduate or graduate students at UA, MSU, or University of Texas-San Antonio who were blind to the experimental conditions from which the video or audio recordings were taken. Below is a description of the manual coding requirements, followed by description of the automated tools that were developed and employed.

1. Manual Coding Requirements

The process of behavioral coding currently requires training human coders to view and listen to recorded interaction between truth tellers and deceivers and to note a host of variations in language usage, specific content details in text messages, minute changes in the voice from audio recordings, and small movements in audiovisual recordings. This is an extremely time- and labor-intensive that involves reviewing audio, video and text for indicators of deception and truthfulness. Lack of valid and reliable physical instrumentation to measure many communication features has meant that most coding is done by trained human coders. Limits of human cognitive ability, mental fatigue, and requirements for independence of judgment and statistical reliability together necessitate a large cadre of coders, multiple passes at coding any segment of potentially deceptive behavior, and a largely serial coding process. To assure independence of judgments and to assess statistical reliability, multiple coders must be used for each indicator and will specialize in one class of cues at a time. To avoid fatigue and achieve highest accuracy, a coder can be expected to rate no more than 5-6 indicators during a session, and no more than 2-3 hours per session. For full audiovisual samples, there are at least three classes of indicators to be coded. Typically, one set of coders will focus on audio indicators, while another set of coders concentrates on visual indicators, and a final set of coders focuses on linguistic features, yielding as much as a 50:1 ratio of coding time to discourse time for any one class of indicators being coded. Thus, to manually code a single 10-minute interchange between two people on all the behaviors currently under consideration may take 20-30 hours of coding time. To code a single experiment with 100 pairs of subjects would then require as many as 3000 hours of coding time. As the size of the interacting group or the number of groups increase, the coding task also expands exponentially.

Training coders is itself also a time-intensive activity, as coders may require up to 40 hours of training to learn a particular coding system (e.g., the Criteria-Based Content Analysis System) and another 10-20 hours of practice coding until they achieve acceptable levels of reliability. The coding task is made more complicated by varying skill levels of coders and by attrition. And for many behaviors, even the most experienced coders are not calibrated well enough to detect deceptive behaviors with complete accuracy. For example, many vocal attributes are more precisely measured by acoustic instrumentation than by human coders (Rockwell, Buller, & Burgoon, 1997). The human coding effort is therefore a monumental one that could be significantly expedited by acquisition of newer tools that ease the process of locating files and segments to be coded, that enable simultaneous coding, that replace subjective human judgment with objective instrumentation, and that automate some aspects of analysis.

2. Behavioral Observation Software: C-BAS

A number of kinesic, proxemic and vocalic cues were coded by trained human coders using two behavioral observation software packages. The initial coding was done using the commercially available program Behavioral Observation system from NOLDUS. Because this program proved difficult to use, the Center for the Management of Information programmers wrote their own program to replace the NOLDUS system. As one of several components of the Agent99 suite of tools, the Behavioral Annotation System was written in C# and called C-BAS for short (http://projectserver.cmi.arizona.edu/cbas/). C-BAS was designed to accommodate the need to record both macroscopic and microscopic behaviors patterns, to record both frequency counts and durations, and to obtain subjective judgments across time.

The layout provides the human coder with a simple yet robust interface that allows the coder to easily focus on the source material at hand. Figure 10 shows a screenshot of the C-BAS interface. In the left-hand window, the video to be observed is displayed. In the right-hand window, a user-defined template of keys and their definitions is provided. For example, if a coder is focusing on left- and right-hand adaptors, those key assignments will appear as a reminder to the coder. In the lower half of the screen, each coder key press is recorded with its time stamp to provide a complete chronological listing of the behaviors as they appear. The screen capture also shows the pop-up box in which coders can record their scoring of subjective measures at set intervals.

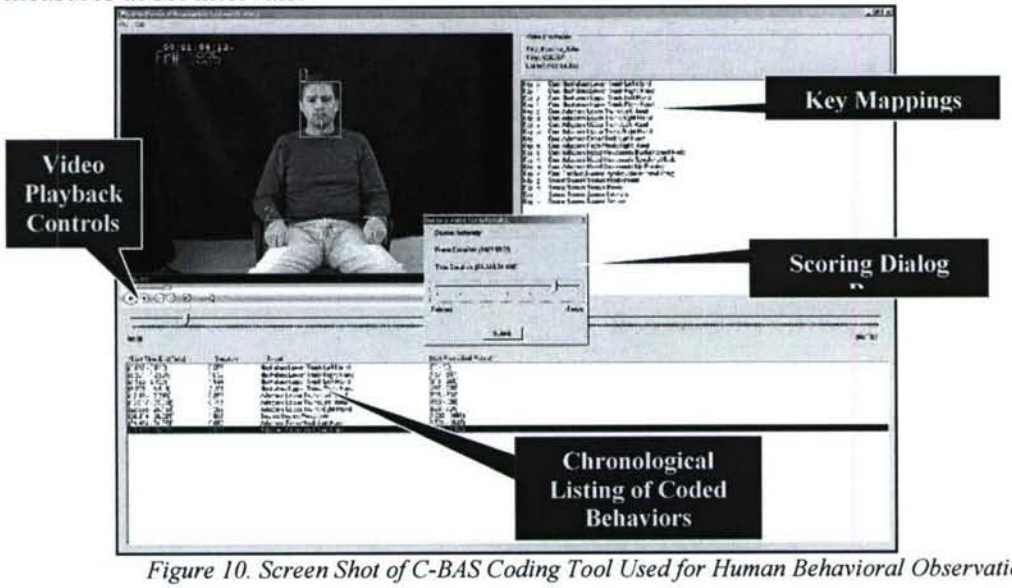


Figure 10. Screen Shot of C-BAS Coding Tool Used for Human Behavioral Observation.

Although the system was developed to specifically aid in the coding of the behaviors of humans engaged in deception, it can easily be modified for use in many other areas of research. C-BAS was designed to provide a balance between flexibility, usability, and a low overhead for users. One key feature of C-BAS is the ability to export the coder's data in XML files, providing an easy, common file format for exchanging data files between C-BAS and other analysis programs.

3. Behavioral Observation Coding Systems

For objectively observable behaviors, human coders were instructed to either do a quick key

press to record each time a brief behavior (such as a nod) occurred or to hold down the assigned key for the duration of longer behaviors (e.g., talk time, illustrator gestures). The specific kinesic behaviors that were coded and their definitions are presented in Table 7.

Table 7. Human-Coded Micro-Level Behavioral Cues.

Adaptors – Hand or facial gestures intended to relieve physical or psychological stress.
1. Self-adaptors: occur when a person brings a hand into contact with their own body, such as scratching or picking lint off of clothing. Hands touching the face were coded as a separate variable, and left- and right-hand adaptors were sometimes differentiated. 2. Lip adaptors: biting, pursing, scrunching, or licking of lips
Gestures – Hand movements that accompany and complement the speech stream.
3. Illustrator gestures: hand movements that accompany and complement the speech stream, including iconics, beats, metaphorics, and cohesives 4. Hand & shoulder shrugs: a specific upward movement of the shoulders or quick rotation of hands outward, with palms open, and back to their original position. Thought to signify uncertainty 5. Emblems: symbolic gestures with specific referents and culture-specific meaning that can substitute for speech (e.g., the AOK sign, the thumbs up sign). Often included with illustrators because of their infrequency.
Head Movements – All movements of the head while in the speaker or auditor role
6. Speaker head movement: nods, shakes, beats, and other head movements accompanying the speaking turn. Includes parakinesic movement that supplies punctuation, signals tense, and serves other syntactic functions 7. Listener backchannel movements: head nods or shakes while in the listener role

For subjectively rated features, coders used a 9-point rating scale (e.g., from not at all to highly involved) and made multiple ratings at specific intervals, such as at the end of an interview question. They were instructed to make their judgments by comparing the materials to how they thought normal people would behave in an interview. Normal behavior was to be rated as the midpoint of the scale. Definitions and instructions for subjective judgments appear in Table 8.

Table 8. Human-Coded, Macro-Level, Subjective Cues.

Involvement–concerns the degree to which the individual seems to be cognitively, emotionally and behaviorally engaged in the interview. People who are involved in an interaction should appear to be interested, attentive, alert, and responsive to the other individual. Those who are uninvolved should appear disinterested, apathetic, distracted, withdrawn, and detached. A person's level of involvement or detachment may be evident through language, voice, and "body language." It is manifested by the following:	
Immediacy–behaviors signaling psychological closeness/approach or psychological distance/avoidance. High immediacy is expressed by:	close proximity/conversational distance
	forward lean
	more frequent and longer gazes/more total mutual gaze
	direct body and face orientation
	use of touch and more intimate or familiar forms of touch
	verbal immediacy–language that uses present (rather than past or future) verb tense, active rather than passive voice, fewer modifiers that qualify or increase uncertainty
Altercentrism–behaviors signaling attentiveness to other (versus egocentrism–high self focus), including:	gaze toward speaker
	postural stillness
	absence of adaptors
	direct body and facial orientation
	backchannel cues of interest and support
	less talk time than the interlocutor
	few interruptions (though overlapped speech may indicate supportiveness)
	group versus self references
Expressiveness–degree to which a person is verbally and nonverbally animated or "flat" and inexpressive, including such indicators as:	frequent illustrator and emblems
	animated facial expressions
	parakinesic head movement
	variation in tempo, loudness and pitch (not monotone)
	louder, more rapid speech
	resonant voice
	vivid, intense language
	emotiveness (high proportion of adjectives and adverbs to nouns and verbs)
	use of metaphor
	self-synchrony
Conversational management–extent to which speaker contributes to a smooth, nonchoppy interaction. Indicators include:	adaptation--mirroring, matching, reciprocity, and interactional synchrony
	fluent speech (few filled and unfilled pauses, few "ah" disfluencies, few "non-ah" disfluencies
	smooth turn switches/short response latencies (no overlong switch pauses)
	verbal coherence mechanisms
	verbal coherence mechanisms
Dominance–One of the primary dimensions of interpersonal relationships. All interactions and relationships can be scaled along a dominance-submission continuum. Interpersonal dominance occurs to the extent that one person exerts influence over another and the other acquiesces, i.e., there is a one-up/one-down pattern between them. Personality dominance occurs when an individual attempts to exert influence over others, regardless of their response. It is sometimes called domineeringness. Here we will consider dominance and will view it in a value-neutral sense. The degree to which the behavior is harsh or hostile, or pleasant or cordial, will be measured by the pleasantness dimension. Indicators to consider:	

Conversational control–individuals may dominate conversations by:	interrupting and taking the speaking role
	talking more frequently and longer
	using direct, unwavering eye contact to gain the floor
	using expressive turn-requesting cues (e.g., raised finger, gesticulating, forward lean)
	using emphatic turn-denying cues (e.g., continued gesturing, louder voice, maintained intonation, gaze aversion while speaking)
	postural shifts
	initiation of topics and changing of topics
Strength and potency–individuals may convey their physical strength, mental prowess, access to resources, etc. through:	more rapid speech
	deeper voice
	stares, glares or unwavering looks
	visual dominance ratio
	expansive gestures
	expansive, open postures that take up more space, seem more "planted"
	all the other features associated with expressivity
	falling rather than rising intonation
	powerful (rather than powerless speech)
	Argumentativeness
	intense language
	intense language
Relaxation–this dimension captures the extent to which a person seems relaxed, calm and poised. The opposite is being tense, nervous, and nocomposed.	
Indicators include:	moderate to slow speaking tempo
	fewer speech disfluencies
	absence of glottal "fry" in the voice
	relaxed muscles in face, jaw, shoulders, arms
	absence of adaptor gestures
	absence of random trunk and limb movements
	asymmetrical posture
	presence of relaxed laughter; absence of nervous laughter
	smooth rather than jerky gestures
	smooth rather than jerky gestures
Activation–this dimension should reflect the degree of arousal and physical activity the person is exhibiting.	
Indicators include:	frequent postural shifts
	Frequent trunk and limb movement
	frequent parakinesic head movement
	frequent gesturing
	rapid movements (e.g., gait, gesturing)
	rapid movements (e.g., gait, gesturing)
Pleasantness–the pleasantness dimension captures the valence dimension of the interaction, the positive or pleasant hedonic tone.	
Indicators include:	"Positive" facial emotions

	Smiling rather than frowning
	warm, resonant voice
	affirmative backchannels (e.g., head nods)
	laughter
	positive rather than negative emotion words in vocabulary

4. Linguistic Analysis Coding Systems

Human coders were also trained to use a variety of coding systems for linguistic and content-related features. Training consisted of over 40 hours of instruction, practice sessions, and feedback in identifying features found in transcripts and text-based communication. The specific coding systems used by trained human coders were CBCA and Reality Monitoring, to which we added other features that had been identified in the literature as possible indicators. Additionally, a software program, Grammatik, was used to obtain initial linguistic features. CMI developed another software tool, the Analyzer, to assist coders in recording and compiling their observations of verbal behavior. The coding systems and Analyzer are described below. Subsequently, we converted to use of automated tools for all verbal analyses.

a) CBCA

Criteria-Based Content Analysis (CBCA) is a sub-component of the Statement Validity Analysis technique that was originally designed to assess the statements of alleged victims of child abuse (Stellar & Koehnken, 1989). This technique was designed to determine if the statements of an alleged victim reflected experienced events or were generated through coaching by someone else. Nineteen criteria are used to score transcripts according to the presence or absence of cognitive and motivational features as well as general characteristics such as logical structure. The approach follows the Undeutsch hypothesis that the more criteria that are met, the more a statement is deemed as truthful. Recent research has applied the CBCA to detecting deception in a variety of circumstances and has included the application of the technique to adults.

In the current investigations, transcripts were coded for the presence or absence of 19 criteria shown in Table 9. Two independent coders coded each transcript on a three-point scale after a brief training session, where 0 indicated that the criterion, 1 indicated it was present, and 2 indicated it was strongly present.

Table 9. CBCA Criteria and Definitions.

Criterion	Definition
Logical Structure	The statement does not contain contradictions or logical inconsistencies
Unstructured Production	Events are sometimes presented in an unsystematic, chronologically disorganized fashion
Quantity of Details	The event, location, and surroundings are described in great detail
Contextual Embedding	The event is described in relation to locations, times, and relationships
Descriptions of Interactions	Description contains descriptions of actions and reactions
Reproduction of Conversations	The statement contains specific accounts of conversations (e.g., "I said" and "he said")
Unexpected Complications	The event does not unfold in the 'normal' way

Unusual Details	Unexpected or surprising details
Superfluous Details	Details that are not absolutely needed to describe the event
Misunderstood Details	The details are accurately reported, but the witness does not understand the meaning or function
Related External Associations	Describes events that are related to the issue but that are not absolutely part of the issue being described
Accounts of Metal State	Reports changes in feelings or thoughts
Perpetrator's Mental State	Descriptions of the emotions, cognitions, or motivations of the offender
Spontaneous Corrections	Corrects previous statements without prodding
Admitted Lack of Memory	Expresses concern that he or she cannot remember all relevant details or that certain details might be incorrect
Raising Testimony Doubts	Indicates that part of the account is odd or that he or she can hardly believe the accounts
Self-deprecation	Mentions unfavorable or incriminating details
Pardoning the Perpetrator	Excuses the behavior of the accused
Characteristic Offense Details	Describes elements that are typical for the type of crime but are not generally known by the public

b) Reality Monitoring (RM)

Reality Monitoring (RM) was originally developed as a technique to discriminate between memories that were produced by external experiences of actual events and those that were produced by internal or imaginary experiences. It operates from the general hypothesis that externally generated events should be rich in sensory and contextual information; whereas, internally derived memories will contain more references to cognitive operations. These general theories have been extended to deception research.

Two coders evaluated transcripts for the presence of seven RM cues. These features and their definitions are listed in Table 10. Each was coded objectively for the total number of times that it is present in a transcript and subjectively on a 7 point scale, with lower scores indicating that the item was less present than higher scores.

Table 10. Reality Monitoring Criteria and Definitions.

Cue	Definition
Visual details	Things the person saw
Sound details	Things the person heard
Taste/Touch/Smell details	Things the person tasted, touched, or smelled
Spatial information	Details about where the event occurred or how things were located relative to each other
Temporal information	Details about the time or time order of the event
Affect	Descriptions of emotion
Cognitive operations	Descriptions of thoughts or though processes (I must have because …)

c) Grammatik

Grammatik is an automated text analysis tool that is part of Word Perfect. It tags parts of speech and calculates a variety of linguistic characteristics of a given document. Both transcribed

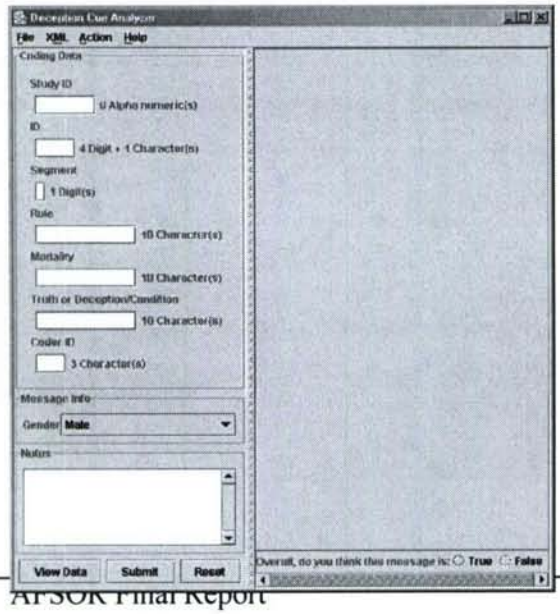
conversations and written documents were subjected to this analysis. Table 11 presents the features that were generated by the program.

Table 11. Linguistic Features Calculated by Grammatik.

Feature	Definition
# of words	The total number of words in the document
# of sentences	The total number of sentences in the document
Short sentences	Sentences that are less than 60 words long
Long sentences	Sentences that are more than 60 words long
Simple sentences	Sentences lacking dependent or independent clauses and phrases
Big words	Words that are 8 characters or longer
Average number or syllables per word	# of syllables/ number of words
Average words per sentence	# of words / # of sentences
Flesch-Kincaid readability index	206.835 – (1.015 x Average sentence length) – (84.6 x Average number of syllables per word)
Passive voice	Sentences written in passive voice
Vocabulary complexity	# of multi-syllabic words
Sentence complexity	# of clauses and phrases per sentence
Total # of flagged errors	Count of spelling, grammatical, punctuation errors
Missing modifiers	Counts of missing articles, adjectives and adverbs that usually precede specific nouns or verbs
Tense change	Changes from one tense to another (e.g., past to present)

5. Linguistic Analysis Software: Agent99 Analyzer

The Analyzer was developed to facilitate the manual coding of linguistic cues from text. Similar to C-BAS, Analyzer provides a simple interface and flexible architecture that enables the analysis of a wide variety of linguistic features. It provides the ability to develop custom templates in XML. These templates are used by the coder to analyze specific linguistic features. As with C-BAS, the templates can be configured to capture a wide variety of information by the coders. Figure 11a shows the basic Analyzer screen before any specific template is selected.



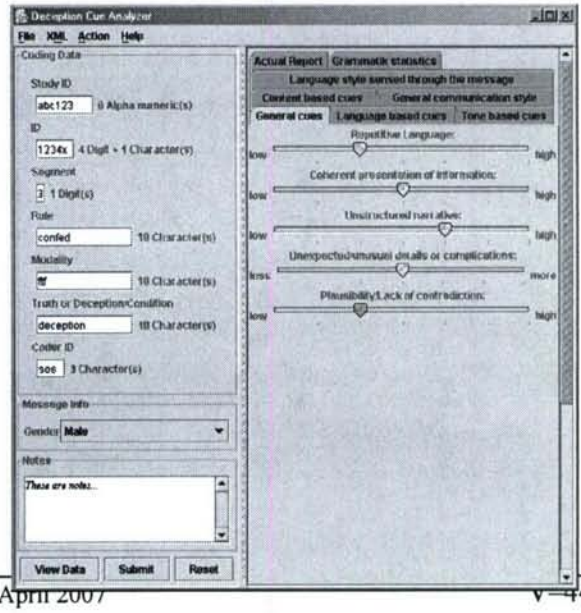


Figure 11. Screenshot of (a) Agent 99 Analyzer(A99A) Interface and (b) Sample of Recorded Codes and Ratings.

The coder then selects a specific template to use, which loads specific features and associated ratings into the Analyzer tool. Figure 11b illustrates what Analyzer looks like after a template has been selected and cues or ratings have been recorded. On the right side of the screen, the coder is presented with rating scales regarding various attributes of the text under examination. In this particular example, the coder is being asked to rate the level of repetitive language, the level of coherence of the information, and so forth. The coder then adjusts the sliders accordingly and submits the ratings.

Lastly, as with C-BAS, the coder's data from Analyzer can be easily exported to XML. XML enables smooth interoperability between different analysis packages. This exported XML is illustrated in Figure 12.

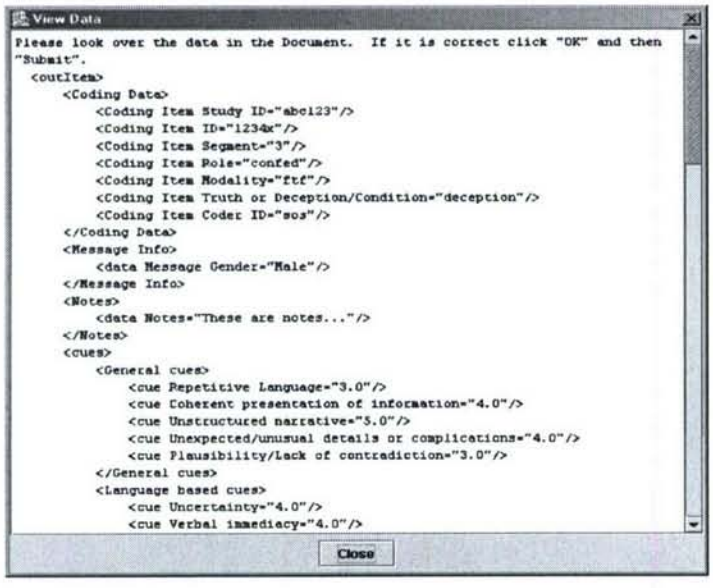


Figure 12. Screenshot of XML Output from A99A.

Both C-BAS and the Analyzer provide researchers with the opportunity to identify hierarchies of cues or features for analysis, enabling coders to manually identify and flag the features with precision down to the second. These two tools allow for manual examination of audio, video, and textual content.

VI. Deception Detection Integrated Multimedia System

As part of the Department of Defense University Research Instrumentation Program, CMI obtained funding to create a Deception Detection Integrated Multimedia System (D-DIMS) to digitally capture, record, store, index, code, retrieve, freeze-frame, and edit high fidelity speech, text, and visual media used in our deception detection research and training. These functions were integrated to capture data from sources such as email, audio transmissions, broadcasts, and

videoconferencing transmissions as well as to measure and analyze configurations of verbal and nonverbal deception indicators in near real time.

More specifically, the D-DIMS, as an integrated multimedia management system,

- Includes video/motion, audio, and lighting subsystems to record deceptive communication,
- Enables synchronous or asynchronous observation, playback and behavioral coding,
- Houses and index all of the data and recorded media from local and remote research sites,
- Allows access to recordings and data among remote archives and research, and
- Establishes a foundation for a national deception and denial repository.

The integrated audio, video, and storage components of the D-DIMS have enhanced significantly the ability to investigate reliable indicators of deception in a comprehensive manner with unprecedented level of granularity. D-DIMS is housed within the Deception Detection Laboratory (DDL) constructed at the University of Arizona to support its Department of Defense research. This specialized research environment is optimal for the detailed study of deception detection in a controlled environment. The various components and capabilities of D-DIMS are shown in Figure 13.

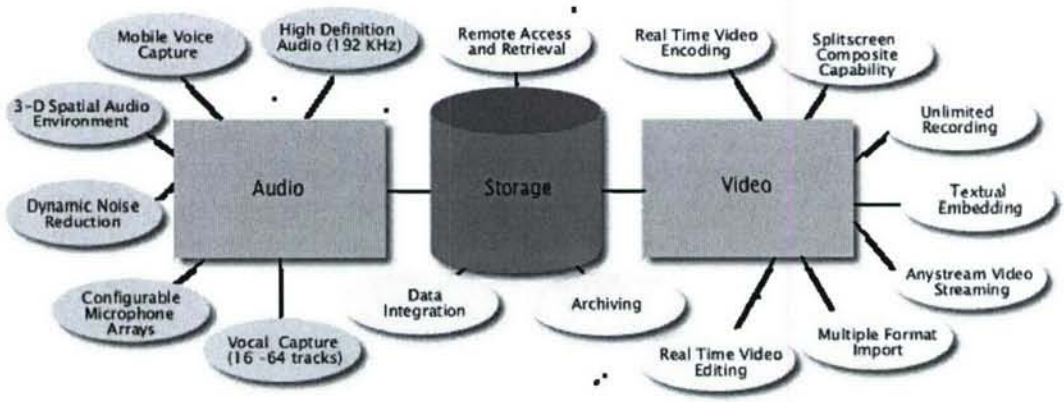


Figure 13. The D-DIMS System.

A. Media Capture and Storage

D-DIMS supports our ever-expanding storage requirements through providing 6 Terabytes of online disk-based storage capacity in a Network Attached Storage (NAS) device and 16 Terabytes of offline storage using a tape storage device. Currently, CMI's storage scalability is limited to a single 6 TB server with 3 TB already in use for existing data. Having multiple TB servers with fast tape storage, the CMI research team can easily save the most frequently used media on the Terabyte server, while saving all complete deception detection media on tape for future editing, research and recall. Through incorporation of a NAS, researchers can obtain flexible data storage connected to multiple servers and applications while enhancing data security through physical separation of hardware and software, as well as accelerate the speed of data storage and retrieval (Apicella, 2001). The NAS interface allows remote sites to access to all

resources incorporated in the storage system. The disk storage system can expand to an almost unlimited amount of tape storage if media storage needs increase.

Table 12 illustrates how the data storage requirements balloon. In a typical four-person group, to record just high definition audio for a group of four people requires 4.8 GB/hour. When video is involved, the data capture procedure requires four cameras, one for each group member plus a camera for the overall group. One can understand how high fidelity video and audio data can exhaust the resources of a terabyte server and quickly necessitate an appropriate archival system. Just one hour of video from 1 camera is 13GB. Several of our current deception detection experimental designs involve four people interacting at once and typically require 52 GB of storage. From an audio standpoint, to store 200 minutes of audio, researchers require 1 GB of space. In our experiments, speech is recorded for each individual as well as for the entire room. To record 10 people, researchers need 3GB – 12GB of storage per hour (3GB for CD quality, 12GB for HD quality). To support multiple cameras and capture multiple video and vocal recordings, the research requires a storage network to support large throughput with a high storage capacity.

Table 12. Multimedia File Sizes.

Time	Number of Cameras	Size	Audio	Total
1 hour	1	13 GB	1.2 GB	14.2 GB
1 hour	5	65 GB	4.8 B	69.8 GB
15.4 hours	5	931.2 GB	68.8 GB	1 TB

Additionally, to meet the training objectives proposed in the original grant, the research team accumulated and archived enormous amounts of communications (transcripts, audio-recordings, video-recordings) from previous field and laboratory experiments as well as samples of naturally occurring discourse that include deception. Along with transcripts and recordings from current experiments, these materials were coded to identify the most reliable verbal and nonverbal indicators of deceit or truthfulness, and relevant exemplars were indexed and edited for use in training programs. The sheer amount of data accumulated throughout the project, from experimental sessions alone, required several terabytes of storage capacity and eliminated the need for constant conversion of tape-recordings to hard storage media (e.g., CDROM) in compressed formats that incur significant loss in resolution.

A top priority is to increase accuracy in distinguishing truthful from deceptive information in a near real-time environment. This requires precise instrumentation for rapid and dynamic computer-assisted analysis of communication streams. Although development of *Agent99* relies on what are regarded as state-of-the-art, well-validated methodologies, data collection, and analysis techniques, current methods are not yet amenable to processing real-time messages. Individual communications are captured as analog recordings or with tape-based digital instruments that must then be compressed and edited before they can be coded and analyzed serially. Recent research literature and our firsthand experience are also indicating that many vocal and motion-oriented deception cues are elusive and warrant deployment of more elaborate means of measurement than traditionally used. Daily advances in digital technology are increasing the fidelity, ease, automaticity, and simultaneity of behavioral data capture which could greatly accelerate the processes of coding and analyzing deceptive messages and could

record audio and visual behavior at a fidelity that can be used to develop algorithms for more accurate, automated behavioral analysis.

B. Media Editing and Indexing

Recorded messages were segmented into naturally occurring units or time-based intervals and indexed so as to retrieve usable segments for creating experimental materials, behavioral coding, training, and annotation. Although state-of-the-art equipment such as an Avid editor make video analysis, storage, and editing a manageable task, editing remains a laborious and complicated task. However, CMI's direct remote video distribution capabilities, the ability to extract video segments and analyze those segments synchronously in a split-screen mode, support for multiple video formats and multiple media streaming formats, and multiple-inputs recording capacity have placed CMI at the (technical) forefront of behavioral research. As a result, as the research endeavors continue into the future, CMI is poised to develop deception detection techniques that can be administered in "real-time." Unfortunately, the upper bounds on achieving such timely processing are set by existing computing technology and storage capacity.

C. Audio Capabilities

D-DIMS implements ProTools HD, a high definition multitrack digital recording system. Current audio capabilities are limited to capture up to 8 simultaneous voices. However, D-DIMS allows the capture of 16 voices simultaneously with the capability to expand to up to 64 voices. This extensibility will enable CMI's deception detection research to include not only dyadic and small group interactions but also larger groups and from multiple locations.

D-DIMS will provide an environment to overcome these barriers, enabling audio capture at a range from 44 KHz (CD quality) to 192 KHz (high definition audio), which provides the ability to discover and detect transitory audio cues such as pitch breaks and vocal inflection. High definition audio allows higher frequency capture for more detailed and lifelike sound, and provides more accurate data for acoustic frequency analysis. Also, the higher sampling rates produce fewer "artifacts" or errors in digital recording. With audio data at 48 KHz (currently used by most academic research facilities), it is difficult to properly represent the true analog wave through digital sampling. Capture at 48 KHz produces a large quantization error, or difference between the actual analog wave and its digital representation. This causes lost data. At higher sampling rates, the quantization error is much lower, allowing for a closer, more accurate representation of the analog signal. Additionally, 192 KHz is becoming a commercial and academic standard. It is inevitable that speech research will soon migrate to the new higher-quality standard as well.

In addition to audio representation, an important aspect of deception detection research is individual positioning within groups. Currently, the DDL spatial arrangement allows for people to sit in two groups of four. D-DIMS audio capabilities would enable many more research seating configurations than currently possible. These configurations could, for example, allow researchers to study the impact of personal space on deception. D-DIMS makes this possible through the incorporation of surround sound, playback, and processing which together create the 3-D audio spatial environment.

The 3-D spatial environment will also allow researchers to record a group in one room and simulate people speaking from those positions in another room. This will allow the virtual placement of people in a room without having them be physically present. For example, to increase the cognitive load of deceivers in experiments, researchers could simulate a crowded room and analyze its impact on vocal cues. Researchers could also run a control condition in which participants are telling the truth under clamorous conditions. D-DIMS has the capacity to screen out background noise or other extraneous audio components during experiments so researchers can truly discern the vocal cues of one participant over another through a dynamic noise reduction processor.

Finally, D-DIMS' high definition, quality microphones enhance participant mobility. Previously, participants had to wear physically attached, wired headsets to capture individual voice. The D-DIMS microphone configurations allow individuals to roam freely through a room without having a physically attached object interfering with their natural movements.

D. Video Capabilities

Several video components are included in the D-DIMS. The system enables our research team to import video from any source (e.g., video tape, DVD, or other media), as well as record videotape lectures, experiments, and training scenarios using a state-of-the-art digital studio. This feature further promotes collaboration on video projects and promotes synergy in collecting and sharing deception detection clips from various sources. Assembling a vast array of examples not only facilitates comparing and contrasting data from various experiments but also enables cross-validation of deception indicators emanating from a wide variety of sources under varying circumstances.

The editing bundle (Avid Xpress PRO Video Workstation) significantly enriches our experimental research capabilities by, for example, enabling us to edit and combine separate video segments into a split-screen composite (see Figure 13) for coding and training purposes. This permits display of a video segment of the same individual under truthful and deceptive conditions, thereby demonstrating uncharacteristic behaviors and discrepancies between baseline (normal) and deceptive communications and behaviors that are a key aspect of identifying deceptive communication. The D-DIMS also allow real time video editing and enables instant video modification, a process that previously took 1-4 hours and is typically performed on a regular basis. This editing bundle also facilitates the application of titles and labels directly onto video, a feature that facilitates indexing and retrieval of relevant segments and can be utilized in the *Agent99 Trainer* for instruction. For instance, in training a student on the deception cue of spatial language, a training scenario could have the phrase "avoidant language" flashing across the bottom of the screen at the point when a deceiver in the video uses indefinite pronouns, thereby alerting the student to the deceptive indicator.

Figure 13. Split Screen Mode.

To address our video distribution requirements, D-DIMS encodes video in real time for direct distribution to remote sites, enabling researchers to avoid long processing times associated with software-based video systems. D-DIMS software also allows authorized, outside groups to remotely access the Network Attached Storage (NAS).

The system's video capabilities allow CMI researchers to bypass the use of tapes that limit recording sessions to 60 minutes. With that exclusion, recording capabilities become unlimited, which is critical in ensuring the experiments are not interrupted to simply change recording media.

Finally, state-of-the-art digital cameras and upgraded lighting included in the D-DIMS allow a higher level of visual and auditory detail, which is needed to uncover microscopic and fleeting deception detection cues. Furthermore, moving from two to three dimensional digital capture and analysis provides a new level of deception detection capability. Digital video equipment that records at a true 30 frames per second progressive scan allows CMI researchers to record, index, and rate potentially relevant deception detection indicators such as changes in respiration and vocal hesitations, feigned smiles that are only evident from barely discernible muscle movements around the eyes and corners of the mouth, and reductions in gesturing or other gross body movements like foot tapping. D-DIMS captures facial and gross body movements and be analyzed in real-time with algorithms to provide deception detectors with information to further probe deceivers.

E. Distributed Access

The D-DIMS also helped address the challenges of sharing files and instrumentation for behavioral coding of verbal and nonverbal communications across multiple sites (Michigan State University, Florida State University, Baylor University, the Air Force Institute of Technology, and Rutgers University). The D-DIMS permits distributed partners to archive, index, edit, and retrieve this massive store of data from remote locations.

The Deception Detection Laboratory in which the D-DIMS is housed is flexibly configured to allow for rapid collection, review and dissemination of data. As such, the component instrumentation systems integrated into the DDL work seamlessly within the requirements of the research program. D-DIMS is designed to simultaneously gather synchronized data from a number of digital, audio and video sources. In addition to its requirements for high reliability, high bandwidth and high fidelity data capture, D-DIMS also provides the capabilities to be unobtrusive and largely invisible to the subjects to avoid instrumentation artifacts.

VII. Tests for Reliable Indicators

A major objective of the research program was to identify reliable indicators of deception and truth. In order to assess the consistency and generalizability of indicators across a variety of contexts, 15 laboratory and field experiments, with a total of 2530 participants, were conducted at University of Arizona, Michigan State University, and Florida State University. The methods of the laboratory and field experiments are described below, followed by detailed descriptions of linguistic, audio, and kinesic-proxemic findings. Studies of moderators appear in Section VII.F.

A. Methodology of Laboratory and Field Experiments

1. Desert Survival Problem Experiments

Two laboratory experiments ($N = 60$, $N = 52$) performed at the University of Arizona utilized a decision-making task that involved a Desert Survival Problem. Participants were asked to imagine the following scenario: Their jeep had overturned in the harsh Kuwaiti desert and they needed to arrive at a consensus with their team mates on prioritizing items to salvage based on the items' value for survival. To aid their decision-making, participants were given a document, *Imperative Information for Surviving in the Desert*, to read prior to their decision-making discussion so that they would have relevant information to utilize in their discussion. Unbeknownst to half the pairs, their partner was enlisted to give deceptive information and arguments that would lead the team to make poor decisions contrary to what experts would advise. Discussions were conducted through email (different-time, text-based communication) or text chat (same-time, text-based communication). Following discussion of each of the 12 items that could be salvaged, individuals independently rank-ordered the items. DSP I limited the time frame for exchanging messages to 3 days and altered the task on the second and third days; DSP II relaxed the time restriction and required prioritizing salvageable items on each of three successive online meetings.

Subsequently, all text was submitted to automated linguistic analysis to determine which features separated truthful from deceptive messages and communicators.

2. Mock Theft Experiments

To simulate the kinds of deceptive contexts in which a crime has been committed, innocent and guilty suspects are interviewed, and interviewees are motivated to evade detection, a mock theft experiment ($N = 240$) was performed at Michigan State University. In this paradigm, some participants "stole" a wallet from a classroom; others were simply present during the theft. All participants were then interviewed by trained and/or untrained interviewers via chat, audio conferencing, or face-to-face interaction. A pilot experiment was first conducted to refine methods and examine deceiver experiences, behaviors, and detectability. In the main experiment, deception was examined under the three different modalities and different levels of motivation. Interviewer ratings, trained coder assessments of verbal and nonverbal behavior, and automated analysis of language, meta-content, and kinesics were all gathered.

3. Real-Jeopardy Interviews

Experimentally generated data often lack the high motivation and jeopardy found in real-world circumstances. To determine ecological validity and triangulate results with automated and human-coded behavior, a seventh investigation ($N = 25$) conducted jointly at the University of Arizona and Michigan State University consisted of secondary analysis of videotapes of criminal suspects who were questioned about bank thefts or similar crimes and for whom ground truth was known. A standard protocol, the Behavioral Analysis Interview, was followed. Manual and automated analysis of kinesic behavior was conducted on these interviews.

4. Deceptive Interviews Laboratory Experiment

This next experiment came from a series of investigations funded by the Army Research Institute's Research and Advanced Concepts Office to develop and test interpersonal deception theory. One of those experiments that were conducted at the University of Arizona was subjected to secondary analysis. The objectives were to determine generalizability of indicators found in the desert survival experiments, to expand analysis beyond text-based features to include nonverbal ones, and to delve deeper into how interviewer style influences deception displays. Nontraditional students and community members (N = 60) were interviewed by naïve interviewers. Interviewees responded to 12 questions during which they alternated between giving blocks of truthful and blocks of deceptive answers. Interviewers adopted one of three interviewing styles indicative of different levels of suspicion and involvement. The videotaped interviews were transcribed for automated linguistic analysis and manually coded on multiple nonverbal behaviors. Viability for automated nonverbal analysis was assessed but the analog videotapes were deemed inadequate for highly reliable automated analyses.

5. Resume Faking Experiments

Another context for deception detection is employment interviews for security forces. Three resume faking experiments (N = 316) were conducted at Florida State University to investigate deceptive communication under different modalities and different levels of suspicion. Business students were recruited ostensibly for a study of business interviewing and were randomly assigned to interviewer or applicant roles. Applicants submitted resumes that had been intentionally falsified. The interviews were conducted via text chat, e-mail, computer-mediated audio, or audio with chat. Half of the interviewers were warned of the possibility that the resumes had been enhanced; the other interviewers were not warned. Interviewer detection performance was compared to that of third-party observers. Text and audio are being manually coded and will be subjected to automated text analysis.

6. ScudHunt, BunkerBuster, and StrikeCom Experiments

Virtually no research has examined deception under conditions of attempting to deceive multiple receivers and using different communication modes. To analyze deceptive communication in chat, audio, and face-to-face communication and to take into account the greater complexity of expanded team size, we first utilized a game developed by DARPA (ScudHunt) for ARI-sponsored research on leadership and shared visualization. A number of challenges posed by the software and lack of flexibility of the game led us to develop our own versions of military game scenarios. Our first product, BunkerBuster, was a four-person, computer-based game similar to *Battleship* in that team participants controlled various information assets that were to guide a series of decisions about where to search for enemy bunkers located on a grid and where to strike to destroy enemy fortifications. Coordination between team participants involves negotiating where to deploy their respective information assets to conduct searches and then reporting back their findings through several search iterations before formulating a final strike plan. Deception is introduced by enlisting one team member to make wrongful reports from his/her assets' data. ScudHunt and BunkerBuster (N = 110) produced analyses of the patterns of communication by truthful and deceptive team members and revealed, as expected, that the presence of deception undermines group performance.

BunkerBuster paved the way for StrikeCom, which is an online, turn-based simulation of a C3ISR (Command, Control, Communication, Intelligence, Surveillance, Reconnaissance) task. The object of the game is for the three-person teams to find and destroy enemy camps that have been hidden on a game board. Like its predecessors, each player controls different intelligence assets. StrikeCom was designed, built and pilot tested during the first year before being revised and upgraded again for final experimental use.

Three experiments were performed by the University of Arizona, Florida State University, and the Air Force Institute of Technology ($N = 655$). Participants in some experiments were U. S. Air Force ROTC cadets who used StrikeCom to conduct mock air operations. In some games, one team member was instructed to be deceptive and purposefully mislead the team away from the enemy camps. StrikeCom also served as a platform for capturing deceptive data in numerous modalities: face-to-face, distal groups, chat room, and voice. In other games, one team member was also made suspicious. All interactions between team members were recorded.

7. Group Resource Allocation Task

Two additional experiments explored deceptive computer-mediated communication under the conditions of attempting to deceive multiple receivers at once ($N = 234$). Additional interests were the influence of proximity between team members and the impact of being made suspicious about possible deceit. Groups conducted a resource allocation task. Dependent variables of interest, beyond the choice of language and content, were the amount of deception voluntarily submitted during group discussions and the success of deceivers in undermining group performance.

8. Enron Field Study

The classification and identification of email messages using more than simple keywords is a difficult task. Previous efforts in the automation of classification based on truth/deception have met with some success. This study ($N = 58$) attempted to show how similar methods might be used to classify authors of text messages in the publicly available Enron email corpus as ingroup members (part of the criminal fraud conspiracy) or outgroup members (uninvolved with the Enron fraud). We defined members of the ingroup as being all Enron employees or associates who pled guilty, were convicted, or are awaiting trial for crimes related to Enron's collapse. We defined members of the outgroup as being all Enron employees or associates who are not members of the ingroup.

The Federal Energy Regulatory Commission (FERC) provided public access to information released in the Western Energy Markets investigation, specifically the Enron investigation. Among other resources, FERC's iCONECT 24/7 portal provided access to Enron emails and included 92% of Enron staff emails. Several versions of the Enron email corpus were available. We used a version created by Jitesh Shetty consisting of the original William Cohen dataset converted to a MySQL database with several duplicate messages removed. This database contains a table identifying 151 distinct email senders. However, by parsing the 'from' address of the messages contained in the database (most of which contain the sender's first and last name using Enron's email address naming convention) we were able to identify 5,209 distinct email senders—including almost all employees we had identified as being members of the ingroup.

Ingroup messages are messages sent by a member of the ingroup only to other members of the ingroup. Outgroup messages are messages sent by a member of the ingroup to at least one member of the outgroup. Messages sent to multiple recipients where at least one recipient is a member of the outgroup were considered to be outgroup messages. We identified 29 ingroup messages and over 600 out-group messages by performing queries against the Enron e-mail database. To make the ingroup and outgroup sample sizes the same, we took a random sample of 29 of the out-group messages for comparison with the 29 ingroup messages.

9. Security Force Squadron Police Statements

To test the applicability of automated linguistic analysis to statements from actual suspects and witnesses, a fourth investigation (N = 383), conducted at University of Arizona and Oklahoma State University, has entailed analysis of written statements collected at two air force bases during investigations by Security Force Squadron personnel. Statements relate to everything from on-base thefts to auto accidents. The statements have been automatically tagged and linguistic analysis tools were being tested for success in discriminating between statements from innocent respondents and statements from guilty ones.

10. Border Security Screening Field Interviews

To further test the generalizability of laboratory-generated data to field settings, videotaped secondary screening interviews (N = 33) were collected at the U.S./Mexico border in cooperation with Customs and Border Protection at the Dennis DeConcini Port of Entry located in Nogales, AZ. Among other duties, officers' are expected "to detect and prevent terrorists and terrorist weapons from entering the United States" (Bonner 2004). Over the course of several years, CMI at UA has developed a relationship with Customs and Border Protection personnel in Tucson, Arizona. After a year of responding to the various legal, jurisdictional, and practical issues, we received approvals from federal and local authorities and from the University of Arizona IRB to videotape interactions between CBP agents and border-crossers.

The typical, legal flow of entry into the U. S. at the DeConcini Port at Nogales, Sonora, Mexico is to (1) secure an appropriate visa at the US consulate, (2) cross into the United States on foot or in a motor vehicle, (3) if needed, apply for an extended stay permit, and 4) proceed to final destination in the U.S. CBP officers interact with entrants at the time they apply for entry and again if they apply for extended stay permits. Among illegal entrants, the four most common categories are impostors, use of fake id, oral false claims, and entries without inspection. CBP officers must make a judgment as to whether or not an individual is an illegal entrant or poses a risk to the U.S. To make this determination, they rely on a variety of techniques, including behavioral observations. When an applicant's conduct is deemed suspicious, they are sent to secondary screening and the Expedited Removal (ER) room where they are interviewed.

For ER interviews, suspects were seated in front of an agent. The ceiling-mounted camera was placed behind and above the head of the agent. Agents were consented prior to any taping, but subjects were consented after the interview. At the conclusion of the interview, officers completed rating scales to assess suspicion, and suspects were shown a video and optionally consented to release their videotaped interaction for use.

All video were digitized at Arizona. Interviews were edited down to the main interaction (excising periods of silence and brief activity such a fingerprinting). Trained bilingual human coders rated the interviews on global patterns of behavior (e.g., involvement, relaxation, pleasantness, submissiveness) to be used to cross-validate automatically generated results. They also identified audible disposition information (denial or granting of permit), confessions, and verified lies so that precise behavioral analyses within interviews could be conducted. The prepared videos were then be processed with the computer imaging techniques (blob analysis) described previously in collaboration with the CBIM at Rutgers University.

VIII. Reliable Linguistic Indicators

The general procedure for analyzing text-based documents follows two major steps, extracting features and classification, each with its own sub-steps. Extraction entails selecting the appropriate features to be examined (some of which depend on the type of discourse being analyzed) and calculating features over the desired text portions.

Classification follows these steps:

1. Manually classify documents.
2. Prepare data for automatic classification.
3. Choose appropriate classification method(s).
4. Train model on portion of data.
5. Test model on remaining data.

Wherever possible, these steps were followed. In some cases, multiple classification models were compared. In some cases, the sample size was too small to justify creating separate training and testing data sets. In such cases, alternative methods for cross-validating training models were employed.

1. Laboratory Experiments

The two Desert Survival, two Mock Theft, and Deceptive Interviews generated laboratory data for testing the effect of deception on various linguistic and meta-content features. The central hypothesis under test was:

> *Deceitful messages display more (a) uncertainty (more modal verbs and fewer modifiers) and (b) non-immediacy, and less (c) quantity, (d) complexity, (e) diversity, (f) specificity, and (g) affect than true ones.*

In order to investigate the reliability of the cues in three experiments, results for individual experiment are first present, and then summarized results across three experiments are shown below.

a) Desert Survival Problem Experiments

Hypothesis 1 was tested with multivariate analyses of variance and follow-up simple effect tests for each dependent variable. Due to the relatively small sample size (N = 30 dyads) in DSP I, analyses were conducted both with the dyad (sender-receiver pair) as the unit of analysis and

messages (N = 180) as the unit of analysis. Results indicated that, compared to truthtellers, deceivers displayed:

1. *more uncertainty* (more modifiers* and modal verbs* or complete absence of modifiers; exception: truthtellers used more adverbs, possibly to intensify verbs)
2. *less personalization/more nonimmediacy* (fewer first person singular—"I"—pronouns, fewer references to others,* more first person plural or group—"we"—pronouns,* more second person—"you"—pronouns, fewer possessives, more passive voice)
3. *more quantity* (more words, verbs,* and longer sentences, to the point of being wordy)
4. *less complexity* (simpler words, fewer compound and complex sentences*)
5. *less diversity* (less varied vocabulary, fewer content words*)
6. *possibly less specificity* (fewer spatial terms and overall sensory terms,* but more modifiers and superlatives or comparatives; truthtellers also use more ellipsis, perhaps to let receiver fill in the blanks)
7. *more negative imagery and affect but more positive pleasantness* (deceivers use fewer positive imagery* and more negative imagery terms, more negative affect relative to content, but also more positive pleasantness terms* and more emotiveness); receivers also matched deceiver language use, so may be a reciprocal effect
8. *more informality* (more misspellings, typos, and other grammatical errors both between and within dyads); receivers also matched deceivers' errors

In the above list, the asterisked items also emerged as key variables in the J48 decision tree analysis or discriminant analyses. With 11 predictors (including activation and negative pleasantness variables that had not appeared in the previous models), the model using messages as the sampling unit accurately classified 72% of the deceivers and 75% of the truthtellers (69% and 72% respectively in the cross-validated model). With subjects within dyads as the sampling unit, the classification improved to 94% of the deceivers and 79% of the truthtellers (88% and 64% respectively in the 10-fold cross-validated model). These results indicate strong promise for using linguistic features to separate truthful from deceptive communication.

DSP II replicated the first experiment with 26 dyads and 204 messages. The smaller sample size for dyads presaged reduced power and fewer significant effects. Nevertheless, where differences emerged, they were largely consistent with DSP I. Deceivers displayed:

1. *less certitude* (fewer modal verbs*, same as DSP I)
2. *less personalization/immediacy* (more group references* and multivariate trend on pronoun use, with highest and least passive voice*, same as DSPI)
3. *more quantity* (more messages, longer messages, more verbs*, same as DSP I)
4. *less complexity (*less complex vocabulary* and syntax, same as DSP I)
5. *more specificity* (more sensory terms* and temporal immediacy*, contrary to DSP I)
6. *less affect/expressiveness* (less imagery, lower emotiveness*, contrary to DSP I)

The items that were important variables in the J48 decision tree analysis and/or significant predictors in the discriminant analysis are again asterisked. With dyads as the unit of analysis, four variables predicted sender truthfulness or deceit: lexical complexity (average word length), modal verbs, content word diversity, and total number of verbs. These variables accurately

classified 85% of the deceivers and 85% of the truthtellers (77% and 85% respectively in the cross-validated analysis). With messages as the unit of analysis, two variables—lexical complexity and group references accuracy classified 86% of the deceivers but only 51% of the truthtellers (84% and 51% respectively in the cross-validated analysis). In this case, the better results obtained with the dyad, not the message, as the unit of analysis.

Most previous deception research suggests that deceptive messages should be shorter than truthful ones because deceivers do not have as many details to put into a message as truthtellers do. However, most of the literature focuses on statements of fact or recollections such as in criminal statement analysis. Results from both DSP investigations reveal a reversal of that pattern, with deceivers saying more rather than less. An informal review of some of the messages showed participants appeared to be trying to boost their credibility to make their proposed bogus rankings plausible. It also seemed that the deceivers were giving more elaborate reasons for their rankings while the truthtellers just ranked the items with short, common sense reasons or no reasons at all. The lengthy responses from deceivers were partly due to superfluous words or meaningless expressions, probably used as fillers to disguise the fact that they did not have much to say but still wanted to give the impression of completeness.

The affect and specificity results are also not consistent across the two studies, although it appears that truthtellers are more inclined to use imagistic terms, especially positive ones, whereas deceivers use more negatively toned ones. However, deceivers in the first study used more positively toned pleasantness language. The specificity results were completely at odds across the two studies, suggesting something specific to the modifications in the task may have influenced these cues. The moderators of these variables warrant further investigation.

A comparison of various classification procedures was also conducted (see Zhou et al., 2004) to determine whether classification accuracy could be improved with a particular method. The results are shown in Table 13. They indicate that no single method emerged as superior. The average performance on DSP I data were 78% for subject data and 71% for message data. Comparable figures for DSP II were 79% and 78% respectively. Not only did the messages capture important differences between deceptive and truthful responding but all but decision trees were consistent across data sets. However, pruning may improve decision tree performance, as other methods also showed substantial improvement when the predictor variable set was reduced to a smaller set of variables. Neural networks also performed well on both training and validation data, giving them a possible edge over the other methods tested. As for the choice between message or subject as the unit of analysis, the results favor aggregating across messages and using the subject (or dyad) as the unit of analysis, a procedure that also doesn't violate assumptions of independence and does not permit verbose individuals to skew the results.

In sum, the DSP results support the hypothesis that messages from deceptive senders differ systematically from truthful ones on numerous linguistic features and can, in combination, significantly classify people on their veracity using several different statistical and machine learning classification methods.

Table 13. Comparison of Classification Methods on DSP I and DSP II Data.

(a) DSP1

Classification methods	Discriminant analysis		Logistic regression		Decision trees		Neural networks	
Test methods	training	cross validation	training	cross validation	training	cross validation	training	cross validation
Overall performance								
Subject	86.7	76.7	100	83.3/83.3	96.7	76.7	100	80
message	76	74	82.3	78.1/79.2	95.8	79.1	92.7	80.2
Truth performance								
Subject	78.6	64.3	100	78.6/78.6	92.9	71.4	100	71.4
message	76.7	74.4	74.4	69.8/72.1	90.7	79.1	95.3	79.1
Deception performance								
Subject	93.8	87.5	100	87.5/87.5	100	81.3	100	87.5
Message	75.5	73.6	88.7	84.9/84.9	100	79.2	90.6	81.1

(b) DSP2

Classification methods	Discriminant analysis		Logistic regression		Decision trees		Neural networks	
Test methods	training	cross validation	training	cross validation	training	cross validation	training	cross validation
Overall performance								
Subject	84.6	80.8	84.6	76.9/76.9	96.1	65.4	100	88.5
message	75.7	72.3	75.5	72.3/74.4	76.6	57.4	79.8	74.5
Truth performance								
Subject	84.6	76.9	84.6	76.9/76.9	100	66.7	100	85.7
message	76.9	74.4	73.5	69.7/72.7	73	48.1	70.8	69.2
Deception performance								
Subject	84.6	84.6	84.6	76.9/76.9	92.9	64.3	100	91.7
Message	74.5	70.9	76.7	73.8/75.4	78.9	61.2	89.1	78.2

As regards the various statistical and machine learning classification methods, Figure 14 reveals that neural networks maintained the greatest consistency across tests, whereas logistic regression achieved the highest absolute accuracy.

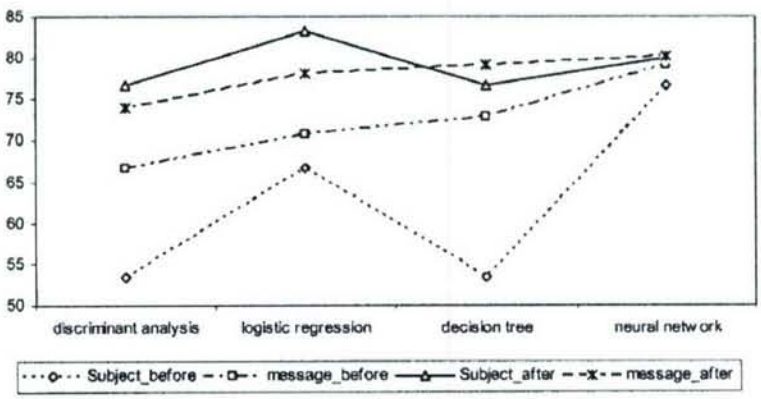


Figure 14. Classification Accuracy Before and After Pruning to Most Significant Cues, with Messages or Subjects as the Unit of Analysis.

b) Mock Theft Experiment

In the mock theft experiment, interviews were conducted either face-to-face or interviewer and interviewee were separated and communicated via text chat or audio communication. Because results can vary according to modality, and each modality can be viewed as a separate replication, results are reported for each one. Additionally, in advance of being questioned about the theft, interviewees were also asked about their favorite high school class and their most recent work experience (describing a typical day and explaining what they liked best). During these question sequences, everyone was to be truthful on the education questions but "thieves" were asked to deceive on the work questions, thus providing two blocks of questions on which they would deceive (work, theft) and on which innocents would be truthful.

Analyses were conducted across the three blocks of questions and also within the work and theft blocks. Results across the three blocks produced a number of interactions between block and guilt, indicating that results varied by question block. In general, deceivers in some respects behaved like truthtellers were predicted to behave. Compared to truthtellers, deceivers' language evinced:

1. *more certitude* (fewer modal verbs and modifiers)
2. *more personalization* (more self references but a mixed pattern on third-person references)
3. *more immediacy* (more temporal immediacy terms)
4. *lower message quantity* (fewer words, verbs, and syllables)
5. *less complexity* (lower average word length, fewer big words, fewer conjunctions, fewer long sentences, shorter average sentence length, less complex sentences)
6. *more diversity* (lexical and content word diversity, but a trend toward less syntactic diversity)
7. *mixed specificity* (more total specificity terms, especially on work questions, but fewer reality-monitoring spatial, visual and overall quantity of details)
8. *mixed affect/expressiveness* (more positive and more negative pleasantness terms but lower negative imagery score and reduced use of affect-laden language over time)
9. *more informality* (trend toward more total grammatical errors, although the difference from truthtellers diminished by the last block)

Thus, linguistic features did differentiate truthtelling from deceit but often opposite in direction from the DSP studies. Specifically, only complexity showed the same pattern throughout. Certitude, personalization/-immediacy, quantity, diversity, and specificity showed reversed patterns. Of the affect/expressiveness measures, only positive pleasantness was consistently higher among deceivers, whereas negative forms of expression, though more often associated with deceit, did not take the same form in each study and even the amount of affective language varied. Analyses within the last (theft) block in the mock theft experiment also revealed that many of the differences between truthtelling and deception had dissipated by this last block, suggesting an adaptation over time by deceivers.

Moreover, many of the above relationships were moderated by the modality in which the interview took place. However, when analyses were conducted within modality, relationships often were only suggestive trends ($p<.10$) due to dividing sample size among the three conditions and the lower statistical power that resulted. Within text, deceivers' language was more certain,

personalized, and specific but also less complex (simpler words and sentences and lower grade-level readability). Within audio, deceivers actually showed mixed complexity (higher average word length but lower readability grade level), and more diversity. Within FtF interviews, deceivers tended to use more negative pleasantness and activation terms.

Overall, these results offer weaker evidence of individual indicators separating truth from deception. However, the discriminant analysis classification models performed reasonably well when conducted within block and modality. In the text modality, five predictors—average word length, number of sentences, sensory terms ratio, negative imagery, and third-person pronouns—were able to correctly classify 75% of deceptive cases and 84% of truthful cases during the work questions. Three predictors—modal verbs, emotiveness, and positive activation—successfully classified 68% of the deceivers and 77% of the truthtellers during the theft block.

Within the audio modality, six predictors—third-person pronouns, number of verbs, positive imagery, passive verbs, negative imagery and first-person-plural pronouns—successfully classified 79% of the deceivers and 79% of the truthtellers during the work question block. Two predictors—temporal immediacy and average word length—successfully classified 59% of the deceivers and 77% of the truthtellers during the theft block of questions.

Within the FtF condition, four predictors—positive pleasantness, temporal immediacy, negative activation, and positive imagery—produced classification accuracies of 59% for deception and 86% for truth during the work questions and a different set of four predictors—negative pleasantness, affect ration, number of verbs, and positive activation—correctly classified 65% of the deceivers and 91% of the truthtellers during the theft block.

A combined model that included all three modalities produced only two predictors and performed less well than the separate models, classifying only 57% of the deceivers and 68% of the truthtellers. These results bolster the need to tailor models to the specific types of questions and modalities under consideration. Linguistic features also played a significant role in a multimodal model presented later.

In summary, the mock theft experiment again demonstrated the ability of linguistic features to distinguish truth from deception but produced a markedly different profile of relevant indicators than the DSP experiments. Moreover, the profiles differed by type of question and modality in use for the questioning. The differences in tasks, synchronicity and richness of the media in use, and possible incentives for successfully evading detection are all possible moderating factors that account for the differences across experiments and conditions. All warrant deeper exploration.

Clearly, no single profile of deceptive language is likely to be discovered. Consistent with IDT (Buller & Burgoon, 1996), deceivers will adapt their language style deliberately according to the task at hand and their interpersonal goals. If the situation does not afford adequate time for more elaborate deceits, one should expect deceivers to say less. But if time permits elaboration, and the situation is one in which persuasive efforts may prove beneficial, deceivers may actually produce longer messages. What may not change, however, is their ability to draw upon more complex representations of reality because they are not accessing reality. In this respect, complexity measures may prove less variable across tasks and other contextual features. The

issue of context invariance thus becomes an extremely important one to investigate as this line of work proceeds.

c) Deceptive Interviews

The experiment's interviewee responses were segmented into 12 interview questions. To provide a general picture of deceptive versus truthful discourse, the 12 segments were then aggregated into 2 blocks. The first block contains the averaged scores of cues recorded for questions 1-3 and 7-9, during which a given respondent was either giving all truthful or all deceptive responses. The second block contains aggregated behaviors during questions 4-6 and 10-12, i.e., the remainder of the interview during which truthtellers switched to deceiving and deceivers switched to telling the truth. The first block also represents early responding and the second, later responding. To parallel previous analyses, only the between-subject effects of each data section were tested.

In the early phase, many indicators were significant. Deceivers said less than truthtellers (fewer words, verbs, and sentences), used less diverse language (lower lexical and content word diversity), had lower specificity (fewer sensory and other specific details), were less certain (fewer modifiers), used fewer pleasantness terms and used less imagistic language but constructed more complicated sentences and words (more complex and compound sentences, longer average word length).

The second block produced far fewer significant differences. The categories of quantity, complexity, uncertainty, nonimmediacy, diversity, and specificity did not produce significant cues. Only affect showed a near-significant trend ($p=0.092$). Deceivers used fewer affect terms than truthtellers. This result implies that dynamic adjustments diminished differences over time.

d) Laboratory Experiments Summary

Table 14 shows the summarized results for classes of cues across the three experiments. The inconsistencies in cue emergence and general directions of classes of cues imply that there are a number of moderating factors that govern what language is in use. Many of the patterns are at odds with previous findings collected under less interactive circumstances. They highlight the critical need for more testing and careful determination of the factors that define a particular situation (e.g., planned or spontaneous discourse, formal or informal interaction, narrative about events versus opinions or feeling states, high or low jeopardy for deceit being detected). With more planning time possible, deceivers could conjure up more details to appear more believable, although the extra information could be superfluous rather than useful. The fact that quantity cues are unreliable can explain the low accuracy in human judgment of deception because humans tend to rely heavily on these convenient (but unreliable) quantity cues.

Table 14. Summary of Deception Effects on Linguistic Categories.

Categories	Desert Survival	Mock Theft	Deceptive Interviews
Quantity	Longer	Shorter	Longer
Complexity	Less	Less	More
Uncertainty	Less certain	More certain	Less certain

Personalization/ Nonimmediacy	Impersonal	Personal/ immediate	
Diversity	Less	More	Less
Specificity	Mixed	Specific	Vague
Affect	Mixed	Mixed	Mixed

Table 15 lists the specific linguistic cues that were significant in between-subject tests. Most linguistic cues were significant at least once, and 17 out of 21 cues that were measured in all investigations turned out to be significant at least once, implying that the defined cues are potentially good predictors.

Table 15. Promising Cues in Each Category.

Linguistic Class	Cues	Linguistic Class	Cues
Quantity	Words	**Diversity**	Lexical diversity
	Content words		Content word diversity
	Verbs		Redundancy
	Wordy	**Informality**	Misspellings, typos, etc.
Complexity	Average word length	**Specificity**	Temporal details
	Average sentence length		Spatial terms
	Pausality (punctuation)		Sensory terms
Uncertainty	Modifiers		Total details
	Modal verbs		Comparatives
	Ellipsis		Superlatives
Personalization/ Nonimmediacy	1st person singular pronouns	**Affect**	Positive pleasantness
	1st person plural pronouns		Negative pleasantness
	2nd person pronouns		Positive imagery
	3rd person pronouns		Negative imagery
	Other references		Affect
	Possessives		Emotiveness
	Passive voice		

Although the current results paint a very complex picture, the problem is not an intractable one. Moreover, the number of linguistic features that emerged in one or more analyses underscores the promise of utilizing language to assess a person's veracity.

2. Field Studies

The two field studies collected data from real-world scenarios and therefore provide an excellent test bed for assessing whether laboratory-generated results generalize to real-world applications.

a) Enron Field Study

In this analysis, ingroup/outgroup status served as a proxy for deception. A J48 decision tree with ten-fold cross-validation correctly classified 48 out of 58 e-mail messages as ingroup or outgroup, attaining 83% accuracy. The classification results are summarized in Table 16.

Table 16. Classification Accuracy for Enron Email Message Corpus.

	Classified as	
Actual Group	**Ingroup**	**Outgroup**
Ingroup	83%	17%
Outgroup	17%	83%

Figure 15 shows the decision tree produced by the J48 algorithm to classify the messages. The decision tree itself required only 5 of the 39 possible cues to perform the classification. They were:

1. *Extreme Positive Pleasantness* (2 standard deviations)
2. *Average Sentence Length*
3. *Verb Quantity*
4. *2nd Person (« you ») References*
5. *Passive Verb Ratio*

```
Pleasantness <= 0.007673: false (19.0/1.0)
Pleasantness > 0.007673
|  Average_Sentence_Length <= 34.5
|  |  Verb_Quantity <= 8: false (3.0)
|  |  Verb_Quantity > 8
|  |  |  You_References <= 0.024155: true (20.0)
|  |  |  You_References > 0.024155
|  |  |  |  passive_verb_ratio <= 0: true (9.0/1.0)
|  |  |  |  passive_verb_ratio > 0: false (2.0)
|  Average_Sentence_Length > 34.5: false (5.0)
```

Figure 15. Decision Tree Output for Enron Email Corpus.

Storyboarding through the decision tree output reveals that a Pleasantness score less than or equal to 0.007673 is first used to identify outgroup email messages (as shown by 'false' in the output). If Pleasantness is greater than 0.007673 *and* the Average Sentence Length is less than or equal to 34.5 words *and* the number of Verbs in the message is less than or equal to eight, the message is classified as outgroup. Similar logic applies throughout; the decision tree reads like a nested if-then-else statement.

b) Security Police Statements

This corpus, for which ground truth was known, was subjected to both A99A and LIWC to calculate the relevant values for the desired variables and compare the performance of each program. For each program, these results were then separately analyzed to determine which variables could be used to distinguish truthful and deceptive statements. For the variables calculated using A99A, significant differences were found for all variables except the ratio of affect-laden terms and modal verbs. Similarly, for the variables calculated using LIWC,

significant differences were found for all variables except the affect ratio and modal verbs (see Table 17). These results show that the programs returned the same variables as discriminators. The direction of the differences between variables was as expected and consistent with previous results for all variables except for the LIWC sensory and perceptual processes variable. While the corresponding sensory ratio variable was significantly greater in truthful than deceptive statements when calculated by A99A (as expected), LIWC found it to be significantly higher for deceptive than truthful statements.

Table 17. Key Predictors from Police Statements Corpus.

A99A Variables	LIWC Variables
Word Count*	Word Count*
Affect Ratio	Affect
Sensory Ratio*	Sensory and Perceptual processes*
Lexical Diversity*	Unique Words*
Non-self References*	Other References*
Second Person Pronouns*	Total Second Person*
Other References*	Total Third Person*
Group Pronouns*	1st Person Plural*
Spatial terms*	Spatial terms*
Modal Verbs	Modal Verbs

*Significant mean difference between truthful and deceptive statements at 0.05 level.

These results lend credibility to the use of these tools in deception detection and other text analysis tasks. The similar results achieved with each tool suggest that cues which have been appropriately defined can be automated to assist investigators. These results might also allow us to draw limited comparisons between different studies using different tools when the variables are defined similarly for both tools. For most of the variables analyzed in this study, the definitions of the variables are relatively straightforward. For example, the list of third person pronouns is fairly well-defined. The results are mixed for less obvious variables such as affect and spatial terms.

Despite these promising findings on most variables, the tools failed to detect significant differences on variables previously suggested to be useful as predictors of deception in text, such as affect and modal verbs (L. Zhou, Burgoon, Nunamaker, & Twitchell, 2004). It may be that the type of statement being analyzed reduced the presence of affective terms such as "good" or "bad" or produced the same amount in both truthful and deceptive statements. Alternatively, the lack of significance in either program may have been the result of looking at this variable at an aggregate level. Some previous studies have separated this variable into more than one variable (Hancock & Dunham, 2001; Zhou, Burgoon, Nunamaker, & Twitchell, 2004), Given that modal verbs have shown to be effective discriminators in other studies, the nonsignificant results on this indicator, like affect, are an argument favoring a multi-indicator model in which only some of the potential indicators are likely to be present in a given statement. Also not to be discounted as an explanation for the nonsignificant findings on these cues is sample size. Only 60 statements were used in this study, which may not be adequate to find significant differences on all cues.

To assess classification accuracy, four different classification tools were compared for their ability to accurately classify deception and truth. Given the somewhat low sample size of 82, cross-validation was accomplished by randomly drawing a second set of 41 truthful statements from the large pool of available truthful statements. Results shown in Table 18 reveal excellent detection accuracies that far exceed those achieved by human judges. The differences among classifiers, especially in the second (cross-validation) sample, are negligible. In the first sample, discriminant analysis and logistic regression (the two statistical methods) performed better in detecting truth whereas decision trees and neural networks (the two machine learning methods) performed better in detecting deception. However, these differences largely disappeared in the second sample.

Table 18. Best Classification Accuracies for Four Classifiers on Two Samples.

	Discriminant Analysis	Logistic Regression	Decision Tree	Neural Network
	Sample 1: N=82			
Truth	.90	.84	.78	.73
Deception	.59	.76	.83	.90
Overall	.74	.80	.80	.82
	Sample 2: N=82			
Truth	.83	.73	.83	.68
Deception	.66	.68	.66	.66
Overall	.74	.71	.74	.67

c) Field Studies Summary

Both the Enron and security police field studies produced accuracies greater than chance (50%). This shows that the ability of machines to detect deception from text is greater than the average human's. Both studies were limited to primarily English speakers within very specific niches. Cultural impacts on text-based deception detection provide potential for further studies. It is likely that with different cultures, cues will perform differently. These field studies also show the potential applications of such a tool. Detectives could use a tool to better understand a suspect; auditors could use a similar tool to identify when companies or individuals are being deceptive.

B. Tests for Vocalic Indicators

1. Mock Theft Experiment

Audio indicators of deception were analyzed in two laboratory experiments. In the mock theft experiment, 33 interviews from the audio-only condition were analyzed. Of these, 20 of the interviewees belonged to the guilty condition and were therefore deceptive during the interview; the remaining 13 belonged to the innocent condition and therefore presumably were truthful. The audio tapes were subjected to both behavioral observation by trained human coders and machine-automated analysis. Human coders rated responses during the work block of questions (during which "thieves" were also asked to lie) and the theft block of questions. To achieve greater

parsimony and interpretability, an exploratory factor analysis with varimax rotation was conducted to reduce the various measures to five composite dimensions. Table 19 lists the dimensions, the specific cues associated with each dimension, the relationship with deception, and whether behaviors are also influenced by modality and motivation.

Table 19. Results for Analysis of Human Behavioral Observation of Audio Features.

Dimension	Indicators	Deception Effects	Modality Effects	Motivation Effects
Quantity	Talk time duration	T > D	Audio > FtF	With Lo, D = T; With Hi, T > D
	Talk time percent	--	--	Affects Lo/Truth
	Frequency of nonfluencies	D > T (work block only)	FtF > audio	--
Turn-taking	Partner turn length	D > T in audio work block T > D in FtF work block	Interacts w/ deception	Hi > Lo
	Switch pause length	--	Interacts w/ motivation	Interacts w/ modality
	# of turns	T > D in audio theft block D > T in FtF theft block	FtF > audio, esp w/ theft block	Hi > Lo
Utterance Length	Mean turn length	T > D	Audio > FtF	Hi > Lo
	# of filled pauses	--	--	Hi > Lo
	Latency of filled pauses (inverse relationship)	--	--	--
Partner Quantity	Other-talk duration	--	--	--
	Own talk latency	--	FtF > audio in Theft block	--
Nonfluencies	Nonfluency Rate	T > D (work block only)	FtF > audio	Hi/Deception < all else

The automated analysis was confined to the theft block of questions and was further segmented into the eight questions that comprised the theft portion of the interview. The questions are listed in Table 20.

Table 20. Mock Theft Questions Evaluated in Vocal Analysis.

ID	Question	Type
Q1	If you had anything to do with taking the wallet you should tell me now.	Short Answer
Q2	Do you know where the wallet is now?	Yes / No
Q3	Walk me through what happened in class.	Narrative
Q4	Did you notice anyone that you suspect?	Yes / No
Q5	Were you near the chalkboard at anytime during class?	Short Answer
Q6	What should happen to the person who stole the wallet?	Short Answer
Q7	Would you ever give that person a break?	Short Answer
Q8	How do you think the investigation will turn out in regards to you?	Short Answer

A stepwise, forward Wald logistic regression was performed for each question in the analysis. Initially, the p-values were set at .05 for entry and .1 for removal of a variable. However, due to the small sample size, many of the models required a more relaxed p-in and p-out before any variables entered the model. Table 21 displays the results for models that were created for each of the 8 questions. All models were significant, many at the $p < .001$ level, and many accounted for substantial variance. While the results appear promising, caution is urged in making generalized interpretations because the size of the data set is small and the results are not cross-validated.

Table 21. Binary Logistic Regression Classification Results for Audio Features in Mock Theft

Question	Overall Correct Classification	Actual Condition	Percentage Predicted		p-value	Cox & Snell R- Square
			Truthful	Deceptive		
Q1	100% [3]	Truthful Deceptive	100% 0%	0% 100%	< .001	.738
Q2	100% [3]	Truthful Deceptive	100% 0%	0% 100%	< .001	. 738
Q3	100% [3]	Truthful Deceptive	100% 0%	0% 100%	< .001	.720
Q4	74.2% [1]	Truthful Deceptive	61% 17%	38% 83%	.004	.239
Q5	100% [2]	Truthful Deceptive	100% 0%	0% 100%	< .001	.743
Q6	75% [3]	Truthful Deceptive	69% 21%	31% 79%	.003	.301
Q7	77.3% [2]	Truthful Deceptive	87% 29%	13% 71%	.003	.325
Q8	100% [3]	Truthful Deceptive	100% 0%	0% 100%	< .001	.743

[1] p-in = .05; p-out = .10 [2] p-in = .1; p-out = .2 [3] p-in = .2; p-out = .3

Cues from all categories can be automatically extracted. However, the difficulty of extraction varies greatly. For example, cues extracted for the frequency and intensity categories are fairly straightforward as they can be directly extracted from the audio signal without any additional contextual information. In contrast, the unfilled pauses cue not only requires identifying when there is silence but it also when it is the subject's turn. Each cue identified in the original taxonomy can also be categorized by the ease of automatic extraction. Figure 16 shows a rough categorization of the cues into one of three categories: easier to extract, medium difficulty to extract, and harder to extract. The categorizations are "rough" because:

1. There is more than one way to measure a cue. Some are more representative than others. However, some measures can be extracted more easily than others.
2. Many of the cues depend on one or more lower-level cues that must be extracted before they can be calculated. The ease of extraction of these cues depends on the ease of extraction of the earlier cues.
3. Some cues require additional contextual information. For example, vocal tension requires a known truthful baseline. While it is possible to create simpler cues that may hint at

tension (e.g. low-pass filter), automatic extraction of such a cue requires additional input besides the audio signal.

4. Some of the cues, such as pleasantness, are largely perceptual. Aspects of the cue might be easy to extract, but a universal measure likely requires the merging of many features.

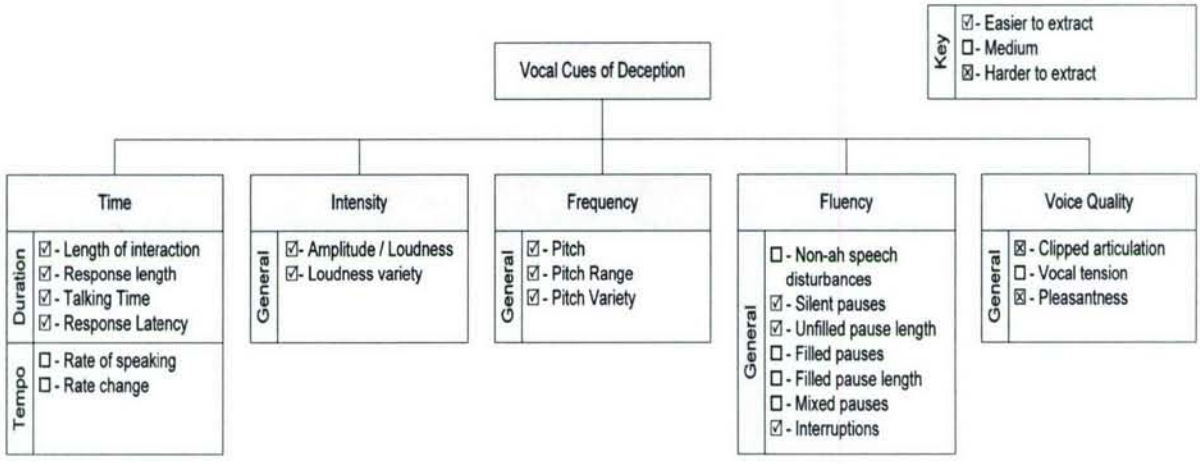


Figure 16. Ease of automatic extraction for each feature in taxonomy

Table 22 summarizes the combination of cues that are predictive for each question and the directionality of those cues. Visual inspection of this table plainly reveals that, even within one dataset, no model is consistent across all questions or question types, though some features appear in multiple models. However, directionality of cues is consistent across questions.

The directionality of most of the cues agrees with findings in the literature. For example, fundamental frequency (pitch) and fundamental frequency variety increase when deception is present (Anolli & Ciceri, 1997; Ekman, Friesen, & Scherer, 1976; Rockwell, Buller, & Burgoon, 1997). Deceivers tend to have increased response latency (Rockwell, Buller, & Burgoon, 1997). Additionally, fluency for deceivers decreases, reflected as a general increase in feature values in our fluency category (Bond, Kahler, & Paolicelli, 1985; DePaulo et al., 2003; Rockwellet al., 1997). Most models combine cues from multiple categories in the taxonomy. It is likely that cues within the same category account for an overlapping amount of variance. Thus, a strategy that incorporates multiple features from multiple categories will likely account for a higher amount of variance and might also more accurately classify deception.

Table 22. Summarization of features and directionality (including question type).

	Time					Intensity					Frequency			Fluency						Voice Quality		
	Interviewer speech average deviation	Interviewer turn duration average	Response latency duration average	Response latency first duration	Subject turns	Energy average	Energy average deviation	Energy minimum	Gain average	Gain range	Fundamental frequency average	Fundamental frequency avg deviation	Fundamental frequency maximum	Non-silence average	Non-silence average deviation	Overlap (interruptions) average	Silence average	Silence average deviation	Unfilled pause length	Adjusted low-pass average deviation	Low-pass average deviation	Low-pass minimum
Q1						↓	↓					↑	↑			↑					↓	
Q2						↓			↑			↑		↓				↑		↑		↓
Q3		↑						↓			↑						↑					
Q4															↑							
Q5			↑	↓									↑		↑				↑			
Q6	↑														↑							
Q7					↑																	
Q8	↑				↑					↑					↑				↑			↓

Question types are color coded as follows:

Yes/No Questions | Short Answer | Narrative

2. Deceptive Interviews

Due to the deceptive interviews being recorded in analog rather than digital format, they were not amenable to automated analysis. Therefore, only behavioral observation was conducted. To reduce some of the variability associated with individual questions, results were analyzed in blocks of questions. Interviewers responded with four blocks of three questions each, beginning either with truth (TTTDDDTTTDDD) or deception (DDDTTTDDDTTT).

Results indicated that relative to truthful responses, deceptive responses were characterized by:

1. *Shorter turn length*
2. *Shorter overall duration during the third block*
3. *More frequent total nonfluencies and "other" nonfluencies*

The total nonfluencies category includes silent pauses, filled pauses, and other types of nonfluencies such as garbled and intrusive sounds and stuttering). Figure 17 shows the pattern for the "other" nonfluencies.

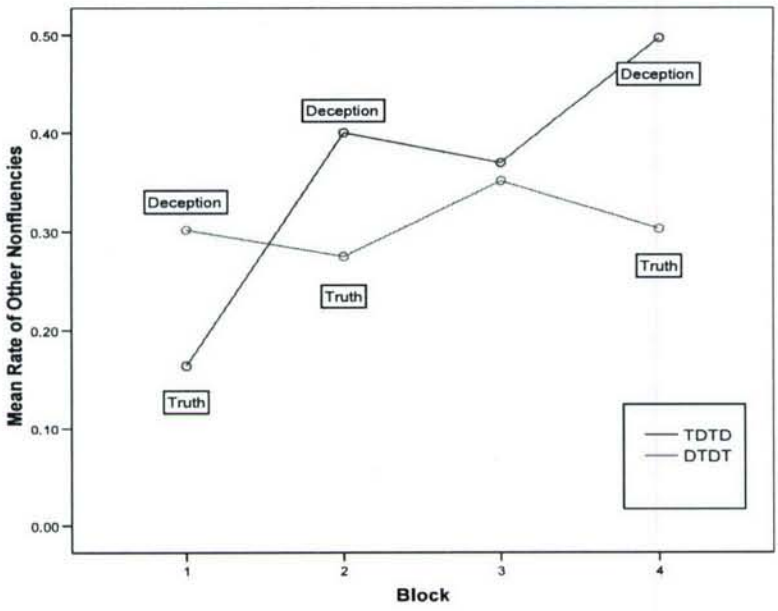


Figure 17. Effects of Deception, Truth/Deception Order and Time on Nonfluencies.

Discriminant analysis was conducted within each block with 15 vocal features as the predictors. Table 23 shows for each significant model the number and specific variables that were significant predictors, the classification accuracies achieved, and the Wilk's Λ for the overall model. It is evident that predictive accuracy was greatest during the second block of questions and that by later in the interview, detection accuracies and number predictive indicators declined. It is also the case that detection accuracy was greater for those interviewees who began with deception rather than truth. Put differently, starting with truthful responding made it easier for interviewees to evade detection when they shifted to deceptive responding.

Table 23. Classification of Truth and Deception from Audio Behavioral Observation

Block	Predictors	Deception Classification Accuracy*	Truth Classification Accuracy*	Overall Classification Accuracy	Wilks' Lambda
1	Partner Turn Length Silent Pauses Frequency Total Nonfluency Rate	.71	.50	.61	.807
2	Turn-switch Pause Length Silent Pauses Frequency Talk duration Response Latency Partner Turn Length	.57	.74	.66	.735
3	Other Nonfluencies Latency	.71	.47	.59	.847
4	Turn-switch Pause Frequency	.40	.84	.63	.887

*cross-validated

The combined results from the audio analyses reveal a large number of features that discriminate truthful from deceptive responses, although the timing of their measurement influences how successful they are at classifying truth versus deception. These features also play a role in the multimodal analysis described in Section VIII-D.

C. Tests for Kinesic Indicators

1. Mock Theft

The face-to-face interactions from the mock theft experiment where used in the current study. There were a total of 42 possible face-to-face interactions that could be included in the study; four were not used because of technical problems with the video work or because the participant did not follow instructions. Each interaction was composed of a number of question-answer exchanges. Two types of analyses were conducted, one with manually coded variables using teams of trained human coders and one with automated blob analysis.

For the human behavioral observation, the variables that were measured were:

1. *adaptor gestures* (frequency count, duration, rate per turn)—gestures that alleviate psychological or physiological discomfort, such as face-touching and scratching
2. *illustrator gestures* (frequency count, duration, percent of turn, rate per turn)--gestures that accompany speech and clarify, complement, or modify verbal statements; also includes interactional gestures that regulate turn-taking
3. *head nods* (frequency count, duration, percent of turn)—all movements of the head during speaking turns and during listening (sometimes divided into speaking and listening roles)
4. *gestural stillness* (duration, percent of turn)—periods of no adaptor or illustrator gesturing
5. *head stillness* (duration, percent of turn)—time periods where no head movement is observed
6. *postural shifts* (frequency count)—number of times a person changes his/her posture, such as crossing legs, shifting in the seat
7. *total movement*—sum of all the gesture, head movement, and postural movement counts

Virtually all of these measures were affected by experimentally induced motivation (see Section E.5.). The frequency of adaptor gestures and postural shifts was highest during the early (truthful) phase of the interview, and deceivers displayed more of them, but by the time the theft portion of the interview commenced, truthtellers actually displayed more (when controlling for the amount of time they talked). However, deceivers did display longer adaptors (partly as a function of the amount of talk time). Postural shifts, head movements, and gestural stillness were also moderated by motivation (i.e., the influence of deception depended on the level of motivation). Finally, illustrator gestures showed a trend toward truthtellers exhibiting more than deceivers during the theft questions. The human-rated behaviors are also discussed in more detail in Section D.1. regarding multimodal analyses.

In the automated analysis, only the theft narrative was included. Of the 38 participants, 16 were truthful and 22 were deceptive.

All relevant variables were included in a stepwise discriminant analysis with p-in set at .10 p-out set at .20. The final model included 16 features:

1. Average difference of the head blob angles between frames
2. Average distance the head blob moved between frames
3. Average x position of the left hand blob
4. Average difference between right hand blob angles between frames
5. Average distance between right hand blob and head blob
6. Average y position of the center of the triangle connecting the three blobs
7. Average number of frames the left hand blob was located in quadrant 2
8. Average amount of down-right movement in the right hand blob
9. Variance of the distance the head blob moved between frames squared
10. Variance of the left hand blob width
11. Variance of the distance the left hand blob moved between frames squared
12. Variance of the right hand blob x position
13. Variance of the right hand blob height
14. Variance of the number of frames the right hand blob was in quadrant 1
15. Variance in head blob angular movement
16. Variance of right hand up movement

The results of the discriminant analysis are shown in Table 24.

Table 24. Classification of Truth and Deception from Kinesic Machine-Coded Indicators

		Predicted	
	Actual	**Truthful**	**Deceptive**
Original Model	Truthful	100%	0%
	Deceptive	0%	100%
Cross-Validated	Truthful	93.8%	6.3%
	Deceptive	0%	100%

The deceptive and truthful participants were classified with an accuracy rate of 100 percent. Cross validating by withholding one subject from the analysis and using that subject for testing caused the accuracy rate to fall to 97.4 percent. The model was significant.

A logistic regression analysis was also conducted. To achieve a cross-validation similar to discriminant analysis, nine randomly selected interviews were used to create a training set model that was then applied to the remaining cases as the test set. The five features in the training model--average distance the head blob moves between frames, average distance between the right and left hand blobs, average number of frames the right hand blob was located in quadrant 3, variance of the distance the head blob moves between frames, variance of the distance the right hand blob moves between frames—correctly classified deceptive and truthful participants

at an accuracy rate of 66.7 percent for each group.

Even in this test with limited sample size, automated extraction and analysis of nonverbal features performs better than typical human judgment. Further, these results demonstrate that automatically extracting nonverbal features for the purpose of deception detection may be feasible. While this experiment is a small initial step bounded by sample size limitations, it does give a glimpse of the promise of blob analysis in analyzing nonverbal behavior.

2. Field Observations

Most of the data we have analyzed has been captured in an experimental setting, which allows us to control many factors. However, we developed our kinesics-based approach with the expectation of applying it in a field environment. Toward that end of testing the feasibility and robustness of our approach, we have begun collecting data at three areas within the Dennis DeConcini Port of Entry, Nogales, Arizona with the approval and cooperation of Customs and Border Protection. Data were collected in three locations—1) at the pedestrian crossing where people queuing in lines or standing in common areas and then approach and interact with a CBP officer, 2) at the permit counter which is a standing interaction, and 3) in the expedited removal room there a seated interview occurs.

Pedestrian Border Crossing – At the pedestrian crossing where individuals queue up to cross into the United States, we are collecting video images from an overhead camera. Figure displays a video frame from the pedestrian lane. The monitoring of people queuing in lines and standing in common areas does not elicit deception per se, as there is no direct interaction between security officials and the subject where overt deception can take place. However, systems can focus on identifying arousal cues that may indicate concealment and avoidance that accompanies deception. For instance, a person in a line with hostile intent may behave differently than others around him. Such a person might be agitated, or perhaps more likely, over-controlled. Tracking the movements of the head and the hands has the potential to identify anomalous behavior even without interaction with humans. When the individual presents him or herself to the officer a certain amount of interaction occurs. These video frames often contain multiple individuals and a skewed angle. Because of the number of individuals who pass through the pedestrian lanes, we are currently not collecting any metrics from the officers. Rather, we will retroactively tag those who were pulled aside for additional questioning. We still need to develop algorithms to track individual movements when multiple individuals are queued in a line.

Figure 18 (a) Pedestrian lane video frame.

(b) Extended stay permit video frame

Permit Window – We are collecting audio and video data at the walk-up counter where individuals apply for permits to travel further into the United States or to stay for an extended period of time. **Error! Reference source not found.**displays a video frame taken from data collected at the extended stay permit. This application process typically takes just a few minutes but often it involves detailed communication about where one is going and why. Officers will complete a simple rating instrument when they are finished with an interaction. This instrument captures how suspicious the officer was of the candidate and also how deceptive the candidate was and any brief notes. During analysis, the behavior of an individual (vocalic and kinesic) will be correlated to the officer's rating of deception and suspicion. In this scenario, the movement-based approach is useful in that it can help to identify heightened arousal on the part of the subject.

Expedited Removal (ER) Room – If the suspicions of security personnel are raised significantly, the subject can be diverted to a structured interview with more controlled conditions. The expedited removal room is an area where individuals who have attempted to enter illegally are processed for "removal" back to Mexico. A third camera and microphone capture the behavior and responses of an individual that has been retained for additional questioning in a seated interview. These interviews last anywhere from 20 minutes to several hours. The ER room has 3 interview stations. After the interview, a researcher consents willing participants to allow us to analyze the captured interview. During the consent process the officer fills out a rating instrument indicating how suspicious they were of the candidate at the beginning, middle, and end of the interview and also how deceptive the candidate was being at the beginning, middle, and end of the interview. Any observations relating to the interviewee are noted as well. **Error! Reference source not found.**displays a snapshot of the rating instrument used during the expedited removal interviews. During analysis, the behavior of the interviewee will be correlated to the metrics collected.

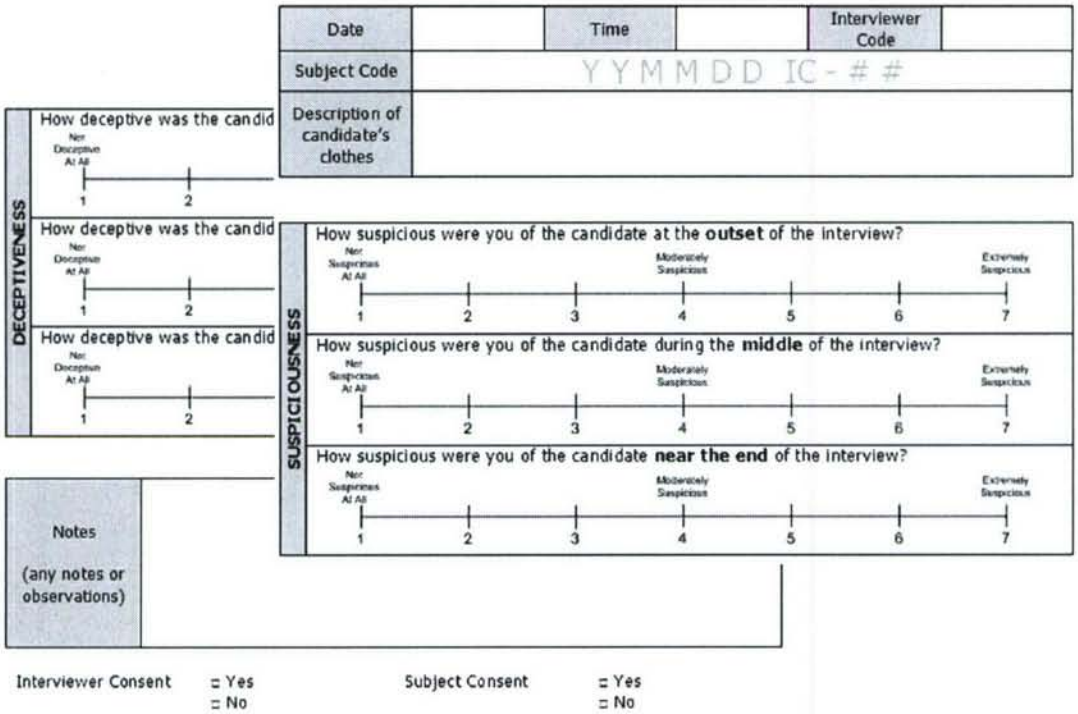


Figure 19. Sample rating instrument used for the permit counter and expedited removal interviews.

3. Macro-Level Analyses of Behavior Cues Extracted from Video

The preceding analyses have all focused on more micro-level and objectively measured cues. This section explores the possibility of using those cues to predict human-interpretable judgments of involvement, dominance, tenseness, and arousal. The approach is based on a Brunswikian lens model that involves distal cues, proximal percepts and a final attribution. A Brunswikian lens model is extremely useful for identifying configurations of micro-level deception cues that predict mid-level percepts which in turn predict attributions. Figure 22 displays an operationalized view of the model using communication dimensions as proximal percepts that can be combined to arrive at an attribution of an individual's level of honesty (on a scale of 0 to 10; 0 being completely deceptive and 10 being completely honest)

In our data sets, humans participate in deception which is represented by the state characteristic in the lens model. The distal indicators are automated features extracted through kinesics analysis (described below). The proximal percepts are communication dimensions (e.g., involvement, dominance, tenseness, arousal) derived from judgments made by third-party observers. The final attribution is a prediction of a self-reported honesty score using proximal percepts as predictors. This attribution is validated through comparison with the original characteristic. In the case of deception detection, both the characteristic and attribution can be viewed as an inverse relationship with the level of honesty. Sample indicators are shown for each of the major components of the model in Figure 22.

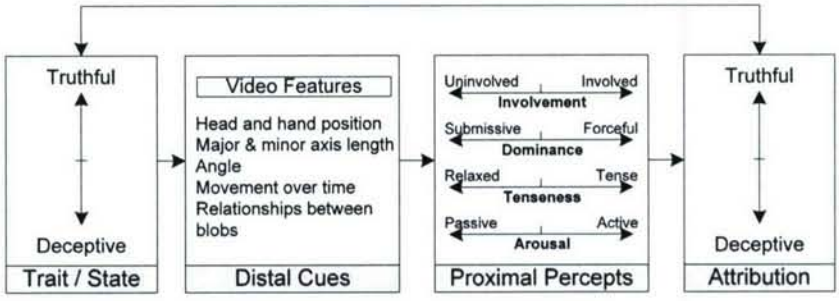


Figure 22. Brunswikian lens model applied to deception detection

When applying the Brunswikian lens model to the problem of deception detection, expected relationships between the components of the lens model need to be specified. Fortunately, past research has provided, to a large extent, empirically-based links between proximal percepts and attributions. In our model, proximal percepts are operationalized as human perceptions of dominance, tenseness, arousal, and involvement. These broad perceptions of human communication encompass a great deal under a single assessment. For example, involvement can be divided into subcomponents of immediacy, altercentrism, expressiveness, conversational management, and social anxiety (Coker & Burgoon, 1987). While more granular measures of the subcomponents can be productively and selectively studied via a Brunswikian lens model, they are conglomerated in the current study.

With the selection of the proximal percepts established, associated distal cues can be determined. There are numerous possible cues which may account for perceived levels of the proximal percepts. The search for relevant distal cues was bounded by existing research in deception detection. Proximal percepts, which were predicted using automatically-extracted distal cues, are used to predict an honesty score which can also be thought of as an estimated level of deception. A sample of proximal percept levels and distal cues that may be associated with deception is shown in Table 25.

Table 25. Sample proximal percepts and distal cues associated with deceptiveness.

Proximal Percepts	Observed Levels	Distal Cues
Dominance	Lowered	Limited hand movement over time
Tenseness	Elevated	Minor hand movements which are close together Rigid head movement
Arousal	Mixed	Frequent hand-to-face gesturing and hand-to-hand movements
Involvement	Lowered	Limited gestures away from the body

We have made initial steps in validating our approach of deception detection via a Brunswikian lens model. Our automatic kinesics analysis is capable of extracting relevant distal cues that can be used to predict perceptual judgments such as involvement (R^2 = .276), tenseness (R^2 =.121), and arousal (R^2 = .447). Additionally, it was shown that the predicted proximal percepts could be used to determine an attribution; in this case, an individual's level of honesty (R^2 =.096). Therefore, the predicted percepts significantly predict self-reported honesty.

By building tools that can better approximate human perceptions of involvement, tenseness, and arousal (and other perceptions), this research lays a foundation to provide answers to such real-world questions as: what is needed for a machine to interact sensibly with a human? What indicators are the prototypical features necessary to simulate real communication? Are there telltale—and automatically detectable—signals from which a machine can infer a human's current internal state? To the extent that scenes have behavioral routines associated with them (e.g., leaders showing dominance in a meeting), such behavioral recognition can ultimately aid scene recognition. More generally, the multifunctional, multimodal, and molar approach represented here offers a more ecologically valid model of how micro-level behaviors create more general perceptions that drive attributions of meaning.

D. Tests for Multimodal and Macro Patterns

1. Multimodal Analyses from Combined Linguistic, Audio and Kinesic Cues

The mock theft experiment afforded the opportunity to determine if stronger predictive models could be developed by combining features from multiple communication modalities. The only condition in which a full multimodal analysis could be conducted was the face-to-face condition. All the linguistic, vocalic and kinesic predictors were entered into that model. Audio and verbal features could be tested separately within the audio condition or with the combined audio and face-to-face condition. Three separate blocks of questions were analyzed. The first, truthful baseline questions, asked about educational experiences and ostensibly should have produced similar behavioral patterns from both innocent and guilty parties inasmuch as the guilty parties were instructed only to lie on the work and theft questions. However, results indicate that even during this ostensible baseline period, discrimination was already occurring (see also the results for the Deception Index). The second block of questions related to work experiences. "Thieves" were asked to lie about their work experiences so as to provide another type of questioning that might discriminate innocent (truthful) from guilty (deceptive) interviewees. The third block of questions surrounded the theft itself.

Table 26 summarizes the detection accuracies and predictors in the stepwise discriminant analyses that were conducted within each of the blocks of questions. In the case of the FtF model during the work questions, the model failed to achieve statistical significance using conventional significance levels, so relaxed *p*-values (.10 for entry into the model, .20) for removal were used. The same relaxed criteria were used for the audio-only condition during the baseline questions.

Table 26. Multimodal Classification of Truth and Deception from Mock Theft.

Modality	Audio Only			FtF Only			Audio and FtF Combined		
Question Block	Base-line	Work	Theft	Base-line	Work	Theft	Base-line	Work	Theft
N	45	42	44	56	51	56	100	96	96
Detection Accuracy									
Truth	0.58	0.50	0.63	0.75	0.84	0.94	0.73	0.48	0.63
Truth Cross Validated	0.58	0.50	0.63	0.72	0.84	0.88	0.73	0.48	0.61

Deception	0.50	0.71	0.82	0.84	0.73	0.77	0.48	0.72	0.67
Deception Cross-Valid	0.50	0.71	0.79	0.69	0.69	0.69	0.48	0.72	0.65
Predictors									
Truthful Baseline	1. "other" references ($p < .10$)			1. turn-switch pause length 2. filled pause rate 3. sensory terms ratio 4. interviewer % of talk time			1. filled pause rate		
Work Questions	1. 3rd person pronouns 2. 1st person singular pronouns			1. negative activation terms (1sd) 2. filled pause rate 3. duration of turn-switch pauses 4. positive imagery terms (1sd) 5. emotiveness 6. 1st person plural pronouns 7. average word length 8. negative pleasantness (1sd)			1. average sentence length 2. complex/compound sentences 3. redundancy		
Theft Questions	1. affect terms ratio 2. average word length 3. turn frequency			1. negative pleasantness (1sd) 2. filled pause ratio 3. passive verb ratio 4. emotiveness 5. spatial far terms ratio 6. affect terms ratio 7. turn length 8. filled pause frequency			1. negative pleasantness (1sd) 2. "other references 3. turn-switch pause length		

Results indicate, first, that detection accuracy was best in the FtF-only condition. It was extremely high (94%) for truthful classifications during both the work and theft questions and was also strong for deception detection for all three question blocks. The strong showing during the baseline is noteworthy in revealing that future deceptive intent was already influencing behavior at the outset of the interview. The fact that classification accuracy worsened when the FtF and audio conditions were combined is probably due systematic discourse differences in each condition. As reported earlier, deception effects were often moderated by modality. Additionally, motivation moderated results in different ways in each condition and so might account for the lack of straightforward deception effects.

The predictor variable set is also most extensive in the FtF condition. Interestingly, it includes a mix of linguistic and vocalic but no kinesic variables. This suggests that language features and voice alone could effectively distinguish truthtellers from deceivers during a FtF interaction. However, quite a few other variables produced significant or near-significant mean differences between the truth and deception conditions, including some kinesic measures. During the baseline, positive pleasantness terms (1 s.d. or higher), filled pause frequency, adaptor gesture rate, amount of time with no gesturing, spatial far terms, pleasantness, percentage of postural

shifts per turn, total nonfluencies, and total amount of movement all showed effects. Within the theft block of questions, lexical diversity, negative activation terms (1 s.d. or greater), duration of turn-switch pausing, duration of illustrator gestures, length of illustrator gestures, and duration of time head is still all produced effects. These findings highlight that many variables could serve as effective predictors and that those that failed to enter the model did so because they were correlated with other measures that had entered the model already.

As regards the specific predictors, virtually all categories of behavior emerge in one of more models. Linguistically and vocalically, there are indicators representing quantity, diversity (lexical and syntactic), specificity, complexity (lexical and syntactic), personalization, immediacy, affect, turn-taking, interviewer behavior, and nonfluencies. Kinesically, all body regions—head, hands, and posture—were potentially implicated. (Many of these variables were also highly sensitive to motivation and modality effects and so would not have shown main effects for deception.) Virtually all the verbal and nonverbal behaviors that were measured, then, have shown promise in the mock theft analyses.

2. Speech Acts for Deception Detection

With the increasing use of computer-mediated communication (CMC) tools such as chat, instant messaging, and e-mail, persistent conversations are becoming more common and are increasingly used as methods for transmitting deception. However, automated tools for studying human behavior in persistent conversations are rare. Even more rare are automated tools for aiding deception detection in these conversations. This section describes how speech act profiling can be used as an automated tool to uncover uncertainty in deceptive conversations. Uncertainty could be added to the set of cues applicable to synchronous CMC already used in message feature mining to increase the accuracy of deception classification models. Speech act profiling can also be used as an aid to deception detection in online CMC. Dominance and uncertainty are two correlates of deception that can be discovered using speech act profiling as shown in the two studies below. The first study uses the large SwitchBoard corpus as a training set while the second uses the smaller, more relevant StrikeCom corpus.

a) Detecting Deception Using Speech Act Profiling Trained on the Switchboard Corpus

These examples come from the StrikeCom corpus. Figure 23 is a speech act profile created from all of the utterances from a single game. In this particular profile, Space2, the deceiver, is behaving submissively. The submission is evident from the slight lack of Statements (sd) and the abundance of Appreciation (ba) and Agree/Accepts (aa). The behavior would also be apparent from a transcript, which would show Space2 several times only saying "ok" while the others are carrying on the conversation. This type of behavior could be an indication of freeloading or cognitive laziness that deceivers sometimes display.

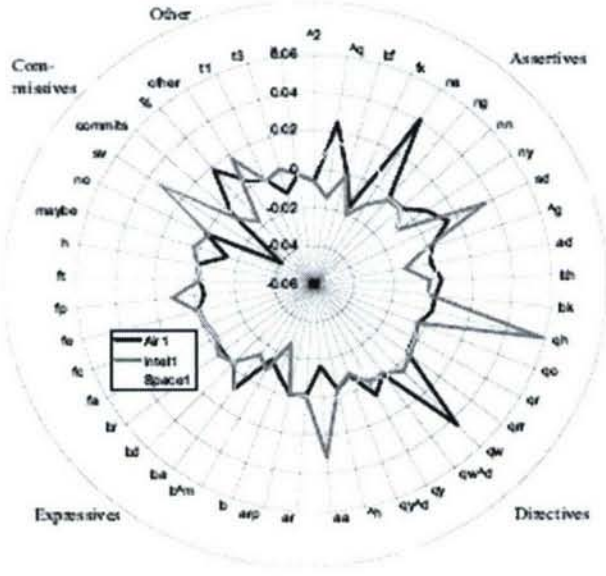


Figure 23. Sample speech act profile from the StrikeCom corpus showing submissive behavior by the deceiver

In another single group interaction depicted in Figure 24, the profile indicates that the participant in the Space1 role has taken a submissive stance compared to the other participants, Air1 and Intel1. Space1 used fewer statements and greater proportion of backchannels and agreements than the other two. In addition to submission, the profile indicates uncertainty in Space1's language. A running transcript reveals that early in the game, Space1 hedges the comment "i got a strike on c2" with the comment "but it says that it can be wrong...". Later Space1 qualifies his advocacy of grid space e3 with "i have a feeling". In reality there was no target at e3, and Space 1 was likely attempting deceive the others as instructed. In the Depaulo et al. (2003) meta-analysis of deception vocal and verbal impressions of uncertainty by a listener were significantly correlated with deception (d = .30). That is, when deception is present, the receiver of the deceptive message often notices uncertainty in the speaker's voice or words. Since the voice channel isn't available in CMC, any uncertainty would have to be transmitted and detected using only the words.

This and the preceding example are useful for envisioning how an investigator might use speech act profiling. However, the probabilities produced by speech act profiling can also be used for statistical comparison. These probabilities represent the probable proportion of utterances that were of a give speech act. These probable proportions can be compared across experimental treatments when attempting to support hypotheses. One way to do this comparison is to obtain the proportion of all speech acts that express uncertainty and test uncertainty across deception treatments. For example, Hedge and Maybe/Accept-part are two speech acts that express uncertainty. A Hedge is used specifically by a speaker to introduce uncertainty into their statement: "I'm not quite sure, but I think we should probably do it. "Maybe/Accept-part, also indicates uncertainty as in the phrase "It might be." The full set of speech acts that often express uncertainty are shown in Table 27. These uncertain speech acts can be combined by summing their probable proportions. The result is the probable proportion of speech acts that express uncertainty.

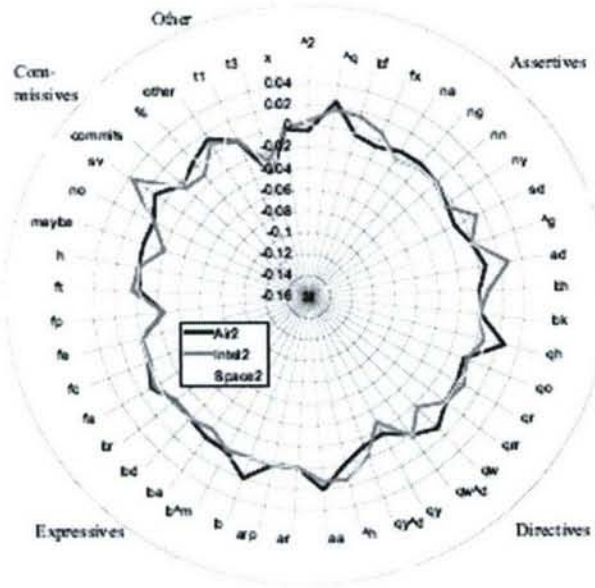


Figure 24. Sample speech act profile from the StrikeCom corpus showing submissive and uncertain behavior by the deceiver.

Table 27. Speech acts that often express uncertainty.

Acknowledge(Backchannel)	Other answers
Appreciation	Open-Question
Backchannel in question form	Or-Question
Declarative Yes-No-Question	Or-Clause

Declarative Wh-Question	Response Acknowledgement
Dispreferred Answers	Signal-non-understanding
Downplayer	Yes-No-question
Negative-non-no answers	Wh-Question

b) Detecting Deception Using Speech Act Profiling Trained on the StrikeCom Corpus

The first study showed that speech act profiling shows promise in deception detection by showing that it can be used to find uncertainty in online conversations. It showed that deceptive participants in three-person online conversations have significantly greater proportion of speech acts that express uncertainty than their partners. The study did have at least one shortcoming. The training corpus, which was used to create the speech act profiling model for detecting uncertainty, was the SwitchBoard corpus of telephone conversations. Though this is the largest corpus to be manually annotated with speech acts to date, it is nevertheless a collection of telephone conversations, not online conversations. Despite the differences in language use between telephone and online conversations, the conversations must still be done in language that is understood by all parties of the conversation and is therefore manageably different for the purposes of speech act classification. Nevertheless, using a corpus that is more similar in language use should still produce better results. Furthermore, in the SwitchBoard corpus participants are dyads discussing a number of general topics, while the data used in the study were chat logs from a three-person online game. Such differences could cause problems in applying SwitchBoard tags to the current data.

A portion of the corpus was annotated with the acts described above. This portion is comprised of 47 games containing a total of 7112 annotated utterances. This portion was used to train the speech act profiling model. The resulting model was then tested on 33 games, 16 of which included a participant who was instructed to be deceptive. Participants in the other 17 games were not given any instructions related to deception. Running speech act profiling on these conversations resulted in estimates of the number of each speech acts uttered by a participant during the game. These were then divided by the total number of utterances produced by that participant during the game resulting in the proportion of each speech act used during the game.

As a measure of uncertainty, the proportion of questions during the online conversations was examined. The previous study had shown a significant difference in uncertainty when measured with a number of speech acts including questions, a distinction between statements and opinions, certain back-channels, and hedges. The current results showed a weak trend on this single crude measure suggestive of deceivers expressing more uncertainty during the game. Deceivers had a smaller proportion of utterances classified as strategy and asset placement than non-deceivers, thus lending support to the premise that deceivers were being cognitively lazy in their choice of how to deceive. Rather than attempting to change the strategy of the group, deceivers simply inserted misinformation into the results or did not follow the strategy of the group when placing assets. However, the fact that deceivers had more total utterances than their group members indicates that even though deceivers didn't participate in strategy and asset placement, they did fully participate in the remainder of the conversation.

The two studies combined illustrate that speech act profiling may be useful in detecting uncertainty as a precursor to detecting deception. The uncertainty uncovered by speech act

profiling could be fused with other cues such as those used in message feature mining to increase deception detection accuracy. It should be promising to take a data mining approach to speech act profiling and deception detection by using all of the speech acts as features for a data mining model such as support vector machines. Then the resulting accuracy rate could be compared to message feature mining alone and speech act profiling and message feature mining combined.

3. Self-reported Communication

Two experiments—BunkerBuster and StrikeCom—investigated the impact of communication modality and deception on the quality of a group's communication and ultimately their performance on a team task. We proposed that the impact of modality richness on group performance is mediated by the quality of the communication that underlies the process and that the effects of CMC on group outcomes can be better understood by considering the qualities of communication that accompany each type of group interaction. The model of mediated interactions in Figure 25, adapted from Stoner (Stoner, 2001), shows the proposed relationships among initial structural affordances of the modality, deception, and communication qualities, and the relationship of communication qualities to group outcomes.

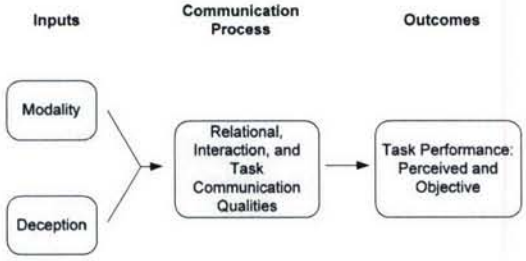


Figure 25. Model of mediated interaction.

Stoner (Stoner, 2001) and others (Burgoon et al., 2000; Burgoon et al., 2002) identified thirteen dimensions of communication quality from previous computer-mediated and group communication research that through principal components factor analysis could be reduced to three supra- or meta-dimensions of relational, interactional, and task communication qualities:

1. *Relational Quality*--addresses the personal relationships between participants. It includes measures of involvement, connectedness, similarity, openness, positivity, composure, and persuasiveness.
2. *Interaction Quality*--measures the team's ability to coordinate and execute the sharing of information during the task. It includes interaction coordination (how well conversation was coordinated, smooth, and fluent), communication appropriateness, expectedness (communication typicality), and richness of the communication itself.
3. *Task Quality*--measures the effectiveness, efficiency, and task focus of the group's communication. It includes task orientation (percentage of communication related to completing the task), efficiency, and level of critical analysis/feedback.

Relevant to deception, it was hypothesized that teams with a deceptive member would be characterized by different communication qualities than groups with no deceivers and would perform less well. Deceptive teams performed significantly worse on the task than groups were no deception was present, yet surprisingly, the presence of deception did not significantly alter the communication quality ratings (see Table 28).

Table 18. BunkerBuster and StrikeCom Communication Quality Means by Truth/Deception Condition

		BunkerBuster		StrikeCom	
Dimension	**Condition**	**N**	**Mean**	**N**	**Mean**
Relational	Truth	35	5.487	47	5.468
	Deception	15	5.430	48	5.401
Interaction	Truth	35	5.498	47	5.570
	Deception	15	5.356	48	5.546
Task	Truth	35	5.663	47	5.645
	Deception	15	5.570	48	5.535

Deceivers were able to successfully execute their deception without the rest of the team members noticing any difference in the quality of the communication. The communication qualities were positively correlated with team performance in the deceptive but not the truthful condition. Where deception adversely affected communication qualities, it also adversely affected team performance. Where teams were able to establish effective communication by achieving high involvement and mutuality, by maintaining a smooth and efficient interaction, and by fulfilling task-related responsibilities they were also able to mitigate the influence of deception.

As regards the influence of modality on communication, we found—as predicted—that the audio modality either exceeded or matched FtF communication in terms of relational, interaction, and task qualities in both the BunkerBuster and StrikeCom data. The implication to be drawn is that loss of visual nonverbal cues from the audio condition does not impair the ability to build involvement or mutuality on the team. Neither does it impair smooth, appropriate, and tension-free interaction or the efficient and effective exchange of information, analysis, and evaluation. The audio condition is sufficient for enabling the smooth coordination of information exchange. The addition of the visual properties of FtF interaction does not yield a sufficiently large benefit to warrant the increased bandwidth or expense needed to make the visual aspects of interaction available and risks worse deception detection. (Of course, with much larger group sizes, such as large team videoconferencing or distance education with large classes, the situation would likely change because the coordination of turn-taking and distinguishing different speakers would become more challenging.) As expected, the text-based modality received the lowest ratings for both relational and interaction quality. These results are consistent with the principles of media richness in that the lack of multiple modes to send cues impaired the ability of the team to build cohesiveness, connection, and positive interpersonal relationships as well as to coordinate smooth message exchange.

The research found no significant difference on average between the deceptive and nondeceptive teams on the communication quality ratings. These results can be seen as both favorable and problematic. Since deception in some shape or form is so prevalent in everyday discourse, it is encouraging that the presence of deception need not impair the quality of the group's communication. Groups can still foster involvement, mutuality, similarity, and coordinate message exchange in spite of the presence of deception. However, from a diagnostic standpoint, the fact that communication qualities did not prove to be a means in and off themselves to identify deceptive communication, means that deceivers can be successful at perpetrating their

deception without naïve team members perceiving their ulterior motives through their communication. Deceivers may even capitalize on the group's communication patterns to achieve their own ends. In groups that are struggling to collaborate and communicate, the deceiver may commit the deceptive act by providing sparse details that are difficult to understand. However, in groups that are achieving high levels of communication quality, the deceiver may use the opposite approach and provide ample information to present a credible appearance and blend in with the team's existing communication norms.

4. Global Assessments

Another approach to analyzing deceptive and truthful behavior patterns at a more macroscopic and multimodal level is to take a "global," subjective approach, i.e., to judge deception in a gestalt fashion the way naive judges do. In the Deceptive Interviews experiment, trained human observers made judgments of interviewees along six dimensions: *involved-uninvolved, dominant-submissive, pleasant-unpleasant, active-passive, relaxed-tense*, and *formal-informal*. Ratings were made after each of the 12 interview questions and averaged across blocks of truthful or deceptive responding.

Multiple regression and correlation analyses were conducted to identify what verbal features contributed these global judgments. The features most responsible for these judgments are listed below. Also listed are nonverbal features associated with each dimension.

1. *Involvement/dominance*: greater verbosity, diversity, specificity, affect-laden language, expressiveness, spatial immediacy, personalization (1st and 3rd person pronouns), simple syntax; fewer errors; more nonverbal immediacy, kinesic and audio expressiveness, altercentrism, composure, and smooth interaction management
2. *Pleasantness*: more diverse, personalized, and active language; more facial pleasantness; warmer voices with more pitch variety and relaxed laugher
3. *Arousal and relaxation*: more redundant affect-laden language; more adaptors and blinking; more random movement (if aroused) or less random movement (if tense); pitch elevation and loudness changes
4. *Formality*: less diverse language, less personalized, lower activation score, more passive voice, more spatially nonimmediate language; more postural symmetry; less physical activity

These global judgments were the dependent measures in repeated measures analysis of variance and discriminant analysis. The objective was to determine if and how these global dimensions relate to actual judgments of truth or deception. Results showed that deceivers were initially *less involved, dominant, pleasant, composed* and *formal* than truthtellers. Over time, deceivers tended to match the communication patterns of truthtellers. In other words, deceptive responding was harder to distinguish from truthful responding as the interview progressed. The implication is that accurate recognition of deception would be optimal early rather than late in an interaction. This conclusion is echoed by the next analysis on the deception index.

5. Multimodal Deception Index

Prior research on Interpersonal Deception Theory (IDT) predicts that as interpersonal interactions provided feedback, deceivers they will strategically adjust their behaviors to feedback cues and become more capable of deceit (Burgoon & Buller, 1994) (Burgoon, Buller,

Dillman, & Walther, 1995b). Research has also shown that planning time can moderate deceptive cues (Miller & Stiff, 1993), indicating that potential deceivers are aware of their pending deception. Researchers believe that the length of an interaction can affect the extent to which deception is required and deceit can be observed, with longer turn exchanges (called interacts) being more likely to generate deceptive cues (Ekman, 2001). This led to a series of hypotheses predicting that:

1. *deceivers will betray their deceptive intentions through subtle indicators even prior to the deceptive portion of an interview,*
2. *indicators of deceit will be more apparent during deceitful phases of the interview,*
3. *longer interacts will generate more deceptive cues than shorter ones, and*
4. *deceiver and truthteller behaviors will converge over time.*

These hypotheses were tested in the mock theft experiment. The first step was to determine which metrics would be indicative of deceit. Past research has failed to identify any single metric which is consistently and strongly associated with deceitful behavior (DePaulo et al., 2003; Vrij, 2000). The 183 available metrics were based on the positions and movement of ellipse tracking of the hands and head of the interview subjects. Metrics were calculated based on the averages and variances that could be derived from positions, relative angles, relative positions, and movement between frames. The resulting data set resulted in metrics that were both highly skewed and often bounded at zero. Although efforts were made to normalize these metrics, it proved intractable. An alternative approach was then taken which dichotomized the data based on the mean across subjects within each interact and then summed indicative metrics to create a deceit index. To determine which metrics were most likely to accurately indicate deceit or truthfulness, each metric was scored as a 0 or a 1 based on whether it was above or below the mean. These scores were then correlated with deceit using the sum of the first 6 interacts of the deceitful phase. Additionally, any metric which had a point-polyserial significance < 0.10 was considered to have a strong correlation with deceit. A set of 19 correlates with deceit was ultimately identified.

Next, each individual interact was scored with a deception index between 0 and 19, based on the dichotomized metrics for that question and whether each one was more or less likely to be indicative of deceit. Metrics that were negatively correlated with deceit were reverse coded. When between-subject factors were analyzed, they were dichotomized for each interact separately to eliminate question effects. When question effects and question-specific moderators were being evaluated, the metrics were dichotomized across all subjects and questions.

An analysis of Question Length and Deception Index revealed that the correlation was not significant. However, a visual representation appeared to show a strong correlation, therefore an alternative approach was taken, where questions were categorized as being either open or closed ended. There was a strong ($p<0.01$) correlation between the deception index and question type, indicating that question effects could significantly impact any analysis of deception metrics across interacts unless controlled for. Several expected interaction effects were also examined, such as Guilt x Question Type x Question Length. While none of these reached a significance of $p<0.05$, this may have been due to the limited sample size.

An analysis of participant deception indexes across the interview was conducted, with all participants expected to experience a lower emotional state of arousal and consequently exhibit fewer indicators of deceit and have a lower deception index over time. However, when question effects were controlled for, there was no significant reduction in the deception index over time.

An analysis of the difference in deception indexes of deceivers and non-deceivers during the non-deception phase (the first 8 questions) showed that they were significantly different ($p < 0.02$), but converging (negative parameter estimate, $p < 0.05$ for guilt by question interaction). This is potentially indicative of several significant things. The deception index was able to distinguish between deceivers and non-deceivers during the non-deceptive portion of the interview, which implies that deceivers may exhibit behavioral differences even when telling the truth if deception is anticipated. In cases where interviewers are attempting to get a baseline reading of truthful behavior before asking probing questions that would require deceit, subjects who are anticipating that deception will be required may already be exhibiting the behaviors that are characteristic of deception for them. This early exhibition of deceitful indicators may invalidate the concept of acquiring a baseline. However, it may make it possible to detect individuals who feel they might need to be deceitful, even if they are never asked questions which directly spark a deceitful response. At the same time, the convergence between deceivers and non-deceivers does imply that there might be a limited interview window during which deception can be detected before deceivers have adjusted to the environment.

The significance of the difference between the deception index of deceivers and non-deceivers appears to increase during the portion of the interview requiring deception. This supports the hypothesis that deceptive cues and the deception index would increase during the deceptive portion of the interview.

Finally, it was believed that the deception index at the start of the deception phase would be moderated by the amount of time spent questioning the subjects during the non-deceptive phase of the interview. A temporal analysis was done to determine if the start time of the deceptive portion of the interview (controlling for guilt) would be a significant independent variable for determining the deception index. The results indicated there was no correlation between the non-deceptive phase of the interview and the deception index. This result is significant as it might imply that regardless of how long deceivers are asked truthful questions, once subjects are in a position where active deception is required, they will still exhibit significant variances in behavior and deceit will still be detectable.

In conclusion, the study determined that the dichotomization of hand and head physical movement metrics and the formation of a Deception Index may serve as a reliable indicator of deceit, including during truthful phases of an interview when the deceiver anticipates the act of deceiving. There are significant question affects that must be controlled for, as well as temporal effects as the interview continues, but when the interview transitions into a phase where active deceit is required deceivers may again reveal significant differences from non-deceivers even if behavior was converging during truthful phases of the interview. It was determined that a prolonged truthful section could generate behavioral differences in potential deceivers between the initial and later portions that would reveal anticipated deception without actual deception ever taking place. In addition, if a deception index can be generated from standardized data

which individuals can then be reliably scored against, a real-time deception index could be generated based on real-time automated analysis of movement.

E. Moderators of Deception

The successful detection of deception depends on many different factors, as articulated in CET (see Carlson & George, 2004) and IDT. The factors that affect the ability of the deceiver to be successful in his task, such as his motivation to lie and his intrinsic ability to do so, are matched by similar factors that affect the success of the receiver in detecting deception. For example, the receiver may or may not be motivated to detect deception, and receivers will vary in their intrinsic abilities to detect deception.

In addition to moderating factors that reflect on the deceiver and the receiver, there are other factors that come into play. One of these is the nature of the relationship between the deceiver and receiver. It intuitively follows that receivers who know the deceiver well should be better able to detect deception when it is present than would receivers who do not know the deceiver at all. Another intervening factor is the communication medium used for the deceptive exchange. For a variety of reasons, reflecting a number of theories about media and their differences, the use of some media should make it easier to detect deception, while for other media, detection should be impeded. Carlson, et al. (Carlson & George, 2004) present many of these reasons why media may differ in their abilities to aid or deter deception detection.

The relationship between media and detection is one of the under-researched areas in study of deception, so investigating how media use affects deception detection was one of the major foci of our research at Florida State University. Another under-researched area in deception is the study of deception in groups. Most all of the deception research in the communication field over the past several decades has focused on dyadic communication, but there is no logical reason to expect that people lie only when communicating with one other person but do not deceive when communicating in groups of three or more people. The study of groups and deception was also one of the major foci of the work at Florida State University (FSU).

Four of the studies that dealt with media and deception and with groups and deception are summarized here. Each study is summarized in terms of its research model, its hypotheses, its research design, and its findings. The first two studies deal with interviews about false résumés, with media as a key moderating variable. Both of these studies also feature warnings to receivers as a moderating variable, and the second résumé study also includes training as a moderator. The third and fourth studies summarized below deal primarily with deception in groups, although group size was not manipulated within these studies. The third study also includes communication media and the number of suspicious receivers as moderators. The fourth study includes group member familiarity and task complexity as moderators. At the conclusion of these four summaries, the overall findings are reviewed.

Two other moderators that were also investigated in the UA and MSU experiments were modality and motivation. Modality effects in the DSP, mock theft, and StrikeCom experiments were described above. They demonstrated that modality significantly moderated results in all of those experiments. Not only did deception verbal and nonverbal displays differ, so did detection accuracy rates.

Motivation was examined in the mock theft experiment. Although the earlier summaries alluded to motivation effects, in this section we summarize the extent to which motivation altered performance of both deceivers and truthtellers.

1. First Résumé Enhancement Experiment

The research question driving this study was whether and how deception detection success varies with media and suspicion. The hypotheses under test were as follows:

1. *Deception detection accuracy will vary with media used for communication, with interviewers using richer media being more accurate than those using leaner media.*
2. *Warned interviewers will be more accurate at detecting deception than unwarned interviewers.*

The research design crossed four types of computer media (e-mail, chat, chat with audio, and audio only) with two categories of induced suspicion, present or absent. All audio conversation was recorded with a tape recorder, and all e-mail and chat transcripts were saved and archived. Participants were randomly assigned to one of the four communication media conditions. There were 20 dyads, or 40 participants, in each condition, with the exception of the audio only condition, in which there were 18 dyads, for a total of 156 individuals. Within each condition, half of the receivers were warned about the possibility of deception and the other half were not.

One participant in each dyad was assigned the role of deceiver. Deceivers were initially told that they were needed to help the department develop a list of minimum requirements for a scholarship under development and to make themselves appear to be as competitive as possible on the application. After completing the application, deceivers were then told that they would be interviewed by another student located elsewhere in the building, and that they would have to convince the interviewer that the application was completely legitimate. In the meantime, the other member of the dyad was assigned the interviewer role and was told that he or she would be interviewing a student applying for an academic scholarship. The application was sent electronically from the deceiver's computer to the interviewer's computer. Interviews were conducted over the assigned computer-mediated medium. The interviews, which averaged 23 minutes in duration, were unscripted. Following the interview, both subjects were given questionnaires to complete. Deceivers were asked to identify all of the deceptive information on the application. Interviewers were asked if they believed the applicant was honest and to recall what information the applicant was lying about.

Analysis of the doctored resumes, in the form of scholarship applications, revealed that they contained as many as 18 deceptions, with an average of 8.6 deceptions per application. The overall deception detection accuracy rate on the part of the interviewers was 8.1%. For interviewers who had been warned about the possibility of deception, the detection accuracy rate was better, at 14.5%. For those who had not been warned, the accuracy was only 6.8%.

As for the hypotheses, there was no support for the hypothesized influence of medium of communication on deception detection accuracy based on media. On the other hand, induced

suspicion did moderate accuracy. Warned interviewers were much better at detecting deception than were interviewers who were not warned.

2. Second Résumé Enhancement Experiment

The next study replicated the previous one while adding the variable of training to the research design. It examined two levels of media (lean and rich), presence or absence of warnings to induce suspicion, and presence or absence of training. It tested four hypotheses:

1. *Deception detection accuracy will be greater for receivers who use richer media than for those who use leaner media.*
2. *Deception detection accuracy will be greater for receivers who are warned of the potential for deceptive communication than for those who are not warned.*
3. *Deception detection accuracy will be greater for receivers who are trained in deception cue recognition than for those who are not trained.*
4. *Deception detection accuracy will be greater for receivers who are trained in deception cue recognition and warned of the potential for deceptive communication than for those who are trained only, warned only, or not trained or warned.*

Students were induced to enhance their resumes and defend those enhancements when communicating with an interviewer via either lean or rich electronic media. If the interviewer was in one of the warning treatments, the researcher conveyed the statistic that about 40% of job applicants lie on their resumes and to be aware of that statistic when interviewing. Subjects in the training cells attended a deception-cues training session one week prior to their scheduled experiment date. Another treatment combined training with warning.

The applicant was asked to do whatever it took to look like the best student for the purpose of setting standards for a scholarship. The scholarship application template included places to put course names and grades along with grade point average, past and present employment, and community service. The applicants were told that during the interview they should be as convincing as possible in defending the information in the enhanced resume. The scholarship application was sent to the receiver via Microsoft NetMeeting. Subjects using lean media used a web-based e-mail provider, Hotmail, with accounts created specially for the experiment. For audio over Internet relay chat, subjects communicated using microphones and headphones. The interviewer asked questions of his or her choice for up to 20 minutes. Before and after the interview, the subjects completed questionnaires.

Of the four hypotheses, only the third—that training was positively related to deception detection accuracy—was supported. This is an encouraging finding that bolsters our previous experiments showing some benefit to training, especially when the information delivered is extensive and tas-relevant. Like the previous resume experiment, the medium over which communication took place did not alter results, but unlike that investigation, induced suspicion failed to improve detection accuracy. These two investigations together mirror the mix of previous findings as regards suspicion. The lack of impact of medium may have been a function of the task and the great difficulty that interviewers had in recognizing embellished resumes.

3. Deception in Groups, Allocation Task

The next experiment addressed several issues. First, like the StrikeCom experiments, it addressed the issue of whether deception performed in groups places new challenges on deceivers or facilitates their deception. Second, it considered the impact of suspicion among group members on detection and whether the number of suspicious members would make a difference. Third, it considered two issues regarding computer-mediated communication--whether groups with members acting deceptively perform and behave differently if in the same room or dispersed, and if using computer-mediated communication or not. Two dependent measures were included to examine both how much deception was produced and how accurately other group members detected it. Figure 26 presents a model of the hypotheses under test. In words, it was hypothesized that:

> *1. Deceivers will submit more deceptive information to group members (a) when using computer-mediated communication than when not using computer-mediated communication, (b) when group members are dispersed than when group members are co-located, and (3) the less suspicious group members are.*
> *2. Deception detection accuracy will be lower with (a) receivers using computer-mediated communication receivers without CMC, (b) when dispersed than when collocated with other group members, and (c) group members are less suspicious.*

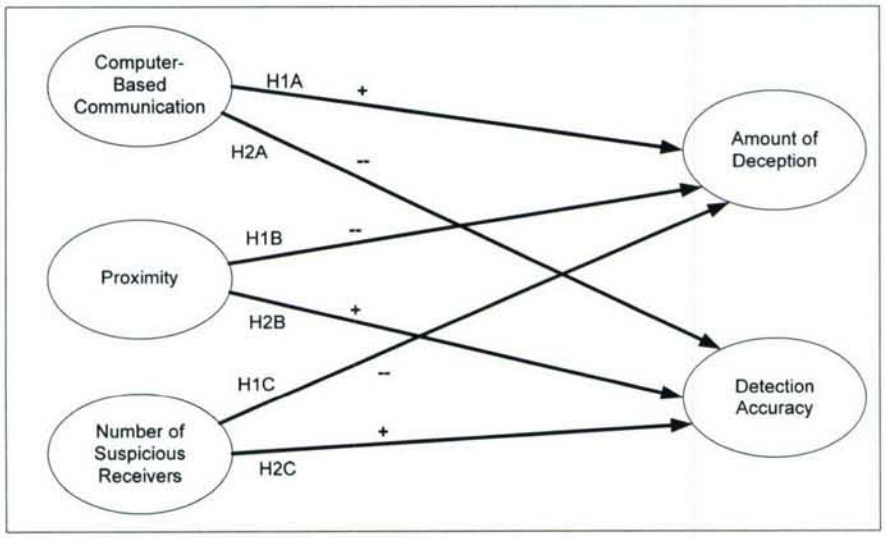


Figure 26. Model of Relationships among Variables in Group Deception Experiment.

The 2x2x3 factorial design crossed the use of computer mediated-communication, the proximity of the group members, and the number of group members who were forewarned about the possibility of deception. One randomly-selected group member was assigned the role of deceiver, unbeknownst to the other two group members. The group was given a resource allocation task to complete, with the deceiver having been previously given one of the allocation options to argue for, and having been promised a financial reward for successfully leading the other group members to choose the deceiver's assigned option. Groups either communicated in a traditional face-to-face manner, over audioconferencing, or via a computer-mediated group support system. Likewise, group members were either collocated or were physically dispersed.

The third factor in the design dealt with forewarning the "receivers" of the deceptive communication, and either both receivers, one randomly-selected receiver, or no receivers were given warnings prior to group discussion. Groups (N = 60) of three undergraduate students completed the task, each with the deceiver submitting at least one purposely false statement. The group members were unfamiliar with one another.

In overview, results regarding production and detection of deception were as follows:

- An average of 1.82 lies were submitted per group session.
- Only 8.8 percent of the lies were detected by the group receivers.
- Deceivers were successful in swaying the group decision in their favor 72 percent of the time.
- There were very few instances (4 times out of the 60 groups) in which false positive judgments were issued by receivers.

Tests of the specific hypotheses produced support for H1a: Deceivers using the group support system submitted 2.2 lies per session, as opposed to non-CMC deceivers, who submitted 1.43 lies. H1b was not supported. There was no significant difference in the amount of deception submitted due to proximity. H1c was supported. Deceivers lied more when there were two forewarned receivers in the group (2.10 lies) or one forewarned receiver (2.00 lies) than in groups with no forewarned receivers (1.35 lies).

H2 was not supported. There was no difference in detection accuracy between computer-supported receivers and non-CMC deceivers, between proximal and distal groups, and among groups with different numbers of forewarned receivers. Put differently, the overall level of detection accuracy was not moderated by medium, proximity, or amount of suspicion present in the group. These and the previous findings would be encouraging that deception can be detected with equal accuracy regardless of modality or use of CMC except for that fact that overall detection was so poor across the board. These results are better interpreted as indicating the ease with which deception can be introduced into group processes and undermining group performance. The lack of impact of number of suspicious members is explainable by deceivers producing more lies when faced with suspicion and thus overcoming them. These results, then, reinforce the greater advantage that deceivers have over receivers in spite of their suspicions.

4. Deception in Groups, StrikeCom III

This next study performed at FSU also focused on deception in groups performing a computer-mediated collaborative task. In this case, the task was the StrikeCom simulation in which teams of players to cooperatively search a grid-like game board for a fixed number of enemy targets, which they attempted to destroy with bombs on their final turn. The moderators of task complexity and team member familiarity were examined. The hypotheses under test were:

1. *Groups facing a complex task will be less accurate at detecting deception than groups facing a less complex task.*
2. *Groups with members that are familiar with each other will be more accurate at detecting deception than groups with members that that are not familiar with each other.*

3. *Groups with members that are familiar with each other and a low-complexity task will be more accurate at detecting deception than groups with members that are not familiar with each other or a high-complexity task.*
4. *Groups with a high-complexity task will (a) have lower task performance than groups with a low-complexity task and (b) suffer worse task performance decrements from the presence of deception than groups with a low-complexity task.*
5. *Groups with members who are familiar with each other will have higher task performance than groups with members who are not familiar with each other.*
6. *Task complexity interacts with member familiarity to affect task performance such that groups with a low-complexity task and group member familiarity will be the least negatively affected by deceivers.*

The relationships among variables and hypotheses are modeled in Figure 27.

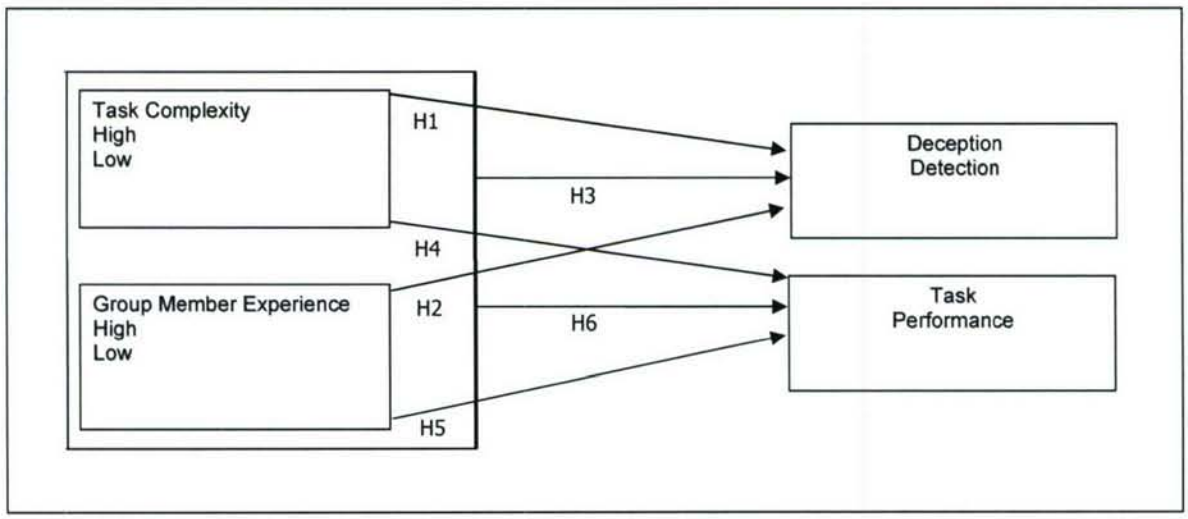


Figure 27. Model of Effects of Task Complexity and Familiarity on Deception Detection and Task Performance under Deception.

For this experiment, the game included a built-in chat area that allowed for real-time computer-mediated communication between players. Participants were students (N = 160) who formed 40 groups for the main experiment. In each group, one member was solicited to be a deceiver. The deceivers were given the target locations that their group members were trying to find, and they were told that their goal in the game was to deceive their team members about the true locations of the enemy targets and to get them to target empty grid squares on the game board. For half of the teams, the number of grid squares and the number of targets in the game were increased to make the game more complex. In addition to task complexity, we manipulated group member experience. Half of groups also had members that had experience with each other, and half of the groups had members with no familiarity. We also warned all participants about the potential of deception in collaborative group settings. Group members could not see each other during the experiment, and so they only communicated using the chat feature of StrikeCom. Groups were scored on their task performance and rated the deceptiveness of their group members after they completed the game. To provide a truthful baseline, an additional 20 groups without deceivers conducted the task so that we could compare the impact of deceivers on group task performance

to this offset control group. We manipulated these groups' task complexity but not their group member experience.

Overall analysis of deception detection and group performance revealed that:

- Even though they were warned, on average, groups judged deceivers as being more honest than deceptive (3.32 on a 7-point level of deceptiveness scale).
- The groups with deceivers were only successful with 48.5% of their final strikes in the game and groups without deceivers were successful with 67.1% of their final strikes.

H1 was supported. Groups performing a low-complexity task had greater deception detection accuracy (-25.41 average detection score) than did groups with a high-complexity task (-43.90 average detection score). This would seem to support the typical expectation that more complex tasks demand more cognitive resources that reduce cognitive investment in deception detection, except that deceivers unexpectedly had a negative impact on the task performance of groups with the low-complexity task and group member experience. It may that familiarity invoked a truth bias so that despite the cognitive capacity to scrutinize message contents, group members did not exert the effort. These results also meant that H6 was not support.

H2 and H3 also were not supported. There was no difference in deception detection accuracy due to group member familiarity, an interaction between familiarity and complexity. However, in support of H5, Familiar groups had higher task performance (0.55 average game score) than unfamiliar groups (0.41 average game score).

H4a was supported. Groups with a low-complexity task had higher task performance (0.53 average game score) than did groups with a high-complexity task (0.43 average game score). H4b was not supported. There was no difference in the effect of deceivers on task performance due to task complexity.

5. Effects of Motivation, Mock Theft Experiment

It will be recalled that one of the variables manipulated in the mock theft experiment was motivation. Those in the high motivation condition were told that success in evading detection is an important social skill to cultivate and that if they succeeded in convincing the interviewer of their innocence and credibility, they would receive a monetary bonus and eligibility for a large prize. Those in the low motivation condition did not receive these instructions or incentives in advance. (However, they received the same amount of money after the fact.)

Because motivation has been identified as a major influence on deception performance (see, e.g., Burgoon, 2005; Burgoon & Floyd, 2000; DePaulo & Kirkendol, 1989), a major objective of the mock theft experiment was to determine the extent to which motivation influenced interpersonal—and interactive—deception. It was hypothesized that:

Deceivers who receive motivation-inducing incentives will (a) show fewer decrements in their verbal and nonverbal deception performance and (b) succeed in their deception more than those who do not receive such incentives.

The analyses of verbal and nonverbal behaviors of truthtellers and deceivers produced a host of main effects and interaction effects for motivation. It is important to put deception results in the context of how much motivation generally affected behavior. Compared to those who did not receive the motivation induction, highly motivated interviewees (truthful and deceptive) displayed:

1. More postural shifts, especially at the beginning
2. More head movement
3. More adaptor gestures, especially during the theft questions
4. Longer illustrator gestures and gesturing a larger percent of the time, especially during the theft questions
5. More talk time
6. Talk for a larger percentage of the interview block
7. More frequent and a higher rate of vocalized pauses
8. More silent pauses (by truthtellers in the audio condition)
9. More total nonfluencies
10. Longer periods of head stillness
11. Longer periods of non-gesturing
12. Longer interviewer turn lengths
13. More rapid paced conversation (more turn exchanges within a block)
14. More sentences, words and verbs
15. More use of spatial sensory terms
16. Less lexical diversity

As for the interactions with deception, all of the following variables were moderated by motivation:

1. visual details
2. imagery
3. redundancy
4. passive verbs
5. talk time duration
6. turn-switch pauses
7. silent pauses, "other" nonfluencies, and total nonfluencies
8. duration and percent of head movement and head stillness
9. postural shifts

In general, relative to highly motivated truthtellers, deceivers:

1. used less imagistic language
2. used more passive voice
3. moved their head for less total time and percent of turn
4. made fewer postural shifts
5. were more gesturally active at the beginning but less so during the work questions
6. talked less and took shorter turns
7. had fewer silent pauses, other speech disfluencies, and total nonfluencies
8. had shorter turn-switch silences (response latencies)

Whereas the general pattern for motivated communicators was one of greater activation in the form of gesturing, head movement, postural shifts, talk time, rapid tempo, and disfluencies associated with more talk, motivated deceivers showed greater restraint. They talked less, had fewer speech impairments, and moved less. From a motivation impairment standpoint, these behaviors do not paint a picture of impaired performance. Adaptor gestures, which are normally expected to be controlled in public, were not more evident. Neither were speech disturbances. It was the motivated truthtellers who exhibited more disfluencies and longer turn-switch pauses. Motivated deceivers' nonverbal behavior showed greater reticence and behavioral control, which is consistent with a strategic perspective. From these data, it would be a mistake to focus on nervous gesturing, postural squirming and restlessness, or speech errors as indicators of deceit when the suspect is likely to be motivated. However, these behavior patterns would characterize the low motivated deceiver.

Linguistically, motivation had less of a direct or moderating impact. In general, motivated communicators were more loquacious, used more spatial language but had less lexical diversity (possibly as a function of talking more and therefore producing an artificially lower index of lexical diversity). Motivation did lead to truthtellers using more vivid, active language in the form of more imagistic vocabulary and more active voice, whereas motivated deceivers used less imagistic language and more passive voice. Other motivation effects for visual details differentiated high- from low-motivated truthtellers and redundancy differentiated low-motivated truthtellers from low-motivated deceivers. Thus, motivation had far less of an impact on verbal than nonverbal behavior. In some respects, this is a benefit from a deception detection standpoint because the same indicators would be relevant regardless of motivation. Using the nonverbal cues as successful discriminators would require identifying or surmising the level of motivation that a suspected deceiver was experiencing.

6. Summary of Moderator Variable Effects

The findings regarding the influence of moderator variables on successful deception detection are reviewed below, organized by moderator.

Communication modality. The DSP, mock theft, and StrikeCom experiments at UA all found numerous main effects and moderating effects for modality. In some cases, modality interacted with both deception and motivation to produce complex effects, but more often the medium in which communication took place exerted direct effects. From a detection standpoint, this means that the backdrop against which any judgments are made must factor in what is typical for a given modality. Among the resume and StrikeCom experiments at FSU, medium did not make a difference but may have been due to the extreme difficulty of the tasks involved. In other words, deception detection was equally difficult regardless of communication medium used. An exception in the latter case is that deceivers communicating via a group support system lied more than deceivers who communicated face-to-face with their groups.

Suspicion. Suspicion was induced in three experiments through warnings about the likelihood of deceit. For the first résumé study, warned interviewers were better able to detect deception than were non-warned interviewers. For the second résumé study, there were no differences between warned and non-warned interviewers in their detection success. Here again, the mixed results may have been due to the difficulty of spotting resume enhancements. In the resource allocation

task, where we manipulated the number of group members who were warned of possible deception, where none, one or both of the receivers were warned. Deceivers lied more where two receivers had been warned compared to when no receivers had been warned.

Training. One study investigated training and found that interviewers trained to detect deception did better than their untrained peers in deception detection.

Task complexity. One study manipulated task complexity and found that groups working on simpler versions of a task were better than their peers working on a more complex task at successfully detecting deception. Incidentally, groups working at a simple task outperformed their peers who were working on a complex task.

Group member familiarity. One study compared groups of people who knew each other to groups that did not know each other. While there were no differences in the groups' abilities to detect deception, groups where members knew each other outperformed groups where members did not know each other.

Motivation. Motivation exerted a stronger moderating influence on nonverbal than verbal variables. In the mock theft experiment, those who received a psychological induction and monetary inducements prior to their interview regarding the theft were generally more active verbally, vocally, and kinesically. Motivated truthtellers showed more speech disturbances, turn-taking delays, and nervous gesturing. Motivated deceivers showed more reticence and behavioral restraint.

Additional research is clearly warranted on all of these moderators to see how well they generalize across different types of tasks and different levels of jeopardy or motivation. All have the potential to alter not only the behavioral displays of truthtellers and deceivers but also the accuracy with which deceit is detected.

IX. Identifying Cognitive Heuristics

A. The Nature of Cognitive Heuristics and Biases

One of the most documented claims in the deception literature is that humans are poor detectors of deception. A recent meta-analysis reveals that although people show a statistically reliable ability to discriminate truths from lies, overall accuracy rates average 54%, or only a little above chance (Bond & DePaulo, 2006).

One reason for this poor detection accuracy rate is that in potentially deceptive situations, people may rely on mental shortcuts to help process information (Burgoon, Blair & Strom, in press). A basic tenet of social cognition is that people are cognitively lazy and will rely on a variety of mental shortcuts to reduce their mental effort rather than process incoming information fully. The most widely cited bias in the deception literature is the truth bias (Levine, Park & McCornack, 1999) but there are a variety of other heuristics that those attempting to detect deception may use. Although heuristics can sometimes lead to correct judgments, they often lead to biases toward over- or underestimating truthfulness.

One objective of the current research was to identify what biases and heuristics are likely to influence deception detection. Review of the research literature produced the following list:

1. *Truth Bias*: The chronic overestimate of truth or the human tendency to initially process all incoming information as truthful

2. *Lie Bias*: The opposite of the truth bias – the tendency to chronically judge incoming information and messages as deceptive

3. *Visual Bias*: the tendency to place greater weight on visual than on auditory or textual information (i.e., "seeing is believing")

4. *Demeanor Bias:* the tendency to judge some senders' communication styles as credible irrespective of their actual truthfulness

5. *Expectancy violation/Infrequency Heuristic*: The tendency to judge unexpected, novel, or infrequent, events as more deceptive (correlates with the expectancy violation theory)

6. *Availability Heuristic*: The tendency to judge the probability of an event by the ease with which similar occurrences come to mind

7. *Falsifiability Heuristic*: The ironic tendency to judge information that could be falsified as less truthful than subjective content that is less amenable to verification

8. *Probing Bias*: The tendency to judge answers in response to probing questions as more truthful, resulting in truthful responses being judged more accurately but lies being judged less accurately

9. *Plausibility Bias*: The tendency to treat content that sounds plausible as truthful

10. *Anchoring/ Adjustment*: The tendency to make estimates by starting with an initial value that is adjusted to yield the final answer.

11. *Familiarity Bias*: The tendency to view acquaintances or liked others as more truthful than strangers (equivalent to the halo effect or leniency bias)

12. *Nonverbal Conspicuousness Heuristic*: The tendency to treat more conspicuous nonverbal cues such as nervous gestures and gaze aversion as diagnostic rather than relying on valid indicators of deception

13. *Framing Bias*: The ability to influence a decision-maker's choice by the way in which the problem is stated, especially if the problem wording can capitalize on people's risk aversion

14. *Representativeness Heuristic*: The tendency to be insensitive to prior probabilities (also referred to as the base-rate fallacy, in which decision-makers fail to ignore prior

probabilities despite presentation of new information), insensitivity to sample size, and misconceptions of chance (also referred to as the gambler's fallacy).

The sheer number of possible biases and cognitive heuristics points to a major cause of detection inaccuracy. An experiment, reported next, examined several of these influences on detection accuracy to determine the severity and direction of their impact as well as their interrelationships.

B. Effects of Heuristics and Biases on Observer Judgments

Four especially salient and potentially interrelated judgment biases that were investigated are *truth bias, visual bias, demeanor bias,* and *expectancy violations bias.* Together, these biases may account not only for poor detection of deception but also more generally for judgments of communicator credibility.

The interrelationships among these biases have not been investigated previously. It may be that some are subordinate to, or artifacts of, others. The visual bias, for example, may be the product of demeanor and expectancy violations biases; or it may be a product of other factors such as the information richness of the medium. Thus, a central objective of the investigation to be reported was to examine the interrelationships among these biases and their ultimate impact on veracity judgments.

A second objective was to test these biases when judgments are applied to the kinds of message exchange that typify normal, ongoing interaction. The Bond and DePaulo (2006) meta-analysis, though quite comprehensive, included very few studies in which the stimuli that were judged were produced under fully interactive conditions, that is, ones in which senders engaged in ongoing and interdependent social interaction with the intended targets of their deceit. Given that deception typically is embedded in ongoing interaction rather than judged in isolation, and given that judgments made of naturalistic interaction differ from those made of brief, experimentally controlled stimuli (Motley & Camden, 1988), knowledge of how people make veracity judgments should be founded on the kinds of stimuli they normally encounter rather than on brief, decontextualized snippets.

The experiment utilized interviews from the mock theft experiment. It varied nonverbal cue availability and deception. Observers saw a complete videotaped interview (full access to visual, vocal and verbal cues), heard the complete interview (vocal and verbal access), or read a transcript (verbal access) of a truthful or deceptive suspect being questioned about the theft then rated the interviewee on information, behavior, and image management and truthfulness.

Four hypotheses tested the presence of each cognitive bias. Results supported the presence of all four biases. These biases were most evident when interviewees were deceptive and observers had access to all visual, vocal and verbal modalities.

As regards truth bias, compared to the 53% of all stimuli that were actually truthful, observers judged 67% to be truthful, and the average truth estimate was far above the midpoint of the scale. These results reinforce what has been a consistent finding in the literature, namely, that people are highly inclined to trust the communication of others and unlikely to question those judgments

unless faced with some major deviation that triggers a reevaluation. The current findings extend this conclusion to messages generated under fully interactive conditions.

As regards the visual bias, judgments of a person's truthfulness increased ordinally with nonverbal cue availability. The truth bias was intensified by modalities that gave observers access to nonverbal cues. Despite the fact that the same verbal content was present in every modality condition, the addition of nonverbal vocal and visual cues increasingly led observers to judge senders' interview answers as truthful.

As regards demeanor bias, results confirmed that deceivers' (but not truthtellers') overall communication was judged more favorably on measures of information, behavior, and image management with increasing availability of nonverbal cues. The communication of deceptive interviewees was seen as the most complete, honest, clear, direct/relevant, involved, dominant, credible, trustworthy, expected, and positively valenced in the AV condition.

As regards the expectancy violations bias, we hypothesized that atypical behaviors would lead to deceivers' communication being judged as a negative violation. The results were only partially supportive. Deception under both text and audio conditions was judged as a negative violation, which implies that deceptive performances can give themselves away by their departures from normative standards for content, language, and voice. Were these the only conditions to qualify as expectancy violations, we would regard the hypothesis as largely supported. However, the truthful responding via text was also among the least expected and desirable combinations. This finding bolsters claims elsewhere about the likely dampening of feelings of involvement, connection, and trust associated with text-based communication (Burgoon, Bonito, & Kam, 2006). At the same time, this finding confirms that the expectancy violations bias is not confined to communicative behavior but may also be applicable to communication channels over which such behavior is transmitted.

The results for the deception/AV condition place a further qualification on the expectancy violations bias. Communication in this condition was judged to be the most normal and positively valenced of any of the combinations, i.e., it was a positive confirmation. This makes sense when considered within the context of the demeanor bias results. Such findings could only be obtained if deceivers were more successful than truthtellers in promulgating an attractive image in the AV condition and if adding visual nonverbal cues enhanced their demeanor relative to the exact same performances in the audio and text conditions. At the same time, the results indicate that abnormal behavior by itself is not the only basis for biased judgment; behavior that is judged as exceedingly normal and appropriate can also lead to biased judgment.

To conclude, deception detection is a complex task and one commonly fraught with cognitive biases. Continued exploration of when these biases are most pronounced and what can mitigate them will aid not only in better detection of deception but also better understanding of how humans come to trust the veracity of others.

X. Training in Deception Detection

A. Curriculum Development

The training curriculum was developed in a format similar to that used at USAF training installations. One of the curriculum developers and researchers was a career Air Force officer and had developed and delivered training at USAF installations in the past, so the research teams were able to generate instructional programs relevant for the students.

The basis for the curriculum was a set of three PowerPoint presentations, each on a different topic: deception detection generally, cues used to detect deception, and heuristics for decision making that are susceptible to deception. Each presentation was designed to last for one hour. The second lecture, on cues, also included deceptive communication examples used by the instructors to illustrate the cues. These examples were either text only, audio only, or video with audio. Most examples came from past studies of deception detection, consisting of experimental subjects trying to deceive their interviewers. Other examples were specifically created and recorded for this study. The lectures, delivered by the training instructors, were also videotaped. The instructors pilot-tested all training materials, including Agent99, weeks before the study began.

For the second study, videotapes were used instead of live instructors. The tapes were professionally produced and edited. The videos were built around a taped lecture featuring an expert giving a scripted and rehearsed talk about deception. The video was inter-cut with PowerPoint slides and video, audio and text examples of both deceptive and truthful behavior. Video lectures were used instead of live instructors to standardize the presentation order and content.

B. Agent99 Trainer Development

1. System Design and Implementation

The Agent99 Trainer that was built on our previous Web-based multimedia training system called LBA (Learning By Asking) (Zhang, 2002). LBA includes "integrated multimedia" and "virtual lecture" (called *Watch Lecture* in LBA) capabilities and provides the basic infrastructure for deception detection training as a general training tool. In order to satisfy the special requirements of deception detection training, we enhanced the architecture of LBA, changed the *Watch Lecture* component, added a *View Examples* component for deception detection practice and feedback, and most importantly seamlessly integrated the two components together to facilitate better deception detection training. The system architecture of Agent99 Trainer is depicted in Figure 28. It is based on a three-layer client/server architecture, which includes client, application and database layers.

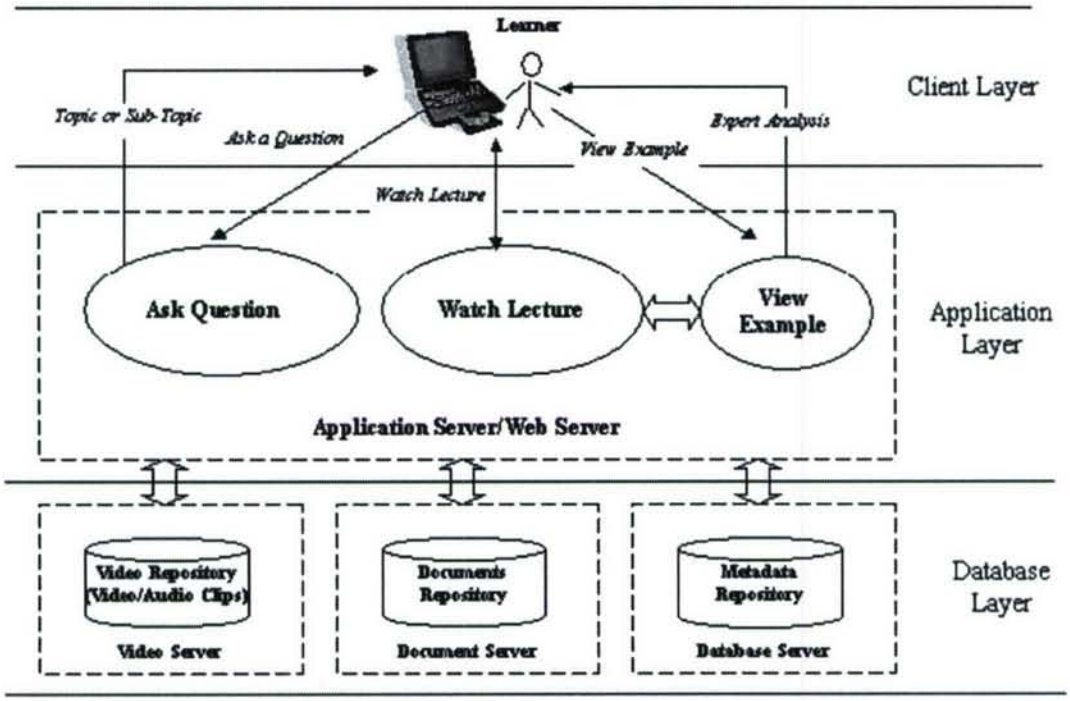


Figure 28. Client/server Architecture for Agent99 Trainer.

Client Layer: Learners access the AGENT99 learning environment through a Web browser. The client side is platform independent, and requires only a web-browser, a video player and a sound card. Application Layer: The application layer includes an application server and a Web server. The application server holds the three major modules: 1) *Watch Lecture* allows learners to watch a lecture similar as if in a traditional classroom, each lecture is divided into topics and sub-topics, 2) *View Example* provides real-life examples and expert analysis to enforce the learning of concepts and theories in the lecture, and 3) *Ask Question* allows learners to ask a question using natural language, and the system returns a list of answers to the question (a list of video clips).

2. Detailed Description of Modules: Watch Lecture

The *Watch Lecture* module provides explicit instructions on deception cues by capturing expert lectures on digital media. In order to provide multiple representations of reality (Jonassen, 1991), we use the combination of instructor's video, slides and transcripts of videos to form a "virtual lecture," which simulates a real lecture in a traditional classroom training. All the learning materials in various media types (video, slides, and transcripts) are well structured and presented in a Web interface. Seeing that an advantage of traditional classroom training is that it supports diverse activities and rich media *simultaneously* and provides an interactive and rich learning environment (Hughes 1998), the *Watch Lecture* module simulates a traditional classroom-learning environment by synchronizing the three cells of instructor's video, slides and transcripts. In the Watch Lecture module, each lecture (a lengthy video) is divided into topics and sub-topics (smaller clips). Navigation buttons and an outline of topics (implemented as a topics drop down menu) are provided so that learners can easily select any topic or subtopic in

the lecture at any time. This provides a non-linear format for instructions and allows learners to control their learning processes.

A unique feature specifically designed for deception detection training is the association of the deception examples with the topics in the lecture in order to combine the explicit instruction and practice. Practice is implemented in the *View Example* module to be discussed next). This association is implemented in two ways: 1) when the lecture (instructor's video) goes from one topic to the next one, links to the *View Example* module are provided so that learners can go directly to viewing the deception examples related to the current topic, and 2) an "Examples" drop-down menu allows learners to select any example to view while they are watching the lecture.

3. Detailed Description of Modules: View Examples and Expert Analysis

Besides the "explicit instruction" implemented in the *Watch Lecture* module, the other two critical components of deception detection training, "practice" and "feedback", are implemented in the *View Examples* module. The *View Examples* module in AGENT99 Trainer is designed to provide various types of real-life examples, scenarios and expert analysis that allow learners to practice and receive immediate and elaborated feedback. When viewing an example, the system allows learners to select different media tracks (audio, video, or text) and thus focus on cues in different communication channels (vocal, visual, or verbal). For instance, the learner may choose to listen to audio without video in order to focus on the vocal cues in deception (e.g. pitch increase) and avoid the distraction of visual cues (e.g., rigid posture). Furthermore, the *View Example* module is designed to provide learners with opportunities for reflection, which is critical for a training environment (Barab & Duffy, 2000).

Reflection is designed and implemented as follows: an example is displayed to learners without expert analysis for a pre-coded "attention span" interval (e.g., a time period of 20 seconds) that forces the trainee to think about the example for a while, and then the system will prompt and permit the learners to view the expert analysis. The expert analysis informs the learner not only of the veracity of the example but also points out the cues used to make the judgment, thereby supporting the learner's refinement of her or his own mental model. In addition, having the example and the expert analysis parallel to each other in one interface allows learners to review and reflect on the example in view of the expert analysis. Overall, this design provides repeatable opportunities for learners to think and reflect before and after viewing the analysis.

4. Detailed Description of Modules: Ask Question

The *Ask Question* module in Agent99 Trainer is the same as what in the LBA system except for new deception detection data and new indices. The *Ask Question* module allows trainees to use natural language to ask a question regarding deception detection. After analyzing the question and searching in the database, the system will return a list of video clips (topics or sub-topics) ranked by their relevance to the question. The returned video clips are presented with associated slides and transcripts in an interface similar to that in the *Watch Lecture* module. The *Ask Question* module enables learning-on-demand by allowing the learner to search for knowledge of deception detection within the lectures. We use a natural language processing (NLP) based two-phase approach for video indexing and retrieval in the *Ask Question* module (Zhang, 2002).

C. Experimental Tests of Training

The experimental tests of training were designed to test three hypotheses:

1. *Users of any version of Agent99 Trainer should perform at least as well as those exposed to traditional training.*
2. *Users of any version of Agent99 Trainer that supports multiple instructional strategies should outperform users of the baseline configuration of Agent99 Trainer, which supports only one strategy.*
3. *Users of any version of Agent99 Trainer that provides support for multiple instructional strategies should outperform users of any version of Agent99 Trainer that supports fewer instructional strategies.*

1. Pilot Tests

Two different measures of performance were developed for the training experiments, knowledge tests and veracity judgment tests. Each knowledge test was composed of 12 multiple choice questions taken directly from the content covered in the training curriculum. Since there were three training sessions scheduled, three knowledge tests were created, each based on the respective content from that day's session. The knowledge pre-test and the knowledge post-test for each session were identical, except for the ordering of questions and the ordering of the choices for each question. Each subject's knowledge was measured as their proficiency on each of the knowledge tests, or the number of correctly answered questions from 0 to 12, with 12 being a perfect score.

Six detection accuracy tests were developed. A common measure in deception detection studies, the judgment tests were designed to test the ability of the participants to judge the veracity (truth or untruth) of statements made by an interviewee in a short interview. Each test consisted of six short interviews in three different media (2 text, 2 audio, and 2 video with audio), culled from twenty real interviews in three separate research studies on deception. Furthermore, the interviews in each test were half truthful and half deceptive and a combination of difficulty levels. The interviews were randomly ordered within each test based on media, veracity, and difficulty. Within each pretest-posttest set (12 interviews), each interviewee was unique to avoid results due to communicator (interviewee) specific cues. Subject performance on a judgment task was the number of correct responses, ranging from 0 to 6, with 6 being a perfect score.

The difficulty and equivalency of the six veracity judgment tests were analyzed in a series of two experiments, with 124 management information systems (MIS) upper-division undergraduate students in Introduction to Business Information Systems as participants. The purpose of the experiments was to test the difficulty of the individual items and the equivalency of the compilation of the six tests. The test forms needed to be of equal average difficulty since the tests were to be used to measure changes in deception detection accuracy. In the first pilot experiment (PE1), 96 students completed one of six veracity judgment test forms, with an average of 16 students completing each test form. Participants took approximately fifteen minutes to complete each test form. The students did not have any previous training in deception detection and thus were expected to achieve approximately 50 percent accuracy. Based on the results of PE1 indicating that the difficulty of the six test forms were not equivalent, items were switched between four of the test forms based on the average item scores in PE1. The second pilot

experiment (PE2) was conducted to collect data on the four revised test forms. The participants were students who had not participated in PE1. Each student completed two test forms, with an unrelated task in between. An average of 14 students completed each form. The PE2 data were combined with the PE1 data for the two test forms not revised and re-analyzed, and the analysis indicated that the accuracy rates achieved on the six test forms were statistically equivalent ($p = .825$). See the data for the revised veracity judgment test forms in **Error! Reference source not found.**.

Table 29. Pilot Experiment 2 – Data from Revised Veracity Judgment Tests.

Test Form	N	Mean (Std Dev)	95% Confidence Interval		Min	Max	ANOVA
			Lower	Upper			p-value*
1	15	3.47 (0.83)	3.00	3.93	2	5	0.825
2	15	3.67 (1.18)	3.02	4.32	2	5	
3	12	3.83 (0.94)	3.24	4.43	3	6	
4	16	3.31 (0.79)	2.89	3.74	2	4	
5	13	3.62 (1.39)	2.78	4.45	1	6	
6	15	3.60 (0.91)	3.10	4.10	1	4	
Total	86	3.57 (1.00)	3.36	3.78	1	6	

* significance of difference between the mean scores on the 6 different test forms

2. USAF I Experiment

The first full experimental test of the training curriculum and Agent99 was conducted in fall 2002 at a large US Air Force (USAF) facility located in the U.S. A total of 125 officers participated; the total number participating per session varied. Participants were already assigned to "blocks," or classes by the USAF Air Education and Training Command, made up of sixteen officers, so blocks were randomly assigned to conditions. Training was delivered in three sessions, with each session on a different topic: introduction to deception and its detection, cues to deception, and heuristics that impede detection. All subjects in all treatments received lectures from live instructions in the first and third sessions, but in the second session, one group received a live lecture, while a second group used the Agent99 Trainer, and a third group had a lecture for part of the time and used Agent99 for the rest of the session. All lectures in all treatments were supported with the same PowerPoint presentations and interview examples.

The control group received no training, but control subjects completed the same measurement instruments as the experimental subjects. A pre-session was used to collect baseline data on all subjects in all four groups. The instructors were two USAF officers completing their masters' degrees at the Air Force Institute of Technology and one MIS doctoral student from a U.S. business school. To avoid any potential instructor effect on performance, the instructors did not train the same blocks of subjects more than once, instead rotating to another treatment with each new session. This was done to avoid a condition by instructor confound.

The basic procedures for the training sessions were as follows: Participants reported to their regular block classroom at the USAF facility. They began by completing a battery of instruments, including a knowledge pre-test and the deception detection accuracy pre-test. After the pre-tests, participants were trained for approximately 45 minutes, which was slightly longer than previous training studies (deTurck, Harszlak, Bodhorn & Texter, 1990). The exception was control participants, who were given a break in the interim. At the end of each session, all subjects completed a knowledge post-test, comprised of the same questions as the pre-test but in a different order. They also completed a deception detection accuracy post-test, similar to the pre-test but consisting of different examples. At this point, participants in all conditions received feedback on the correct responses to the deception detection pre- and post-tests.

The knowledge tests can be used as a manipulation check, comparing the control group to the treatment groups, to determine if training was effective in imparting information about deception and its detection. Performance on the knowledge tests was measured by taking the difference between pre-test and post-test scores within each session. Independent *t*-tests showed that the treatment groups differed from the control group for all three sessions. For each session, the control group did not improve, while the training session groups did. Trained individuals, then, did appear to learn about deception and its detection through the training program.

However, for the second training session, where there was variation in delivery methods, there were no differences among the groups that were exposed to traditional lectures, Agent99 Trainer, or the lecture and system combination. Performance on the judgment tests was also measured by taking the difference between pre-test and post-test scores within the session. There were no statistically significant differences between the treatment groups and the control group on deception detection accuracy.

These results provided partial insight into a comparison of e-training and traditional classroom instruction. Subjects who went through the training program on deception detection increased their knowledge about deception and its detection, compared to the control group. That there were no differences between the groups that used Agent99 Trainer and those that did not implies that the e-training delivery mode worked just as well as traditional classroom training, also providing support for Hypothesis 1.

That there was no improvement in deception detection accuracy for the trained groups was the most troubling finding from this study. While accuracy performance improved for all groups for the first training session (the introduction), performance declined for the final two sessions. We found afterwards that the post-test for the cues session was more difficult than intended, even though it had been rigorously tested before the study was conducted. The design of the second study was altered to address some of the issues uncovered in the first study.

3. USAF II Experiment

Major changes made in the study design after the first study was completed included: 1) dropping the heuristics content in order to simplify the second study by having only introductory and cues content; 2) increasing the number of items in the judgment tests from 6 to 10 and decreasing the number of items in the knowledge tests from 12 to 10; and 3) using videotaped

lectures instead of live instructors in order to minimize variance in content presentation across subjects.

The primary purpose of the second study was to help determine whether some configurations of e-training systems were better than others, and to test the three hypotheses listed previously. The second study was conducted at the same USAF base as the first study in the fall of 2003. A total of 177 officers participated. They attended two separate training sessions, one covering an introduction to deception and its detection, the other covering specific cues that have been demonstrated to be effective indicators of the presence of deception (see DePaulo, et al., 2003, for examples). For each group, the second training session was held five days after the first.

Participants were randomly assigned to either the control group, which featured a videotaped lecture on each topic, or to one of four treatments that featured a different version of the Agent99 Trainer system. The control group viewed a professionally videotaped and edited lecture. The other groups used one of four configurations of the Agent99 Trainer. The four different configurations of Agent99 were configured as follows:

Linear Agent99: This version of Agent99 Trainer included the same lecture video, Powerpoint presentation, and examples as the video lecture used in the control. However, the material was displayed in the synchronized multimedia interface of the Agent99 Trainer. Users could only access the material in a linear manner, in the same order in which the lecture had been organized. Only support for multimodal delivery of instruction was provided.

Agent99 + Ask-A-Question: This version of Agent99 Trainer allowed users to jump to any topic listed in the index, allowing them to move though the training material at their own pace, governed by their own interests and priorities. This added support for the self-directed instructional strategy. This version also added the Ask-A-Question feature. Ask-A-Question allowed users to enter a question about the content in a natural language format. The system response lists locations in the content where more information about the topic of the question can be found. Adding Ask-A-Question provided support for the learner-instructor interaction instructional strategy. We put these two strategies in one configuration because we believed that together they explained how learners can interact in the Agent99 Trainer system: interact with learning materials or interact with instructor (virtually).

Agent99 + Ask-A-Question + More Examples/Multimedia Cases: This version of Agent99 Trainer was exactly like the former version except that one additional feature was added: More examples of cues to deception than were included in the prior versions or the video lecture. This added more support for the fourth instructional strategy: practice and feedback of skills.

Agent99 + Ask-A-Question + More Examples/Multimedia Cases + Quizzes: This version added quizzes designed to test the user's comprehension of what he or she had been exposed to thus far. Quizzes provided support for the fifth instructional strategy, practice and feedback for knowledge. The quizzes appeared intermittently throughout the lesson and had

to be answered before the student could proceed. Students received immediate feedback about whether or not their answer was correct.

Experimental procedures were similar to those used in the first USAF-based study. One major difference had to do with the number of items in the knowledge tests – decreased from 12 to 10 – and the number of items in the veracity judgment tests – increased from 6 to 10. After completing the pre-tests, participants were exposed to the video lecture or to one of the Agent99 Trainer configurations. After instruction, participants completed knowledge and judgment post-tests, identical in format to the pre-tests. The knowledge post-tests included the same items as the pre-test but in a different order. The judgment post-tests were made up of 10 items totally different from the pre-tests.

To understand the relationship between the different combinations of system functions and training effectiveness, we conducted a planned comparison analysis using Helmert contrasts. Helmert contrasts test whether treatments with additive features are differentially effective by comparing each level or group with the average of the more sophisticated remaining levels (Stevens, 1986). These comparisons are meaningful when the effectiveness of a combination of treatments is being evaluated. Such is the case with the different packages of training delivery used in here in Study Two, as each treatment packaged an additional software feature to its previous instantiation of Agent99 Trainer (A99). Contrast 1 compared the video lecture condition with all the other conditions, allowing a test of Hypothesis 1. Contrast 2 compared the linear A99 group with the average of the other A99 conditions that have more functionality, allowing a test of Hypothesis 2. Contrast 3 compared the A99+AAQ with the other A99 groups with more content and quizzes, providing for a test of Hypothesis 3. We expected that versions of the Trainer with more content would result in better performance than versions with less content. Finally, Contrast 4 tested whether the addition of quizzes in A99 can produce better training effectiveness than the less-equipped versions of the software, providing an additional test of Hypothesis 3.

In the introduction lecture session, there was significant improvement for both knowledge and judgment performance following the treatments. The within-subjects comparison for the knowledge tests and the judgment tests showed that the training was effective for everyone in the first lecture. The different configurations of training software did not have any effect on knowledge test performance. For the judgment tests, the software treatment did make a significant difference. The participants using linear Agent99 were outperformed by those using Agent99 with more features (conditions 3, 4 and 5), supporting Hypothesis 2. Also, subjects using the version of Agent99 with AAQ, more content, and quizzes outperformed those who used the version with AAQ and more content but no quizzes, providing some support for Hypothesis 3. There were no differences in performance when comparing conditions 4 and 5 to condition 3. Taken together, these results also support Hypothesis 1, as users of the Agent99 Trainer performed as least as well as students who received only the video lecture.

In the second session, the cues lecture, there was a similar improvement between pre-test and post-test performance for both knowledge tests and judgment tests overall. Another similarity between the introduction and cues lecture was that the software configurations had no effect on knowledge test performance, but this time there was not a software-related effect for judgment

test performance either. There was, however, an interaction between the time of performance observation and software condition for knowledge test performance. The post-test improvement for the video lecture subjects was minimal, the subjects in condition 4 (the package of A99, Ask-A-Question, and extra content) performed the same on the post-test as the pre-test, and subjects in the other three software conditions performed better on the knowledge post-test. Taken together, these findings provide support for Hypothesis 1 but not for the other hypotheses.

To summarize the findings from the second study, all groups, whether exposed to the video lecture or to one of the four configurations of Agent99 Trainer, increased their knowledge about deception and its detection, and they improved their ability to detect deception on the judgment tests. The e-training effort was successful for helping students learn more about deception and its detection and for helping them use their new knowledge to better detect deception in real-life examples. Hypothesis 1 was supported for both outcome measures for both the introductory and cues lectures. In addition, for the introductory session, students using the three versions of Agent99 Trainer that supported multiple instructional strategies all outperformed users of the linear Agent99 Trainer configuration. As for judgment tests, users of the Agent99 Trainer configuration with the most features outperformed users of the Agent99 Trainer that was similarly configured but did not include quizzes. Hypotheses 2 and 3 were partially supported, then.

For comparisons across the entire training period, instead of across a single training session, all students improved, based on their performance on the knowledge and judgment tests. For knowledge, users of Agent99 Trainer with the most features outperformed those who used the similar configuration but lacking quizzes, providing additional support for Hypothesis 3. For judgment, students who used the Agent99 Trainer outperformed those who were exposed to the video lecture, again supporting Hypothesis 1.

4. FSU Experimental Test of Agent99

A laboratory experiment was conducted to test several hypotheses related to deception detection and media, warnings and training. The specific hypothesis related to training was:

> *Deception detection accuracy will be greater for receivers who are trained in deception cue recognition than for those who are not trained.*

The study was designed as a 2 X 2 X 2 factorial, with conditions of induced suspicion (warning or no warning), training (either training or no training), and two types of media, specifically lean (e-mail) or rich (audio over Internet chat relay). This experiment required students to enhance their resumes and defend those enhancements when communicating with an interviewer via either lean or rich electronic media. Subjects in the training cells attended a deception-cues training session one week prior to their scheduled experiment date. The training materials consisted of 20 minutes of the video lecture created for the Keesler training experiments. The applicant was asked to do whatever it took to look like the best student for the purpose of setting standards for a scholarship. The applicants were told that during the interview they should be as convincing as possible in defending the information in the enhanced resume. The scholarship application was sent to the receiver via Microsoft NetMeeting. Applicants were interviewed remotely by another subject, either by e-mail or voice-over-IP. The interviewer asked questions

of his or her choice for up to 20 minutes. Before and after the interview, the subjects completed questionnaires for data collection.

The hypothesis regarding training was supported: Subjects who had been trained were better able to detect deception than were subjects who had not been trained. Coupled with the preceding findings, the results offer support for the value of training generally and for computer-based training specifically.

D. Usability Test, UA

To test whether the Agent99 Trainer improves deception detection accuracy and whether the performance of Web-based training system was better than performance under lecture-based training, we conducted usability tests at the University of Arizona. A pretest-posttest comparison was conducted between two treatment groups: Lecture group and Agent99 group. Training using Agent99 Trainer significantly improved the detection accuracy from the pretest to the post test and produced somewhat better (though not statistically different) detection accuracy than the lecture group (Cao, et. al, 2003).

To test the subjective effectiveness of Agent99 Trainer, participants completed a questionnaire after the posttest judgment test. Only the subjects in the Agent99 group were asked to answer questions related to the usability test. The results are shown in Table 30 (where means are based on a 1 = strongly agree to 5 = strongly disagree rating scale).

Table 30. Participants Responses in the Usability Test (Questionnaire).

Questions	Mean
1. The overall training content is interesting to me.	2.33
2. The video/audio quality of the lecture is satisfactory.	2.83
3. It is easy to learn how to use the system.	1.83
4. During the learning process, I think that accessing of various parts of the system or navigating through the system is easy.	2.33
5. The structured and synchronized multimedia content provides aid in my understanding of the subject matter.	2.11
6. I enjoy the self-paced control I have in the selection of what I want to access in the learning process (be capable of watching any part of the lecture and any example at any time).	1.78
7. The View Example and Expert Analysis module helps me better understand the content of the lecture.	1.67
8. The knowledge I learn from the lecture(s) helps me analyze the examples I view.	1.65
9. Completing the training make me feel more confident in my ability to accurately detect deception.	2.28
10. I am enthusiastic/genuinely interested in utilizing this format of learning again.	2.41

The results were highly positive, justifying our system design from a subjective view. The numbers indicate that the Agent99 Trainer system was interesting, easy to use, the structure and synchronization of multimedia contents and self-based learner control was helpful (question 5 and 6), and more importantly, the method of "view examples with expert analysis" and the association of explicit instructions (lecture) with practice (examples) helped the learning of deception detection (question 7 and 8).

In sum, the tests of Agent99 Trainer and associated deception training curriculum were quite supportive of their utility in improving knowledge of deception and applying that knowledge in realistic judgment tasks. The Trainer itself has many features that recommend it for use in domains other than deception.

XI. Lessons Learned

A five-year project naturally produces voluminous findings and insights. The following sections describe the most important lessons learned from the project.

A. Conclusions

Our research has created an enhanced framework for understanding deception and its detection. This has been applied with great success in our audio and video deception detection research efforts.

We have continued to revise interpersonal deception theory (IDT) to produce a much more robust model that has high applicability to deception over electronic communication. Our taxonomy of deception indicators has provided valuable direction in selecting indicators to analyze, and our 27 experimental studies, with nearly 3400 subjects, have confirmed a host of reliable text-based and nonverbal deception indicators. Our desert survival experiments uncovered several text-based cues differentiating truthful and deceptive senders and cues differentiating deceptive senders and truthful receivers.

Text-based indicators have related to message length, syntactic and lexical complexity, lexical diversity, specificity, certainty, immediacy, affect, and dominance. Nonverbal indicators have related to arousal, expressiveness, and dominance. We have found that neural networks, multiple discriminant analysis, and Bayesian analysis offer promise for identifying clusters of cues that accurately predict truth or deception. These lessons learned are now being applied in our textual analyses of witness statements at two pilot Air Force security squadrons.

Additionally, we have confirmed the influence of several cognitive heuristics and biases on detection accuracy. Visual bias and truth bias are among the two that reduce detection accuracy. These findings highlight major considerations in automating deception detection.

The following items represent the fundamental knowledge gains resulting from the research conducted over the course of this project:

1. Computer Tools Can Assist Users in Detection

Software tools were built for detecting deception through linguistics, vocalics and kinesics. In all cases, the detection of deception was improved through the use of computer-aided systems. The use of software tools increased the accuracy of detection to as high as 80-90% accuracy.

2. Biases Exist

Everyone has biases, even experts lean toward truth or deception. Many people are not even aware of the biases that they bring to a given context. Training can make an individual aware of their biases and help them overcome the effects of these predispositions and improve their detection accuracy.

3. No Single Cue Is Sufficient for Detection

Everyone is looking for the "silver bullet" of deception detection, however no single cue is adequate when used alone. Regardless of the modality, multiple cues (taken together as a set) are better predictors of truth or deception. Someone who is an expert in deception can control a few cues, but it is nearly impossible (under stressful conditions) to control seven or eight cues. Some cues will eventually "leak out" no matter how hard a suspect tries to control them.

4. Context Must Be Considered

The communication modality employed and the interview environment both have an effect on which cues are the most salient in detecting deception. Thus, successful detection requires customizing the cues utilized based upon the particular interview context. For example, a checkpoint at an airport is a different environment than a bank, and security personnel should be aware of the differences in expected behavior between the two.

5. Culture Must Be Considered

The cues that are utilized to detect deceptive communication in one culture may not work well for another. Different cultures express themselves in very different ways – linguistically and kinesically. For example, a non-verbal gesture that is considered to be innocuous by the members of one culture may be considered to be highly offensive to the members of another. Thus, the cultural background of an individual must be considered when attempting to identify the meaning of a particular phrase or gesture that they use.

6. Ground Truth Is Difficult to Obtain

Ground truth is absolutely necessary for testing and developing deception detection algorithms, however many sources of ground truth in criminal investigations (e.g., indictments, convictions, etc.) are imperfect and even courtroom verdicts may be incorrect. Often, researchers cannot obtain the video and audio because of legal considerations and privacy concerns, which makes the determination of ground truth more difficult. For example, there is no feedback mechanism on people crossing the border – it is not a "closed-loop" system of reporting. The only way to know if a person who has crossed the border was intending to do something malicious is if they are later caught and/or apprehended after the fact.

7. More Data Is Better

The algorithms and techniques for deception detection can be improved with larger volumes of data, as the increased volume significantly helps improve data mining efforts. Larger data sets provide the opportunity to better train computer models, which will ultimately improve their predictive accuracy. Again, the legal and privacy concerns often restrict the amount of data that can be made available.

8. The Multi-Disciplinary Approach Is Valuable

Contributions of multiple disciplines help shed light on the complex problem of deception detection, as one discipline informs another. An engineer looks at a problem from a systems perspective, while the psychologist may see the same problem from a cognitive perspective. The different scientific methods employed by both researchers helps guide the development efforts of the other, making the resultant artifact more robust and accurate.

9. Research Methods Should Be Theory-Driven

Software, lab experiments and field studies require solid theoretical foundation. Without theory, we don't know what we're observing and are less likely to draw relevant conclusions from those observations. Thus, theory should drive the methods that are used to produce and analyze the results.

10. Both Laboratory & Field Testing Are Necessary

A combination of controlled lab experiments and naturalistic field observations are essential in deception detection research. In our experience, combinations of these two methods have helped deliver more accurate and robust results.

B. Next Steps

The following items summarize the suggested extensions of the research conducted during this project, and in all cases additional experimental and field data collection is suggested:

1. Create Test Beds for Continued Annotation and Analysis

The data collected throughout this research project will be valuable to other researchers who seek to build upon our existing research efforts. These data sets provide myriad opportunities for future research in many disciplines. Although the time required to prepare these test beds for use by others is considerable, we are committed to making them available to potential research partners.

To illustrate the time needed to prepare these data sets, consider that each hour of video collected requires about 4 hours of preparation. This process includes the following steps (along with an average time to complete each):

1. Import the recorded video into the AVID Xpress editing suite (60 mins.)
2. Apply time codes to the imported media for clip identification/ extraction (15 mins.)
3. Manually code the video to identify unnecessary/irrelevant material (120 mins.)
4. Manually edit the video to remove unnecessary material (30 mins.)
5. Apply additional post-editing time codes for ease of analysis (15 mins.)

In the near future, we hope to make each of the following sets of data available to researchers:

1. Desert Survival 1 & 2 (text only)
2. StrikeCom 1 & 2 (text, audio, video, self-reports)
3. Mock Theft (text, audio, video, self-reports)
4. Deceptive Interviews (text, video)
5. Resume faking (text, audio)
6. Cross-cultural deceptive interviews pilot (video)
7. Cheating experiment 1 & 2
8. Pedestrian border crossings
9. Visa interviews
10. CBP secondary screenings
11. Counter-Crime Consortium criminal interviews

2. Deception Detection Tool Development

The next step is to continue the development of a system (capable of processing both video and audio) that operates in near-real time for the purpose of streamlining the process of deception detection. The tool should incorporate additional feature sets that approximate coding semantic level units automatically, such as:

1. Video features
   a. Identifying specific gestures, rather than recognition of low-level kinemes
   b. Multiple person tracking in single video frame
2. Audio features
   a. Perceptual approximations (voice quality)
   b. Specific non-fluencies
   c. Enhanced turn-taking tracking
   d. Multiple voice segmentation in a single audio channel

As part of the Agent99 suite of tools, we continue to work on developing a system for extracting, analyzing, and fusing vocalic, kinesic, and linguistic features. Figure 29 below graphically depicts how the software is intended to work.

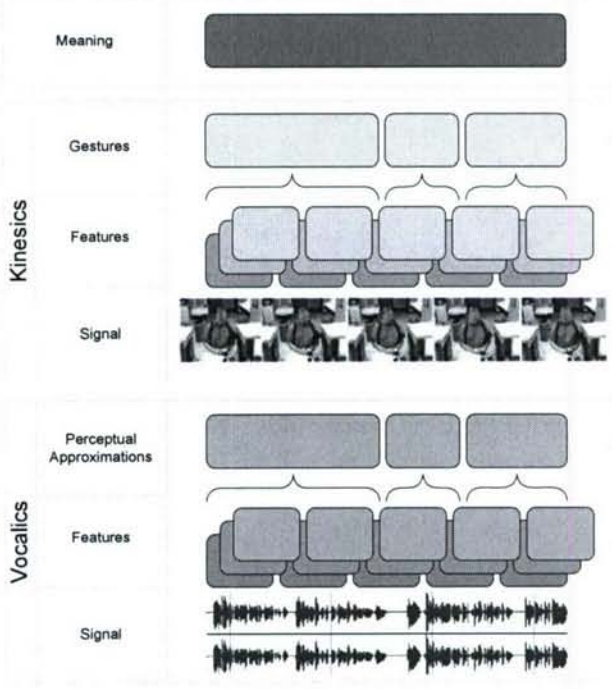


Figure 29. Proposed Agent99 Tool Schematic

The application is designed to extract kinesic and vocalic features from a multimedia stream of data. As shown in the diagram, the video signal is systematically broken down into a series of kinesic features, which can be grouped into related sets that approximate an individual's gestures, and these gestures, taken together can be used to determine the individual's intended

meaning. Likewise, the audio signal is analyzed into vocalic features which approximate phonemes and their related emotive qualities (e.g., emphasis, stress, or accentuation). These features are then taken together and grouped into perceptual approximations that reflect the components of their speech, helping identify the content and patterns of speech used throughout an interview which can then be used to infer the subject's intended meaning.

With additional research and development, the software ultimately is designed to work autonomously and in near-real time. However, the underpinning technologies are currently insufficient to operate independently, thus in the interim, the software requires human-coded data and interpretations to identify the features of interest to be extracted. A Behavioral Analysis System (BAS) is envisioned that will provide visualization capabilities in multiple modalities in addition to capturing and annotating behavior. A sample screenshot of the BAS interface is shown in Figure 30.

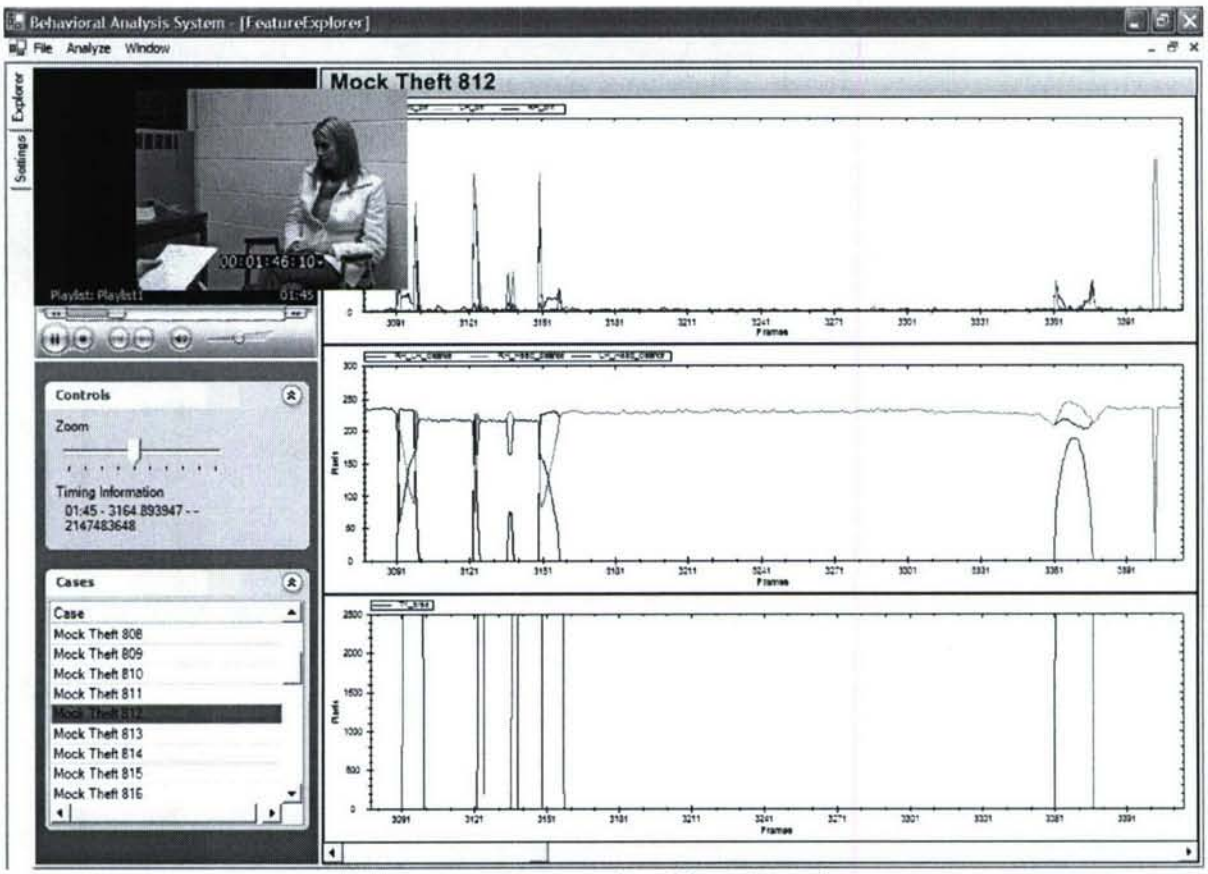


Figure 30. Screenshot of the BAS Interface

As the above screenshot illustrates, the human coder's observations of an individual's behavior (which are keyed in manually) can be represented in a graphical form – displaying the kinesic (and even the vocalic) features of the individual subject throughout a videotaped interview. The data are then processed by the Agent99 Analyzer application as it attempts to determine the meaning of an individual's kinesic and vocalic behaviors – ultimately augmenting the human user's ability to ascertain the truth or deception present in the individual's interview responses.

3. Experimental and Field Data Collection

The studies conducted over the course of this research project provided significant contributions to the body of knowledge related to deception detection and helped guide our efforts toward developing designs for how the process should be automated. However, these preliminary studies represent only the first steps in understanding the problem and developing automated solutions. We suggest that the following data collection efforts be undertaken to further research the problem, as the results of these studies will likely be critical in the continuing development of a practical solution that can be implemented in the future.

a) MSU Cheating Study (in progress)

In the cheating study's pilot test, participants were asked to play a trivia game with a confederate, who tried to induce cheating while the game master is out of the room. At the end of the game, all participants were interviewed about their "game strategy" and whether cheating occurred. These interviews were videotaped for subsequent analysis and increased scrutiny. But in the replication currently underway, the interactions are designed to be longer and the subjects are given larger monetary inducements to cheat. It is hypothesized that the increased opportunities for deception (i.e., the longer list of interview questions and the greater amount of time to provide convincing answers that may be deceptive), coupled with the increased motivational stimulus, will provide a better corpus of data to analyze. The improved quality and quantity of data collected will help us better identify the most salient cues that are indicative of deception in similar situations and foster the development of the automated tools to detect these cues. The net result of this laboratory experiment will be a collection of results that is more generalizable to real-world scenarios, which will improve the predictive power of our proposed software tools under development.

b) Field Studies of Customs & Border Protection (in progress)

Despite our close proximity to the CBP's border operations, this data collection effort has proven to be highly labor-intensive. The field studies that we have been conducting at the Nogales, Mexico, border crossing station have yielded some very promising preliminary results, but it takes a considerable amount of time to collect a sufficient number of videotaped examples for a more thorough analysis. We have to record a large number of subjects in order to get a handful of data points of interest – "consented" videos, those approved for release/analysis by the interviewed subject.

To illustrate the time-intensive nature of this project consider that (to date) the field studies of the Customs & Border Protection facilities required over 230 hours of labor to collect the 177 hours of videotaped recordings that have been generated. But of this total, only 25% of the data has been approved for use by our researchers: 10.5 hours of pedestrian crossing (captured in 21 half-hour shifts); 18 hours of "permit counter" interviews (144 subjects); and 45 hours of expedited removal interviews (of which only 33 were "consented"). In fact, our researchers have driven over 3700 miles (nearly 56 hours) thus far in order to supervise these data collection efforts. Thus, a significant amount of time is required to improve the robustness of this corpus of data and insure its appropriateness for further scientific scrutiny.

c) Deceptive Interviews (new experiment)

Our research to date has focused on "seated" interview scenarios, and the kinesic analysis of the videotaped recordings of these interactions has generally focused on the behaviors of the interviewee's upper torso (including arms/hands, head movement, and postural changes). However, the ever-growing body of research in the field of deception suggests that some salient cues may exist in the lower limbs (legs and feet). Also, there is some evidence to suggest that certain interviewing styles may result in reciprocal behaviors (where the interviewee strategically mimics the behavior of the interviewer) that may confound the positive identification of some deception cues. Additionally, the vast majority of subjects that we have analyzed to date are North American natives. Thus, we have designed a laboratory experiment that will allow us to investigate the full-body kinesic behaviors of a wider variety of participants from a broader cross-section of cultural backgrounds across different interviewing styles.

The first replication (pilot study) of this proposed experiment requires the interviewees to alternate between providing truthful and deceptive responses across a series of 12 questions, in which the order of responses varied (truth-first or deception first). The interviewers in these interactions will be naïve to interviewee behavior (i.e., the interviewer will be blind to the truth condition provided to the interviewee, so as not to evoke any particular behavior) and all interviews videotaped for verbal, vocal & visual analysis. The visual analysis will be facilitated by four camera views of the same interaction – a facial close-up (to capture more minute facial expressions), an upper torso view (similar to past experiments), a full-body view (to capture the movements of the interviewee's lower limbs, and a wide view of both the interviewer and interviewee (to capture any reciprocal behaviors that may be employed).

The second replication of the study will require alterations in the interviewer's questioning style and begin to investigate the cultural variability in kinesic behaviors. We will conduct interviews with bilingual subjects from 3 different cultures, asking them to provide truthful and deceptive responses in two languages (some in their native language and some in English). Additional replications will feature a larger sample of cultures, and incorporate the addition of biometrics (pupillometry, gaze and eye blink behaviors, etc.). We plan to study 600 participants throughout the course of this experiment, and each will be interviewed for approximately 20 minutes. This will result in 300 hours of videotaped interactions (which can be viewed in any of the four camera angles).

It is hypothesized that this will be the most feature-rich videotaped deception-based data set in existence. This corpus will significantly help us determine the appropriate direction for future studies and guide in the development of the automated software application's "fusion engine," which will provide appropriate weights and metrics to be used by the software in determining the truth or deception in a given respondent's answers.

d) Consulate Visa Interviews (new field experiment)

Another suggested field study involves recording interviews in a different context than our prior work: Visa interviews conducted at the U.S. Consulate in Nogales, Mexico. These interviews will primarily feature Mexican citizens requesting temporary U.S. visas that will allow them to enter the country to work, shop, etc. It is important to note that these interactions are not

conducted at the busy border crossing station, rather they are conducted in the less-busy, more quiet and orderly setting of the consulate offices.

Like the border crossing interviews, these interactions will be brief (typically 1½ to 5 minutes) and feature standing participants. However, these interviews differ from the existing border interactions in that they are conducted in consolidated "blocks" each morning and they require internal approval to collect data on each subject. We plan to collect video of approximately 500 interviews (or about 20 hours worth of interactions), and this data could be cross-validated with the Customs and Border Protection permit counter interviews to improve our analyses and interpretations of the results.

e) Computer-Aided Decision-Making (new field experiment)

Our last suggested experiment will investigate the degree to which human users will accept the output of a software application that is designed to augment their decision-making process in determining whether an individual is being deceptive. This is critically important in the successful transitioning of the technology that is developed as a result of this research project – we need to develop an application that provides the human user with accurate, objective information that actually helps them achieve their goal. However, if the human does not trust the application's suggestions (i.e., its automated interpretations of the interaction), the technology transition will ultimately fail, as the perceived value of the application will be greatly reduced.

It is widely accepted that machines are best suited to analyze discrete, micro-level cues such as the number of modal verbs used in speech, the fundamental frequency of the interviewee's voice, the velocity of gestures and hand movements, etc. Alternatively, humans are best suited to analyze general, macro-level cues such as general tension throughout the interview, the interviewee's level of involvement and overall cooperativeness. It is the goal of this exercise to determine the appropriate combination of these capabilities into an effective human-computer system of deception detection.

Toward this end, we have designed experimental treatments that investigate an individual's level of comfort with an intelligent agent (system use), dependent upon their current level of detection training and expertise. We will vary these factors in both lab and field experiments to see how they affect the human user's judgment performance, accuracy, confidence, decision-making strategies, level of effort exerted, and their trust in the system. Figure 31 graphically depicts the relationships between these factors of interest in this suggested experiment.

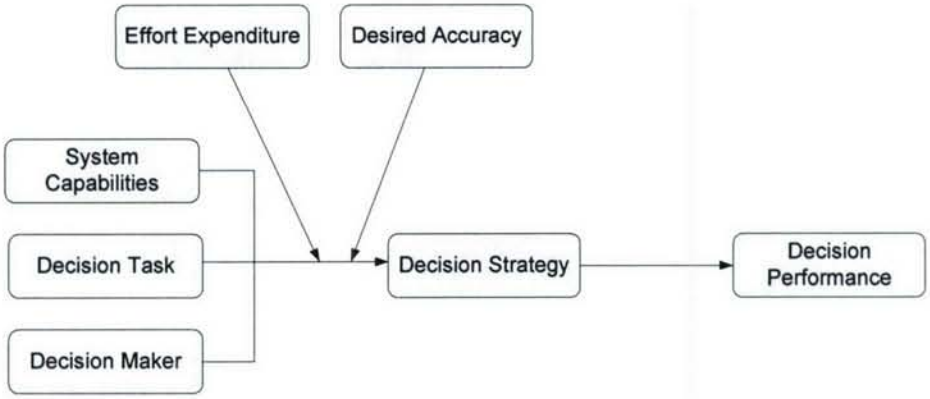


Figure 31. Decision-Making Experiment Design

XII. Transitions

Technological innovations developed through this program continue to migrate into other venues and organizations. StrikeCom has been used in the Office of Secretary of Defense's Network Centric Warfare workshops throughout the world. Over the course of this research project, it has been used as a key teaching tool in the U.S. and NATO workshops in Portugal, the Netherlands, and Germany. Currently, the tool is being upgraded to ease setup requirements and lower operational overhead. It is thought that this will allow StrikeCom to be used with distance learning classes throughout the world. This will significantly enhance the ability of OSD to lower costs and broaden participation in these workshops.

We have made considerable progress in integrating video and audio deception detection into the program. This has opened new doors for transition. Currently, we are working to build and test the Agent99 textual analysis tools with Air Force security squadrons in Arizona and Oklahoma. Ultimately, we hope to field a functioning prototype to these units that can be used to analyze witness statements in near real time. We will also collaborate with other researchers to extend theory and practice in the area of deception detection.

A. AGENT99 Suite

Components of the Agent99 Suite have already been implemented for research and training purposes. The tools for text analysis—Agent99 Parser, Client and Analyzer—and for nonverbal analysis—C-BAS and AutoID—are described below.

1. Agent99 Parser, Client and Analyzer

We have used the Agent99 Suite to conduct all our text analysis and have made it available for other researchers to analyze text. We have integrated the part-of-speech parser with GATE (General Architecture for Text Extraction) and Weka (the back-end processing and classifier tools) so that we can batch-process XML-formatted text files.

We have continued to extract a number of low-level features such as average sentence length, number of words, and other computed values (e.g., emotiveness) as well as to calculate higher

level features such as speech acts (e.g., non-opinion statements, backchannels) that may be indicative of states like uncertainty. For the more complex lexical analysis the Agent99 Analyzer tool uses machine learning techniques such as decision trees and neural networks to analyze text. Once again, we have been able to harness free and open source software to meet our ends. The GATE toolset performs all of the tasks previously accomplished by the Grok software and a few more. It is flexible and allows the addition of new lexical and grammatical bases of evaluation.

After performing the GATE analysis on the data, the Agent99 Client uses Weka, yet another open source software application, to perform advanced statistical and machine learning analyses. Weka can easily be set up to execute any number of algorithms for machine learning and solving real-world data mining problems through the use of data pre-processing, classification, regression, clustering, association rules and visualization.

The extensive use of GATE and Weka has freed our research group from building our own analysis tools and has afforded us the opportunity to instead focus on identifying and detecting reliable deception cues. This year we have compared the sensitivity and specificity of discriminant analysis, neural networks, decision trees, and support vector machines to determine deception from an original segment. We are now exploring hidden Markov models as a potentially superior method for achieving greatest sensitivity and specificity.

2. C-BAS

C-BAS, our Behavioral Annotation System written in C#, has been used extensively in our program of research to annotate frequencies, durations, and ratings of observed behaviors, which can include both verbal and nonverbal features. We have modified the system so that coders can use a standard keyboard and computer (unlike proprietary tools that require a special keyboard). The interface includes a video display frame with controller, a frame displaying the template of assigned keys, and a frame that displays the time-synced key presses as they are made. Multiple behaviors can be interleaved in the same file. This tool has been transitioned to other universities, a demonstration of it at a European conference attracted significant attention, and it will be demonstrated (as a featured, invited tool) at the International Society for Gesture Studies, June 2007. We have continued to refine it so that it has maximum flexibility for users and could become a tool for an intelligence analyst to add commentary to analyses of multimodal intelligence.

3. AutoID Behavioral Analysis System

The manual approach still remains extraordinarily time-consuming, and human coding cannot escape some degree of subjectivity. The AutoID tool offers the potential to replace human coding for observations that can be totally objective and to identify complex patterns that go unrecognized by humans. Researchers at the University of Arizona and Rutgers University are developing a knowledge-based system which analyzes kinesic and linguistic behavior in search of deceptive cues. The system, known as the behavioral analysis system (BAS), analyzes the movements and linguistic properties of communication from one person engaged in a recorded face-to-face interaction. The BAS tracks the head and hands as they move throughout a recorded segment and analyzes linguistic characteristics from a transcript of the interaction and calculates features that give insight into whether or not the observed person is being deceitful.

a) Kinesics

The BAS utilizes a tracking method developed by Computational Biomedicine Imaging and Modeling Center (CBIM) at Rutgers University (Lu, Tsechpenakis, Metaxas, Jensen, & Kruse, 2005). The method extracts hand and face regions using the color distribution from a digital image sequence. A three-dimensional look-up-table (3-D LUT) is prepared to set the color distribution of the face and hands. This 3-D LUT is created in advance of any tracking using skin color samples. After extracting the hand and face regions from an image sequence, the system computes elliptical "blobs" identifying candidates for the face and hands. The 3-D LUT may incorrectly identify candidate regions which are similar to skin color, however, these candidates are disregarded through fine segmentation and comparing the subspaces of the face and hand candidates. Thus, the most face-like and hand-like regions in a video sequence are identified. From the blobs, the left hand, right hand and face can be tracked continuously. A complete technical description of the kinesics portion of the BAS system is beyond the scope of this study, however, the interested reader is directed to (Lu, Tsechpenakis, Metaxas, Jensen, & Kruse, 2005; Meservy et al., 2005).

b) Linguistics

The BAS also is capable of analyzing linguistic features of interactions. These features are derived from transcripts of each interaction and are created using a method called message feature mining. Message feature mining (Adkins, Twitchell, Burgoon, & Nunamaker Jr., 2004) is a method for classifying messages as deceptive or truthful based on content-independent message features.

The reliability of the BAS (both the kinesic and the linguistic components) is currently being explored. Various experiments have shown that reliability rates vary between 60-90% (Burgoon, Jensen, Kruse, Meservy, & Jay F. Nunamaker, forthcoming) and field tests are currently ongoing. The variation in reliability may be the result of a host of influential factors including: environmental constraints, BAS's ability to track human movement, variation of human behavior during various interactions, motivation of liars to succeed, and possible consequences if the liar is caught. Therefore, researchers expect the reliability of the BAS to change as it is used in different environments (Swets, 1986).

In each location where the BAS may be used, careful calibration must be completed. The calibration should follow the steps of signal detection theory (SDT) in diagnostic decision-making (Swets, 2000). First, behaviors that are most closely associated with deception in the new location should be identified. This step is guided by existing work in deception detection. Second, a proper threshold must be determined which will balance hits and false positives. For example in deception detection, a strict threshold would only classify those who exhibit a large amount of behavior associated with deception as deceptive. In determining this threshold, all costs (e.g., time, resources, legal consequences, etc.) of misdiagnosis should be considered.

B. Experimental Interface

To test the usefulness of the BAS in judgments, an experimental interface has been developed and is shown in Figure 32. The interface consists of a series of screens and forms that present the BAS output in a logical way. The interface provides explanations about the BAS in natural language and will also provide help if the users have additional questions about how the BAS

operates. The interface is designed specifically to capture and record all the explanations that the user accesses in formulating a decision.

Experiments are being conducted to determine how users respond to information delivered in this manner and their reliance on system-returned information versus their own judgment.

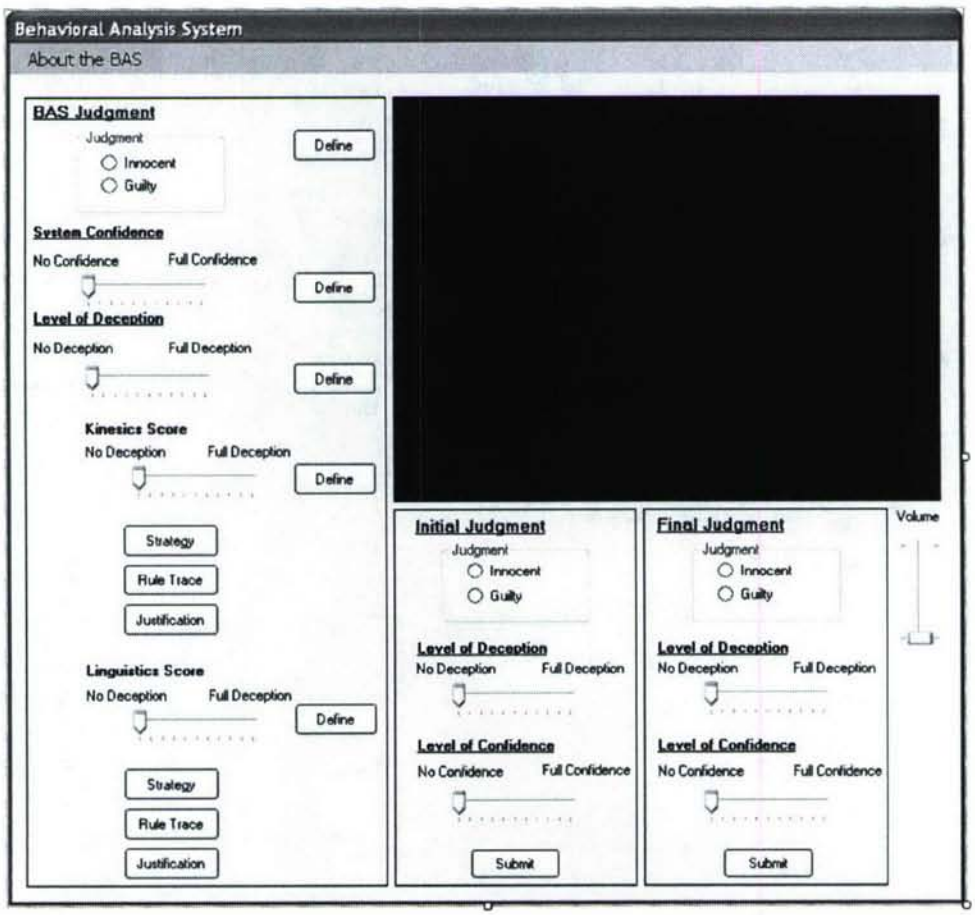


Figure 32. Sample BAS interface.

C. Fusion of indicators

A primary goal of our research has been to develop a software application that helps augment the deception detection abilities of military, security, and law enforcement personnel. Creating this application involves two different software development tasks: Designing a "back-end" system that analyzes the various streams of data for deterministic (predictive) cues of deception and incorporates them into an actionable suggested course of action (otherwise known as the "fusion engine"), and a "front-end" graphical user interface (GUI) to display representations of the software's various levels of analysis and its recommended course of action.

Unfortunately, the results of our research to date are insufficient to begin work on a fusion engine. This is because no set of deception cues has proven to be reliable enough to be used in a

majority of situations–the high degree of variability in deception cues observed across modalities, context, culture, and so on make the development of a universal fusion engine impractical at this time.

However, we have made progress in the development of the software's front-end GUI, and this preliminary research artifact will ultimately help guide the development of the fusion engine and (ideally) promote higher levels of usability of (and satisfaction in) the final software deliverable.

To date, we have relied upon two well-established software development techniques to guide our GUI development process: User-Centered Design (UCD) (see Figure 33) and iterative prototyping.

User Centered-Design (UCD) is both a philosophy and a process. It is a philosophy that places the user at the center as opposed to the product. UCD as a process focuses on factors such as perception, memory, learning and other cognitive factors, as they come into play during peoples' interactions with things. The process seeks to answer questions about users and their tasks and goals, and then use the findings to drive development and design.

Figure 33. The User-Centered Design Model Framework.

UCD seeks to answer questions such as the following:

1. Who are the users of the product?
2. What are the tasks the users will perform and their goals?
3. What is the users' experience level with this product and others like it?
4. What functions do the users need from this product?
5. What information do users need, and in what form do they need it in?
6. How do users think this product should work?
7. How can the design of this product facilitate the user's cognitive process?

To this end, our preliminary research into the GUI development yielded the following observations (arranged according to the conceptual categories prescribed by the UCD model):

TARGET USER – The target user group is defined as a Transportation Hub Security Screener. We have stereotyped our prototypical user as an individual with the following characteristics:

- 35-year-old male
- Little or no college (community college or trade school)
- "Task-oriented" nature
- Developed "people skills"
- "Resilient" personality
- Basic computer skills

SOCIAL ISSUES – The social issues in this context are of great concern, in that many personal freedoms are often infringed upon in the guise of public safety. Intensive training will be the key factor to keep the Transportation Screeners from engaging in stereotypical screening, and prejudicial harassment.

- Public safety (Threat of terrorism)
- Impact of stereotypes & biases on judgment
- Trade-off: Security vs. personal privacy

ORGANIZATIONAL ISSUES – Oversights are major concerns. In the aftermath of the 9-11 tragedy Congress acted swiftly to require certain standards for screeners, both in intellect and experience. As with any Governmental oversight, deadlines and requirements are not always funded, or managed to the extent needed. Thus, requirements for training and implementation could be a moving target.

- Governmental oversight (via DoD, DHS, etc.)
- No profit motive!
- Deep pockets for tech investment
- Operations/processes not mature
- Specialized job roles
- Formal training provided

TECHNOLOGY FACTORS – The technology must not interfere with the process, instead be a tool to increase the opportunity of discovering deception, while speeding the processes for the majority of people that will pass through the screeners. The equipment will support touch screen, non-evasive, easily recognizable cognitive readouts that will aid users at all levels of expertise.

- User interface & interaction
- Non-invasive, non-interactive
- Input devices
- Audio/video surveillance equipment
- Touch screens
- Use of color/icons/graphics
- Color-coded read-outs reduce cognitive load
- User support materials

- Comprehensive! (In-line help tools required)
- Minimal physical limitations
- Can disabled persons perform duties?
- Low experience level
- Employee churn?
- Reactive nature of threat detection?
- Low enjoyment/satisfaction
- Average traveler demeanor?

HUMAN FACTORS – The target users of this interface are security personnel and analysts with moderate salaries and motivation. Their jobs require that they be detail-oriented, constantly aware of what is going on, and alert of their surroundings. The cognitive load in this field is fairly high because there are multiple areas of focus in a distracting environment. The tasks they perform are very time-sensitive and crucial to the success of the organization.

- Moderate motivation
- Moderate salaries
- Moderate personality
- Job alternatives?
- High amount cognitive processes
- Time-sensitive tasks
- Multiple areas of focus

TASK FACTORS – The tasks that must be performed by our users are complex and have numerous components. Users must simultaneously monitor multiple channels using standardized and regimented processes. The tasks require a fairly high level of skill and training but are also extremely repetitive. Because of these task factors, our interface needs to be advanced enough to allow for various complexities of tasks, but must be able to clearly signal alerts to an operator who may have become jaded to the presence of the interface.

- High number of components
- Standardized & regimented processes
- High complexity
- Monitoring multiple channels
- High repetitiveness
- High level of skill required
- Job role relies on learned skill set

ENVIRONMENTAL FACTORS – The users of this system will usually be working in high-stress situations and have numerous responsibilities. They will also be in a public place (such as an airport) where there are many potential distractions. The interface will have to be one that is easy and quick, read, and interact with.

- High-stress responsibilities
- Countless potential distractions

- Low signal-to-noise ratio

REQUIRED SYSTEM CAPABILITIES – Additionally, the ideal transitioned product/system will feature the following characteristics:

- Simple-to-use
- Standard Windows interface controls
- Interactivity NOT required
- Ideally, a "PASSIVE" application
- Large, clear, colorful displays & read-outs
- User-customizable display granularity
- "Quick Reset" function
- "Sterile" presentation schema
- Not overly "busy" … to increase focus & aid DM
- In-line help tools
- Pop-up information screens

Following this preliminary analysis of the application environment, we began work on our initial prototype – a series of thumbnail diagrams that approximated a representational appearance of the desired end product. The initial wireframe drawings created were as shown in Figure 34.

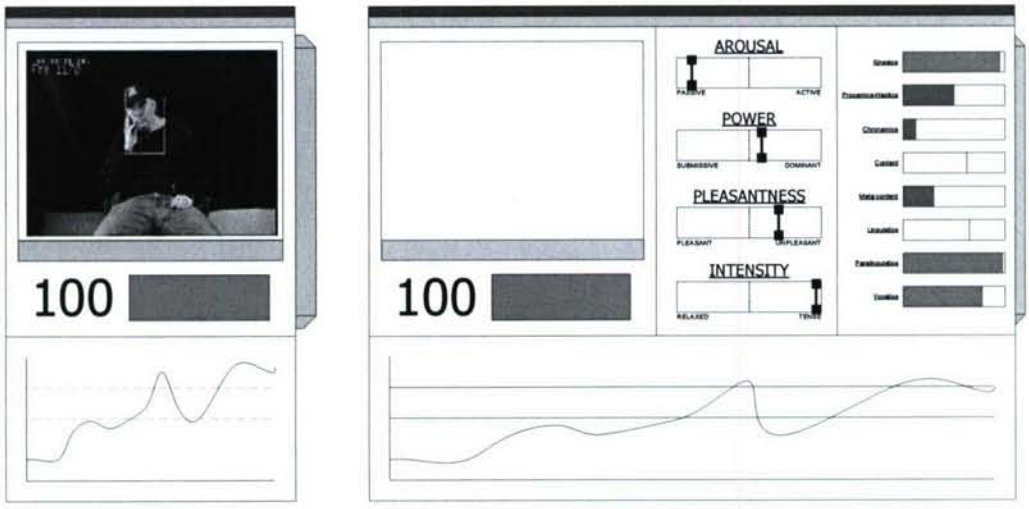


Figure 34. Preliminary Interface Design for Field-Deployable Application.

However, the process of iterative prototyping calls for continuously making revisions (whenever required) at various intervals throughout the development process, gradually improving their quality, until the ultimate objectives are sufficiently achieved. Our most recent versions of the GUI front-end are as shown in Figure 35.

It is important to note that the final transitioned product may not resemble any of the above representations at all. Simply put, we plan to iteratively revise the prototype design, content, and appearance regularly (as our research results dictate). Thus, as our lab experiments and field

studies continue to reveal more insight into what specific information is required by user of this application to make an accurate determination of an individual's level of deception, we will iteratively incorporate those findings into the software application's design until the ultimate objectives are sufficiently addressed.

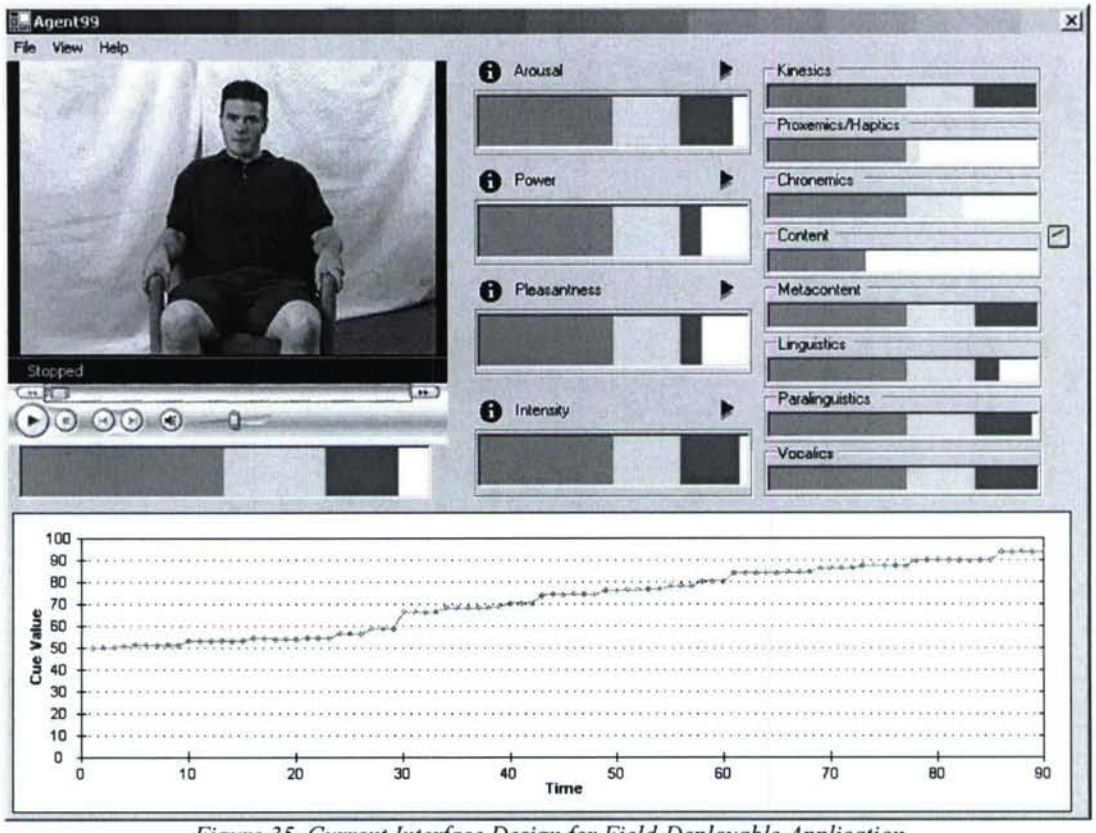


Figure 35. Current Interface Design for Field-Deployable Application

D. Trainer

Another goal of our research has been to improve human detection capabilities through development of training tools. To this end, we aimed to build a tool that is platform independent, easy to integrate, supports multi-users for a server version and that also offers a single-client version. The result was Agent99Trainer, a multi-perspective training tool. Agent99 Trainer provides explicit instruction on deception detection knowledge through the use of organized video and different types of real-life examples to help users get a broader, deeper understanding of concepts and theories. The modules of Agent99 Trainer provide a videotaped lecture, hands-on experience evaluating actual communications, interaction through the ability to ask questions, the ability to view examples, and a self-test via a pop-up quiz. The premise of the tool is that by providing explicit instruction, practice, feedback and interaction, effective training in deception detection can be accomplished.

Deception detection presents an ill-defined problem with no perfectly reliable cues. To glean a deep understanding requires extensive experience and high-level cognitive processing. Taking this into account, we designed Agent99 Trainer to be a learner-centered training system: a stand-alone system that is suitable for various environments that can be easily customized.

Training experiments were conducted to determine whether training with the Agent99 Trainer improves deception detection and what features of the tool are most beneficial. We conducted

two training pilots designed to compare Agent99 Trainer with traditional lecture groups. Findings showed that Agent99 students' detection accuracy improved by 63%, versus traditional lecture students' 46%. Further experiments were conducted at a U. S. Air Force base with officers in the basic communications training program. A control group was compared to treatment groups representing inclusion of different interface features. A pre-test, post-test design was utilized to determine improvements in knowledge (multiple choice questions) and judgment (tasks assessing truthful versus deceptive stimuli). The AFB results indicated that trained groups performed better than untrained groups on knowledge tests, all groups improved over time, and those trained with the full functionality of the Agent99 Trainer showed the most gains in knowledge and judgment. The implication was that training aided immediate knowledge gains, that computerized training was at least effective as traditional lecture, and that adding features such as pop-up quizzes and navigational flexibility improved performance.

Usability tests revealed that the attractive system features included ease of use, structured and synchronized multimedia lecture capability, multiple channels of training (e.g., video, audio, slides and text), self-paced learning, the ability to view examples and analysis; and practice and feedback components. The user comments provided important insights into future system design efforts.

We continue to test the viability of curriculum implementation in various delivery modes, including instructor-based classroom training, Web-based training, and a stand-alone computer program with different conditions (linear playback of video, user self-paced control, availability of the "Ask a Question" functionality (natural language querying and keyword searching) to validate our initial findings that online learning using the Agent99 Trainer was as effective as classroom training, and whether a user's ability to detect deception increases after training using the curriculum. This will allow us to use the existing curriculum as a template for creating different deception detection training programs for the Air Force, as well as civilian organizations.

1. StrikeCom

Virtually no research has examined deception under conditions of attempting to deceive multiple receivers and using different communication modes. To analyze deceptive communication in chat, audio, and face-to-face communication and to take into account the greater complexity of expanded team size, three experiments were performed at the University of Arizona, Florida State University, and in conjunction with Air Force Institute of Technology using StrikeCom, a simulation developed by the University of Arizona team. Participants in some experiments were U. S. Air Force ROTC cadets who used StrikeCom to conduct mock air operations. StrikeCom is an online, turn-based simulation of a C3ISR (Command, Control, Communication, Intelligence, Surveillance, Reconnaissance) task. The object of the game was for the three-person teams to find and destroy enemy camps that have been hidden on a game board. Each player controlled different intelligence assets. In some games, one team member was instructed to be deceptive and purposefully mislead the team away from the enemy camps. In other games, one team member was also made suspicious. All interactions between team members were recorded. Verbal and nonverbal behaviors of all three members are being analyzed.

Results to date indicate that team members became distrustful of deceivers, so something in their behavior cued receivers, but deceiver's information was still accepted and resulted in poorer team performance. This suggests that humans often do not act on their suspicions and continue to show biased information processing. Coding of speech acts has shown that utterances such as questions or backchannels can discriminate different facets of deception such as indications of uncertainty.

StrikeCom is being used by the U.S. Office of the Secretary of Defense Office of Force Transformation as a tool to illustrate key conceptual concepts of Network Centric Operations. NATO/OTAN Allied Command Transformation is using StrikeCom for a "hands on experience" to transform NATO's military capabilities. The United States Marine Corp's Expeditionary Warfare School identified a potential use of StrikeCom to teach concepts to their globally distributed distance learning classes. The U.S. Naval Post Graduate School is scheduled to use StrikeCom in December 2005 as a tool for Network Centric Warfare instruction. CMI will also use StrikeCom at the Network Centric Warfare Asia 2005 conference to highlight critical concepts. Finally, when funding is available in OCT05, StrikeCom will be transitioned to the U.S. Air Force Research Laboratory/Information Directorate as a networked multiplayer command and control game to explore linkages between C2 concepts and network centric operations.

E. Repository of Research to Facilitate Knowledge-Sharing

An integral part of our work has been to build a repository to facilitate the sharing of knowledge, both internally and externally. Requirements of the repository include the ability to handle all major file formats (e.g., Video – MPEG, AVI, MOV; Papers – DOC, PDF, WPD, HTML; Citations – Endnote; Presentations – PPT; Experimental Data – XML), as well as the ability to provide full-text searches and property searches. It must also be capable of providing varying levels of security (e.g., Public vs. CMI).

By incorporating best practices in document management, the repository will allow for locating documents and files quickly and accurately. We are also creating a deception test bed comprised of articles, working papers, citations, video, audio, text, experimental data and scenarios to enhance future ongoing and future research efforts. This will facilitate literature reviews; product, software and hardware reviews; training material reviews; and allows queries and updates via a web interface. CMI is continually redesigning our website to enable easier access to the data repository.

It is also our goal to extract additional indicators of deception by conducting further experiments and assessing the cross-contextual validity and reliability of resultant body movement, lexical and speech act indicators. We will also continue to develop prototypes (e.g., training and extraction enhancements) and further improve our deception detection integrated multimedia system.

XIII. References

Adami, A., Kajarekar, S., & Hermansky, H. (2002). *A new speaker change detection method for two-speaker segmentation.* Paper presented at the IEEE International Conference on Acoustics, Speech, Orlando, Florida.

Adkins, M., Twitchell, D. P., Burgoon, J. K., & Nunamaker Jr., J. F. (2004, April, 2004). *Advances in automated deception detection in text-based computer-mediated communication.* Paper presented at the Proceedings of the SPIE Defense and Security Symposium, Orlando, Florida.

Anolli, L., & Ciceri, R. (1997). The voice of deception: Vocal strategies of naive and able liars. *Journal of Nonverbal Behavior, 21*(4), 259-284.

Biros, D., George, J., & Zmud, R. (2002). Inducing sensitivity to deception in order to improve decision making performance: A field study. *MIS Quarterly, 26*(2), 119-144.

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

Bond, C. F., Kahler, K. N., & Paolicelli, L. M. (1985). The miscommunication of deception: An adaptive perspective. *Journal of Experimental Social Psychology, 21*, 331-345.

Buller, D., & Burgoon, J. (1996). Interpersonal deception theory. *Communication Theory, 6*, 203-242.

Buller, D., Burgoon, J., White, C., & Ebesu, A. (1994). Interpersonal deception: Vii. Behavioral profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology, 13*(4), 366-395.

Buller, D. B. (1987, November). *Detecting deception by strangers, friends, and intimates: Attributional biases due to relationship development.* Paper presented at the Speech Communication Association, Boston, MA.

Buller, D. B., & Burgoon, J. K. (1986). The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. *Human Communication Research, 13*(1), 126-144.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 3*, 203–242.

Buller, D. B., Burgoon, J. K., Buslig, A., & Roiger, J. (1996). Testing interpersonal deception theory: The language of interpersonal deception. *Communication Theory, 6*, 268-289.

Buller, D. B., Burgoon, J. K., White, C., & Ebesu, A. S. (1994). Interpersonal deception: Vii. Behavioral profiles of falsification, concealment, and equivocation. *Journal of Language and Social Psychology, 13*, 366-395.

Burgoon, J. K. (2005). The future of motivated deception detection. In P. Kalbfleisch (Ed.), *Communication Yearbook 29* (pp. 49-95). Mahway, NJ: Erlbaum.

Burgoon, J. K., & Burgoon, M. (2001). Expectancy theories. In P. R. H. Giles (Ed.), *Handbook of language and social psychology* (2nd ed., pp. 79-101). Sussex, England: John Wiley & Sons.

Burgoon, J. K., Blair, J. P., & Moyer, E. (2003). *Effects of communication modality on arousal, cognitive complexity, behavioral control and deception detection during deceptive episodes.* Paper presented at the Annual Meeting of the National Communication Association, Miami Beach, Florida.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). *Detecting deception through linguistic analysis.* Paper presented at the NSF/NIJ Symposium on Intelligence and Security Informatics.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F., Jr. (2003). Detecting deception through linguistic analysis. In H. Chen, R. Miranda, D. D. Zeng, C. Demchak & T. Madhusudan (Eds.), *Proceedings of the first NSF/NIJ symposium on intelligence and security informatics, tucson, az, USA* (pp. 91-101). Berlin: Springer-Verlag.

Burgoon, J. K., Blair, J. P., & Strom, R. E. (in press). Cognitive biases and nonverbal cue availability in detecting deception. *Human Communication Research.*

Burgoon, J. K., Bonito, J., & Kam, K. (Eds.). (2006). *Communication and trust under face-to-face and mediated conditions: Implications for leading from a distance*: Mahwah, NJ: LEA.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Ramirez, A., Dunbar, N., & Miczo, N. (2000). Testing the interactivity model: Communication processes, partner assessments, and the quality of collaborative work. *Journal of Management and Information Systems, 16*(3), 33-56.

Burgoon, J. K., Bonito, J. A., Ramirez, A., Jr., Dunbar, N. E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication, 52*(3), 657-677.

Burgoon, J. K., & Buller, D. B. (1994). Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior, 18*(2), 155-184.

Burgoon, J. K., & Buller, D. B. (2004). Interpersonal deception theory. In J. S. Seiter & R. H. Gass (Eds.), *Readings in persuasion, social influence and compliance-gaining* (pp. 239-264). Boston: Allyn & Bacon.

Burgoon, J. K., Buller, D. B., Dillman, L., & Walther, J. B. (1995a). Interpersonal deception .4. Effects of suspicion on perceived communication and nonverbal behavior dynamics. *Human Communication Research, 22*(2), 163-196.

Burgoon, J. K., Buller, D. B., Dillman, L., & Walther, J. B. (1995b). Interpersonal deception: Iv. Effects of suspicion on perceived communication and nonverbal behavior dynamics. *Human Communication Research, 22*(2), 163-196.

Burgoon, J. K., Buller, D. B., Ebesu, A. S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection. *Communication Monographs, 61*(4), 303-325.

Burgoon, J. K., Buller, D. B., Guerrero, L. K., & Afifi, W. A. (1996). Interpersonal deception: Xii, information management dimensions underlying deceptive and truthful messages. *Communication Monographs, 63*(1), 50-69.

Burgoon, J. K., Buller, D.B. (2004). Interpersonal deception theory. In J. S. S. R. H. Gass (Ed.), *Readings in persuasion, social influence and compliance-gaining* (pp. 239-264). Boston: Allyn & Bacon.

Burgoon, J. K., Buller, D.B., White, C.H., Afifi, W., & Buslig, A.L.S. (1999). The role of conversational involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin, 25*, 669-686.

Burgoon, J. K., Burgoon, M., & Sunnafrank, M. J. (1981). Learning theory approaches to persuasion. *Human Communication Research, 7*(2), 161-179.

Burgoon, J. K., & Floyd, K. (2000). Testing for the motivation impairment effect during deceptive and truthful interaction. *Western Journal of Communication, 64*, 243-267.

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs, 55*(1), 58-79.

Burgoon, J. K., & Hoobler, G. (2002). Nonverbal signals. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (3rd ed., pp. 240-299). Thousand Oaks, CA: Sage Publications.

Burgoon, J. K., Jensen, M. L., Kruse, J., Meservy, T. O., & Jay F. Nunamaker, J. (Forthcoming). Deception and intention detection. In H. Chen, T. S. Raghu, R. Ramesh, A. Vinze & D. Zeng (Eds.), *Handbooks in information systems---national security*.

Burgoon, J. K., & LePoire, B. A. (1993). Effects of communication expectancies, actual communication, and expectancy disconfirmation on evaluations of communicators and their communication behavior. *Human Communication Research, 20*(1), 67-96.

Burgoon, J. K., & Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology, 25*(1), 76-96.

Burgoon, J. K., & Walther, J. B. (1990). Nonverbal expectancies and the evaluative consequences of violations. *Human Communication Research, 17*(2), 232-265.

Burgoon, M., & Doran, N. E. (1983). Friendly or unfriendly persuasion: The effects of violations of expectations by males and females. *Human Communication Research, 10*(2), 283-294.

Carlson, J. R., & George, J. F. (2004). Media appropriateness in the conduct and discovery of deceptive communication: The relative influence of richness and synchronicity. *Group Decision and Negotiation, 13*(2), 191-210.

Carlson, J. R., George, J. F., Burgoon, J. K., Adkins, M., & White, C. H. (2004). Deception in computer-mediated communication. *Group Decision and Negotiation, 13*, 5 - 28.

Coker, D. A., & Burgoon, J. K. (1987). The nature of conversational involvement and nonverbal encoding patterns. *Human Communication Research, 13*(4), 463-494.

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

DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

DePaulo, B., Lindsay, J., Malone, B., Muhlenbruck, L, Charlton, K, & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

DePaulo, B., Stone, J., & Lassiter, D. (1985). Deceiving and detecting deceit. In (pp. 323-370).

DePaulo, B. M., Blank, A. L., Swain, G. W., & Hairfield, J. G. (1992). Expressiveness and expressive control. *Personality and Social Psychology Bulletin, 18*, 276–285.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality & Social Psychology, 70*(5), 979-995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

Ekman, P., Friesen, W., O'Sullivan, M., & Scherer, K. (1980). Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality & Social Psychology, 38*(2), 270-277.

Ekman, P., Friesen, W. V., & Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica, 16*, 23-27.

Feeley, T. H., & deTurck, M. A. (1995). Global cue usage in behavioral lie detection. *Communication Quarterly, 43*(4), 420-430.

Fiedler, K. (1993). Training lie detectors to use nonverbal cues instead of global heuristics. *Human Communication Research, 20*(2), 199-223.

Freud, S. (1959). *Collected papers*. New York, NY: Basic Books.

George, J., & Carlson, J. (1999). *Group support systems and deceptive communication.* Paper presented at the 32nd Hawaii International Conference on Systems Sciences, Maui, HI.

Greene, J. O., O'Hair, H. D., & Yen, C. (1985). Planning and control of behavior during deception. *Human Communication Research, 11*(3), 335-364.

Hancock, J. T., & Dunham, P. J. (2001). Impression formation in computer-mediated communication revisted. *Communication Research, 28*(3), 325-347.

Hocking, J., & Leatheres, D. (1980). Nonverbal indicators of deception: A new theoretical perspective. *Communication Monographs, 47,* 119-131.

Hollien, H., & Harnsberger, J. (2006). Voice stress analyzer instrumentation evaluation. *CIFA Final Report* (Contract FA 4814-04-0011). February.

Jensen, M. L., Meservy, T. O., Kruse, J., Burgoon, J. K., & Jay F. Nunamaker, J. (2005). *Identification of deceptive behavioral cues extracted from video.* Paper presented at the International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*, 253–260.

Kemp, T., Schmidt, M., Westphal, M., & Waibel, A. (2000). *Strategies for automatic segmentation of audio data.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey.

Levine, T. R., Anders, L. N., Banas, J., Baum, K. L., Endo, K., Hu, A. D. S., et al. (2000). Norms, expectations, and deception: A norm violation model of veracity judgments. *Communication Monographs, 67*(2), 123-137.

Levine, T. R., Kim, R., Park, H. S., & Hughes, M. (2006). Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model. *Communication Monographs, 73,* 243-260.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the 'veracity effect'. *Communication Monographs, 66*, 125-144.

Lu, S., Tsechpenakis, G., Metaxas, D. N., Jensen, M. L., & Kruse, J. (2005). *Blob analysis of the head and hands: A method for deception detection.* Paper presented at the Proceedings of the 38th Annual Hawaii International Conference on System Science (CD/ROM), Big Island, Hawaii.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: The University of Chicago Press.

Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Nunamaker, J. F. (2005). Automatic extraction of deceptive behavioral cues from video. In P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen & R. Merkle (Eds.), *Intelligence and security informatics: Proceedings of the ieee international conference on intelligence and security informatics isi 2005, atlanta, ga, USA* (Vol. 3495, pp. 198-208). Berlin: Springer - Verlag.

Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Nunamaker, J. F. (2006). *Detecting deception in a security screening scenario.* Paper presented at the Proceedings of the 38th Annual Hawaii International Conference on System Science (CD/ROM), Hawaii.

Meservy, T. O., Jensen, M. L., Kruse, J., Twitchell, D. P., Tsechpenakis, G., Burgoon, J. K., et al. (2005, Sept/Oct 2005). Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems, 20*.

Miller, G. R., & Burgoon, J. K. (1982). Factors influencing judgments of witness credibility and truthfulness. In N. L. Kerr & R. M. Bray (Eds.), *The psychology of the courtroom* (pp. 169-194). New York: Academic Press.

Nass, C. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research, 19*(4), 504-527.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 669-678.

Nass, C., Steuer, J., & Tauber, E. R. (1994, 1994). *Computers are social actors.* Boston, MA.

Nass, C. I., & Reeves, B. (1991). Combining, distinguishing, and generating theories in communication: A domains of analysis framework. *Communication Research, 18*, 240-261.

Nathan, J. (1998). *Back-to-basics audio*: Newnes.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Qin, T., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004, January 5-8). *An exploratory study on promising cues in deception detection and application of decision tree.* Paper presented at the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI.

Quek, F., Ma, X.-F., & Bryll, R. (1999). *A parallel algorithm for dynamic gesture tracking.* Paper presented to the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece.

Rockwell, P., Buller, D. B., & Burgoon, J. K. (1997). The voice of deceit: Refining and expanding vocal cues to deception. *Communication Research Reports, 14*, 451-459.

Steller, M., & Kohnken, G. (1989). Criteria-based content analysis. In D. C. Raskin (Ed.), *Psychological Methodsl in Criminal Investigation and Evidence* (pp., 217-245). New York: Springer-Verlag..

Stoner, G. M. (2001). *Decision-making via mediated communication: Effects of mediation and time pressure.* Unpublished Masters Thesis, University of Arizona, Tucson, AZ.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their rocs and implied models. *Psychological Bulletin, 99*, 301-340.

Swets, J. A. (2000). Enhancing diagnostic decisions. In T. Connolly, H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader.* Cambridge, England: Cambridge University Press.

Swets, J. A. (2000b). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.

Tung, P., Lam, R., & Tsang, W. K. (1997). English as a medium of instruction in post-1997 hong kong: What students, teachers, and parents think. *Journal of Pragmatics, 28*(4), 441-459.

Turner, R. E. E., C.; Olmstead, G. (1975). Information control in conversations: Honesty is not always the best policy. *Kansas Journal of Speech, 11*, 69-89.

Vrij, A. (1994). The impact of information and setting on detection of deception by police detectives. *Journal of Nonverbal Behavior, 18*, 117–136.

Vrij, A. (1999). *Interviewing to detect deception. Handbook of the psychology of interviewing*: Chichester, UKs: John Wiley & Sons Ltd.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Chichester: John Wiley & Sons.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practices*. West Sussex, UK: John Wiley and Sons.

Vrij, A., Akehurst, L., & Morris, P. (1997). Individual differences in hand movements during deception. *Journal of Nonverbal Behavior, 21*, 87-103.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239-263.

White, C. H., & Burgoon, J. K. (2001). Adaptation and communicative design: Patterns of interaction in truthful and deceptive conversations. *Human Communication Research, 27*(1), 9-37.

Wren, C. R., Azarbayejani, A. J., Darrell, T., & Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 780-785.

Zhou, L., Burgoon, J. K., Nunamaker, J., Jay F., & Twitchell, D. P. (2004). Automated linguistics based cues for detecting deception in text-based asynchronous computer-mediated communication: An empirical investigation. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., Jr., & Twitchell, D. P. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, L., Burgoon, J. K., & Twitchell, D. (2003). *A longitudinal analysis of language behavior of deception in e-mail.* Paper presented at the Proceedings of the Symposium on Intelligence and Security Informatics, Tucson, AZ.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., & Nunamaker Jr., J. F. (2003, January 6-9, 2003). *An exploratory study into deception detection in text-based computer-mediated communication.* Paper presented at the Thirty-Sixth Annual Hawaii International Conference on System Sciences (CD/ROM), Big Island, Hawaii.

Zuckerman, M., DePaulo, B., & 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., & Driver, R. (1985). Telling lies: Verbal and nonverbal correlates of deception. In A. W. Siegman & S. Feldstein (Eds.), *Nonverbal communication: An integrated perspective* (pp. 129-147). Hillsdale, NJ: Erlbaum.

XIV. Appendix A: Publications

1 Adkins, M., Burgoon, M., & Nunamaker, J. F., Jr. (2003). Using group support systems for strategic planning with the United States Air Force. *Decision Support Systems, 34*(3), 315-337.

2 Adkins, M., Kruse, J., Damianos, L. E., Brooks, J., Younger, R. E., Rasmussen, E., et al. (2001, January 3-6). *Experience using collaborative technology with the United Nations and multi-national militaries: Rim of the Pacific 2000 Strong Angel exercise in humanitarian assistance.* Proceedings of the 34th Annual Hawaii International Conference on System Sciences (CD-ROM), Wailea, Maui, HI, p. 10.

3 Adkins, M., Kruse, J., McKenna, T., Cuyugan, A., Nunamaker, J. F., Jr., Miller, S., et al. (2000, January 4-7). *Experiences developing a virtual environment to support disaster relief with the United States Navy's Commander Third Fleet.* Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (CD-ROM), Wailea, Maui, HI, p. 9.

4 Adkins, M., Kruse, J., & Younger, R. E. (2002, January 7-10). *Ubiquitous computing: Omnipresent technology in support of network centric warfare.* Proceedings of the 35th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, pp. 551-559.

5 Adkins, M., Kruse, J., & Younger, R. E. (2004, January 5-8). *A language technology toolset for development of a large group augmented facilitation system.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 9.

6 Adkins, M., & Schwarz, R. (2002, January 7-10). *Embedded facilitation requirements using the Skilled Facilitator approach: With and across time and space.* Proceedings of the 35th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, pp. 560-567.

7 Adkins, M., Twitchell, D. P., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004, August). *Advances in automated deception detection in text-based computer-mediated communication.* Enabling Technologies for Simulation Science VIII, Orlando, Florida, pp. 122-129.

8 Adkins, M., Younger, R. E., & Schwarz, R. (2003, January 6-9). *Information technology augmentation of the Skilled Facilitator approach.* Proceedings of the 36th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 10.

9 Allspach, L. E., & Burgoon, J. K. (2003). Nonverbal displays. In *Encyclopedia Of Human Emotions*. New York: Macmillan.

10 Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & Virtual Environments, 12*(5), 456-480.

11 Biocca, F., Harms, C., & Burgoon, J. K. (in press). Criteria and scope conditions for a theory and measure of social presence. *Presence: Teleoperators and Virtual Environments.*

12 Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation, 13*(2), 173-189.

13 Biros, D. P., Hass, M. C., Wiers, K., Twitchell, D. P., Adkins, M., Burgoon, J. K., et al. (2005, January 3-6). *Task performance under deceptive conditions: Using military scenarios in deception detection research.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 22.

14 Biros, D. P., Sakamoto, J., George, J. F., Adkins, M., Kruse, J., Burgoon, J. K., et al. (2005, January 3-6). *A quasi-experiment to determine the impact of a computer-based deception detection training system: The use of Agent99 Trainer in the U.S. military.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 24.

15 Blair, J. P., & Horvath, F. (in press). Detection of deception accuracy using the verbal component of the Behavioral Analysis Interview model. *Paper submitted for publication consideration.*

16 Briggs, R. O., De Vreede, G.-J., & Nunamaker, J. F., Jr. (2003). Collaboration engineering with ThinkLets to pursue sustained success with group support systems. *Journal of Management Information Systems, 19*(4), 31-64.

17 Briggs, R. O., De Vreede, G.-J., Nunamaker, J. F., Jr., & Sprague, R. H., Jr. (2002). Special issue: Decision-making and a hierarchy of understanding. *Journal of Management Information Systems, 18*(4), 5-9.

18 Briggs, R. O., De Vreede, G.-J., Nunamaker, J. F., Jr., & Sprague, R. H., Jr. (2003). Special issue: Information systems success. *Journal of Management Information Systems, 19*(4), 5-8.

19 Burgoon, J. K. (2001). The challenge of writing the theoretical essay. In J. Potter & A. Alexander (Eds.), *Editor's guide to scholarly publishing* (pp. 47-56). Thousand Oaks, CA: Sage.

20 Burgoon, J. K. (2005). The future of motivated deception detection. In P. Kalbfleisch (Ed.), *Communication Yearbook* (Vol. 29, pp. 49-95). Mahway, NJ: Erlbaum.

21 Burgoon, J. K. (2005). Nonverbal measurement of deceit. In V. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 237-250). Hillsdale, NJ: Erlbaum.

22 Burgoon, J. K., Adkins, M., Kruse, J., Jensen, M. L., Deokar, A. V., Twitchell, D. P., et al. (2004, October 3-6). *Moving toward intent detection: A tool-based approach.* Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, Washington,

DC, pp. 124-129.

23 Burgoon, J. K., Adkins, M., Kruse, J., Jensen, M. L., Deokar, A. V., Twitchell, D. P., et al. (2004). *The decision for detecting intent.* Global Business and Finance Research Conference, London, England.

24 Burgoon, J. K., Adkins, M., Kruse, J., Jensen, M. L., Meservy, T. O., Twitchell, D. P., et al. (2005, January 3-6). *An approach for intent identification by building on deception detection.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 21a.

25 Burgoon, J. K., & Bacue, A. (2003). Nonverbal communication skills. In B. R. Burleson & J. O. Greene (Eds.), *Handbook of Communication and Social Interaction Skills* (pp. 179-219). Mahwah, NJ: Erlbaum.

26 Burgoon, J. K., Blair, J. P., & Moyer, E. (2003, November 19-23). *Effects of communication modality on arousal, cognitive complexity, behavioral control and deception detection during deceptive episodes.* National Communication Association 2003 Convention, Miami, FL.

27 Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F., Jr. . (2003, June 2-3). *Detecting deception through linguistic analysis.* Proceedings of First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI) 2003, Tucson, AZ, pp. 91-101.

28 Burgoon, J. K., Blair, J. P., & Strom, R. E. (2005, January 3-6). *Heuristics and modalities in determining truth versus deception.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 19a.

29 Burgoon, J. K., Bonito, J. A., Ramirez, A., Jr., Dunbar, N. E., Kam, K., & Fischer, J. (2002). Testing the Interactivity Principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication, 52*(3), 657.

30 Burgoon, J. K., & Buller, D. B. (2004). Interpersonal deception theory. In J. S. Seiter & R. H. Gass (Eds.), *Readings in persuasion, social influence and compliance-gaining* (pp. 239-264). Boston: Allyn & Bacon.

31 Burgoon, J. K., Buller, D. B., Blair, J. P., & Tilley, P. (2006). Sex differences in presenting and detecting deceptive messages. In D. Canary & K. Dindia (Eds.), *Sex differences and similarities in communication* (2nd ed.). Mahwah, NJ: LEA.

32 Burgoon, J. K., Buller, D. B., & Floyd, K. (2001). Does participation affect deception success? A test of the Interactivity Principle. *Human Communication Research, 27*(4), 503-534.

33 Burgoon, J. K., & Burgoon, M. (2001). Expectancy theories. In P. Robinson & H. Giles (Eds.), *Handbook of language and social psychology* (2nd ed., pp. 79-101). Sussex, England.

34 Burgoon, J. K., Burgoon, M., Broneck, K., Alvaro, E., & Nunamaker, J. F., Jr. (2002, November 21-24). *Effects of synchronicity and proximity on group communication.* 2002 Annual Meeting of the National Communication Association, New Orleans, LA.

35 Burgoon, J. K., Chen, F., & Twitchell, D. P. (in press). Deception and its detection under synchronous and asynchronous computer-mediated communication. *Group Decision and Negotiation.*

36 Burgoon, J. K., & Dunbar, N. E. (2006). Dominance, power and influence. In V. Manusov & M. L. Patterson (Eds.), *The SAGE Handbook of Nonverbal Communication.* Thousand Oaks, CA: Sage Publications.

37 Burgoon, J. K., & Ebesu-Hubbard, E. (2004). Expectancy violations theory and interaction adaptation theory. In W. B. Gudykunst (Ed.), *Theorizing About Intercultural Communication.* Thousand Oaks, CA: Sage Publications.

38 Burgoon, J. K., George, J. F., Carlson, J. R., Blair, J. P., & Adkins, M. (in press). Deception and trust in online communication. In J. F. Nunamaker, Jr. & R. O. Briggs (Eds.), *Advances in MIS: Foundations of collaboration.* Armonk, NY: M.E. Sharpe Inc.

39 Burgoon, J. K., & Hoobler, G. (2002). Nonverbal signals. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (3rd ed., pp. 240-299). Thousand Oaks, CA: Sage Publications.

40 Burgoon, J. K., Jensen, M. L., Meservy, T. O., Kruse, J., & Nunamaker, J. F., Jr. (2005, May 2-6). *Augmenting human identification of emotional states in video.* Intelligence Analysis Conference, McLean, VA.

41 Burgoon, J. K., Marett, L. K., & Blair, J. P. (2004). Detecting deception in computer-mediated communication. In J. F. George (Ed.), *Computers in society: Privacy, ethics & the Internet* (pp. 154-166). Upper Saddle River, NJ: Prentice-Hall.

42 Burgoon, J. K., & Nunamaker, J. F., Jr. (2004). Computer-aided support of the detection of deception. *Group Decision and Negotiation, 13*(2), 107.

43 Burgoon, J. K., & Nunamaker, J. F., Jr. (2004). Toward computer-aided support for the detection of deception. *Group Decision and Negotiation, 13*(1), 1.

44 Burgoon, J. K., & Qin, T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology, 25*(1), 76-96.

45 Burgoon, J. K., Segrin, C., & Dunbar, N. E. (2002). Nonverbal communication and social influence. In J. P. Dillard & M. Pfau (Eds.), *The persuasion handbook: Developments in theory and practice* (pp. 445-473). Thousand Oaks, CA: Sage Publications.

46 Burgoon, J. K., Stoner, G. A., Bonito, J. A., & Dunbar, N. E. (2003, January 6-9). *Trust and deception in mediated communication.* Proceedings of the 36th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 11.

47 Burgoon, J. K., Twitchell, D. P., Jensen, M. L., Adkins, M., Kruse, J., Deokar, A. V., et al. (in press). Detecting concealment of intent in transportation screening: a proof-of-concept. *IEEE Intelligent Systems.*

48 Cao, J., Crews, J. M., Lin, M., Burgoon, J. K., & Nunamaker, J. F., Jr. (2003, August 3-5). *Can people be trained to better detect deception?: Instructor-led vs. web-based training.* Ninth Americas Conference on Information Systems (AMCIS 2003), Tampa, FL.

49 Cao, J., Crews, J. M., Lin, M., Burgoon, J. K., & Nunamaker, J. F., Jr. (2003). Designing Agent99 Trainer: A learner-centered, web-based training system for deception detection In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2-3, 2003. Proceedings* (Vol. 2665/2003, pp. 358-365): Springer Berlin / Heidelberg.

50 Cao, J., Crews, J. M., Lin, M., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004, January 5-8). *Evaluation of users' experiences in Agent99 Trainer: A web-based multimedia training system for deception detection training.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI.

51 Cao, J., Crews, J. M., Nunamaker, J. F., Jr., Burgoon, J. K., & Lin, M. (2004, January 5-8). *User experience with Agent99Trainer: A usability study.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, Hawaii.

52 Cao, J., Lin, M., Deokar, A. V., Burgoon, J. K., Crews, J. M., & Adkins, M. (2004). Computer-based training for deception detection: What users want? In *Intelligence and Security Informatics* (Vol. 3073, pp. 163-175). Heidelberg: Springer Berlin.

53 Cao, J., Lin, M., Deokar, A. V., Burgoon, J. K., Crews, J. M., & Adkins, M. (2004). Computer-based training for deception detection: What users want? . In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics* (Vol. 3073/2004, pp. 163-175): Springer Berlin / Heidelberg.

54 Cao, J., & Zhang, D. (2001). *An intelligent learning assistant in multimedia-based interactive e-learning.* Seventh Americas Conference on Information Systems (AMCIS 2001), Boston, MA, pp. 127-133.

55 Carlson, J. R., & George, J. F. (2004). Media appropriateness in the conduct and discovery of deceptive communication: The relative influence of richness and synchronicity. *Group Decision and Negotiation, 13*(2), 191-210.

56 Carlson, J. R., George, J. F., Burgoon, J. K., Adkins, M., & White, C. H. (2002, August). *Deception in computer-mediated communication.* Annual meeting of the Academy of Management, Denver, CO.

57 Carlson, J. R., George, J. F., Burgoon, J. K., Adkins, M., & White, C. H. (2004). Deception in computer-mediated communication. *Group Decision and Negotiation, 13*(1), 5.

58 den Hengst, M., & Adkins, M. (2005, January 3-6). *The demand rate of facilitation functions.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 43c.

59 den Hengst, M., & Adkins, M. (2005, July 10-13). *Which facilitation functions are most challenging: A global survey of facilitators.* Group Decision and Negotiation Conference (GDN 2005), Vienna, Austria.

60 den Hengst, M., & Adkins, M. (2007, January 3-6). *Which collaboration patters are most challenging: A global survey of facilitators.* Proceedings of the 40th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, Kauai, HI.

61 Dunbar, N. E., & Burgoon, J. K. (2005). Nonverbal measurement of dominance. In V. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 361-374). Mahwah, NJ: Erlbaum.

62 Dunbar, N. E., & Burgoon, J. K. (2005). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships, 22*(2), 207-233.

63 Dunbar, N. E., Ramirez, A., Jr., & Burgoon, J. K. (2003). The effects of participation on the ability to judge deceit. *Communication Reports, 16*(1), 23-33.

64 Fuller, C., Biros, D. P., Adkins, M., Burgoon, J. K., Nunamaker, J. F., Jr., & Coulon, S. (2006). Detecting deception in person-of-interest statements. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: Fourth NSF/NIJ Symposium, ISI 2006, San Diego, CA, USA, May 23-24, 2006. Proceedings* (Vol. 3975/2006, pp. 504-509): Springer Berlin / Heidelberg.

65 Fuller, C., Biros, D. P., Burgoon, J. K., Adkins, M., & Twitchell, D. P. (2006, August 4-6). *An analysis of text-based deception detection tools.* 12th Americas Conference on Information Systems (AMCIS 2006), Acapulco, Mexico.

66 George, J. F. (2004). The theory of planned behavior and Internet purchasing. *Internet Research, 14*(3), 198.

67 George, J. F., Biros, D. P., Adkins, M., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004). Testing various modes of computer-based training for deception detection. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: Second NSF/NIJ Symposium, ISI 2004, Tucson, AZ, USA, June 10-11, 2004. Proceedings* (Vol. 3073/2004, pp. 411-417): Springer Berlin / Heidelberg.

68 George, J. F., Biros, D. P., Burgoon, J. K., & Nunamaker, J. F., Jr. (2003). Training professionals to detect deception. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2-3, 2003. Proceedings* (Vol. 2665/2003, pp. 366-370): Springer Berlin / Heidelberg.

69 George, J. F., & Carlson, J. R. (2005, January 3-6). *Media selection for deceptive communication.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 21b.

70 George, J. F., & Marett, L. K. (2004, January 5-8). *Inhibiting deception and its detection.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 10.

71 George, J. F., & Marett, L. K. (2004, May). The truth about lies. *HR Magazine, 49.*

72 George, J. F., & Marett, L. K. (2005). Deception: The dark side of e-collaboration. *International Journal of e-Collaboration, 1*(4), 24-37.

73 George, J. F., Marett, L. K., Burgoon, J. K., Crews, J. M., Cao, J., Lin, M., et al. (2004, January 5-8). *Training to detect deception: An experimental investigation.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 10023a.

74 George, J. F., Marett, L. K., & Tilley, P. (2004, January 5-8). *Deception detection under varying electronic media and warning conditions.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 9.

75 Hale, J. L., Burgoon, J. K., & Householder, B. (2005). Nonverbal measurement of relational communication. In V. Manusov (Ed.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 127-139). Hillsdale, NJ: Erlbaum.

76 Helquist, J. H., Jensen, M. L., Meservy, T. O., Deokar, A. V., Diller, C. B. R., Burgoon, J. K., et al. (2006, January 4-7). *A framework for studying human-agent teamwork.* Proceedings of the 39th Annual Hawaii International Conference on System Sciences (CD-ROM), Koloa, Kauai, HI.

77 Helquist, J. H., Kruse, J., & Adkins, M. (2006, January 4-7). *Developing large scale participant-driven group support systems: An approach to facilitating large groups.* First HICSS Symposium on Field and Case Studies of Collaboration, Koloa, Kauai, HI.

78 Helquist, J. H., Kruse, J., & Adkins, M. (2006). *Group support systems for very large groups: A peer review process to filter brainstorming input.* Americas Conference on Information Systems, Acapulco, Mexico.

79 Hender, J. M., Dean, D. L., Rodgers, T. L., & Nunamaker, J. F., Jr. (2002). An examination of the impact of stimuli type and GSS structure on creativity: Brainstorming versus non-brainstorming techniques in a GSS environment. *Journal of Management Information Systems, 18*(4), 59-85.

80 Jensen, M. L., Meservy, T. O., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, April 27-28). *Automated deception detection from nonverbal behavioral cues (Presentation).* Department of Homeland Security Research and Development Conference, Boston, MA.

81 Jensen, M. L., Meservy, T. O., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, September 13-16). *Identification of deceptive behavioral cues extracted from video.* International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria.

82 Jensen, M. L., Meservy, T. O., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, June 27-29). *Lessons learned from developing systems to automatically detect deception in various modalities.* Global Business and Finance Research Conference, London, England.

83 Kruse, J., Adkins, M., & Holloman, K. A. (2005, January 3-6). *Network centric warfare in the U.S. Navy's fifth fleet: Web-supported operational level command and control in Operation Enduring Freedom.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 45a.

84 Le Poire, B., Duggan, A., Shepard, C., & Burgoon, J. K. (2002). Relational messages associated with nonverbal involvement, pleasantness, and expressiveness in romantic couples. *Communication Research Reports, 19*(3), 195-206.

85 Lin, M., Crews, J. M., Cao, J., Nunamaker, J. F., Jr., & Burgoon, J. K. (2003, August 3-5). *Agent99 Trainer: Designing a web-based multimedia training system for deception detection knowledge transfer.* Ninth Americas Conference on Information Systems (AMCIS 2003), Tampa, FL.

86 Lowry, P. B., Albrecht, C. C., Nunamaker, J. F., Jr., & Lee, J. D. (2003). Evolutionary development and research on Internet-based collaborative writing tools and processes to enhance eWriting in an eGovernment setting. *Decision Support Systems, 34*(3), 229-252.

87 Lu, S., Tsechpenakis, G., Metaxas, D. N., Jensen, M. L., & Kruse, J. (2005, January 3-6). *Blob analysis of the head and hands: A method for deception detection.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 20c.

88 Marett, L. K., & George, J. F. (2004). Deception in the case of one sender and multiple receivers. *Group Decision and Negotiation, 13*(1), 29.

89 Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, May 19-20). *Automatic extraction of deceptive behavioral cues from video.* The IEEE International Conference on Intelligence and Security Informatics (ISI 2005), Atlanta, GA, pp. 198-208.

90 Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005). Automatic extraction of deceptive behavioral cues from video. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: Third NSF/NIJ Symposium, ISI 2005, Atlanta, GA, USA, May 19-20, 2005. Proceedings* (Vol. 3495/2005, pp. 198-208): Springer Berlin / Heidelberg.

91 Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Nunamaker, J. F., Jr. (2006, January 4-7). *Detecting deception in a security screening scenario.* Proceedings of the 39th Annual Hawaii International Conference on System Sciences (CD-ROM), Koloa, Kauai, HI.

92 Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., Nunamaker, J. F., Jr., Twitchell, D. P., et al. (2005). Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems, 20*(5), 36-43.

93 Nunamaker, J. F., & Briggs, R. O. (2003). Collaboration technology research: Current topics and applications. In J. F. Nunamaker, Jr. & R. O. Briggs (Eds.), *Advances in MIS: Foundations of collaboration* (Vol. 2). Armonk, NY: M.E. Sharpe Inc.

94 Nunamaker, J. F., & Briggs, R. O. (2003). Collaboration technology research: Foundations. In J. F. Nunamaker, Jr. & R. O. Briggs (Eds.), *Advances in MIS: Foundations of collaboration* (Vol. 1). Armonk, NY: M.E. Sharpe Inc.

95 Nunamaker, J. F., Jr., Romano, N. C., & Briggs, R. O. (2002). Increasing intellectual bandwidth: Generating value from intellectual capital with information technology. *Group Decision and Negotiation, 11*(2), 69-86.

96 Qin, T., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004, January 5-8). *An exploratory study on promising cues in deception detection and application of decision tree.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, pp. 23-32.

97 Qureshi, S., Hlupic, V., De Vreede, G.-J., Briggs, R. O., & Nunamaker, J. F. (2002). Managing knowledge in a collaborative context: How may intellectual resources be harnessed towards joint effect? In V. Hlupic (Ed.), *Knowledge and business process management* (pp. 174-195). Hershey, PA: Idea Group Publishing.

98 Ramirez, A., Jr., & Burgoon, J. K. (2004). The effect of interactivity on initial interactions: the influence of information valence and modality and information richness on computer-mediated interaction. *Communication Monographs, 71*(4), 422-447.

99 Ramirez, A., Jr., Walther, J. B., Burgoon, J. K., & Sunnafrank, M. (2002). Information-seeking strategies, uncertainty, and computer-mediated communication: Toward a conceptual model. *Human Communication Research, 28*(2), 213-228.

100 Romano, N. C., Donovan, C., Chen, H., & Nunamaker, J. F., Jr. (2003). A methodology for analyzing web-based qualitative data. *Journal of Management Information Systems, 19*(4), 213-246.

101 Thatcher, J. B., & George, J. F. (2004). Commitment, trust, and social involvement: An exploratory study of antecedents to web shopper loyalty. *Journal of Organizational Computing & Electronic Commerce, 14*(4), 243-268.

102 Tsechpenakis, G., Metaxas, D. N., Adkins, M., Kruse, J., Burgoon, J. K., Jensen, M. L., et al. (2005, July 6-8). *HMM-based deception recognition from visual cues.* IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, The Netherlands, pp. 824-827.

103 Twitchell, D. P., Adkins, M., Nunamaker, J. F., Jr., & Burgoon, J. K. (2004, June 2-3). *Using Speech Act Theory to model conversations for automated classification and retrieval.* Proceedings of the International Working Conference Language Action Perspective Communication Modelling (LAP 2004), New Brunswick, NJ, pp. 121-130.

104 Twitchell, D. P., Biros, D. P., Adkins, M., Meek, N. F., Burgoon, J. K., & Nunamaker, J. F., Jr. (2006, January 4-7). *Automated determination of the veracity of interview statements from people of interest to an operational security force.* Proceedings of the 39th Annual Hawaii International Conference on System Sciences (CD-ROM), Koloa, Kauai, HI, p. 17a.

105 Twitchell, D. P., Jensen, M. L., Burgoon, J. K., & Nunamaker, J. F., Jr. (2004, October 3-6). *Detecting deception in secondary screening interviews using linguistic analysis.* Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, Washington, DC, pp. 118-123.

106 Twitchell, D. P., & Nunamaker, J. F., Jr. (2004, January 5-8). *Speech act profiling: A probabilistic method for analyzing persistent conversations and their participants.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI, p. 10.

107 Twitchell, D. P., Nunamaker, J. F., Jr., & Burgoon, J. K. (2004). Using speech act profiling for deception detection. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: Second NSF/NIJ Symposium, ISI 2004, Tucson, AZ, USA, June 10-11, 2004. Proceedings* (Vol. 3073/2004, pp. 403-410): Springer Berlin / Heidelberg.

108 Twitchell, D. P., Wiers, K., Adkins, M., Burgoon, J. K., & Nunamaker, J. F., Jr. (2005, January 3-6). *StrikeCOM: A multi-player online strategy game for researching and teaching group dynamics.* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI.

109 White, C. H., & Burgoon, J. K. (2001). Adaptation and communicative design: Patterns of interaction in truthful and deceptive conversations. *Human Communication Research, 27*(1), 9-37.

110 Zhang, D., & Nunamaker, J. F., Jr. (2004). A natural language approach to content-based video indexing and retrieval for interactive e-learning. *IEEE Transactions on Multimedia, 6*(3), 450-458.

111 Zhou, L., Burgoon, J. K., Nunamaker, J. F., Jr., & Twitchell, D. P. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*(1), 81-106.

112 Zhou, L., Burgoon, J. K., & Twitchell, D. P. (2003). A longitudinal analysis of language behavior of deception in e-mail. In H. Chen, R. Moore, D. Zeng & J. Leavitt (Eds.), *Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003, Tucson, AZ, USA, June 2-3, 2003. Proceedings* (Vol. 2665/2003, pp. 102-110): Springer Berlin / Heidelberg.

113 Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F., Jr. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139-165.

114 Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F., Jr. (in press). Towards the automatic prediction of deception: An empirical comparison of classification methods. *Journal of Management Information Systems.*

115 Zhou, L., Burgoon, J. K., Zhang, D., & Nunamaker, J. F., Jr. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior, 20*(3), 381-402.

116 Zhou, L., Twitchell, D. P., Tiantian, Q., Burgoon, J. K., & Nunamaker, J. F., Jr. (2003, January 6-9). *An exploratory study into deception detection in text-based computer-mediated communication.* Proceedings of the 36th Annual Hawaii International Conference on System Sciences (CD-ROM), Waikoloa, HI.