THE EFFECTS OF AN EXPERT SYSTEM ON NOVICE AND PROFESSIONAL DECISION MAKING WITH APPLICATION IN DECEPTION DETECTION

by

MATTHEW LYNN JENSEN

A Dissertation Submitted to the Faculty of the

COMMITTEE ON BUSINESS ADMINISTRATION

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2007

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by MATTHEW LYNN JENSEN

entitled THE EFFECTS OF AN EXPERT SYSTEM ON NOVICE AND PROFESSIONAL DECISION MAKING WITH APPLICATION IN DECEPTION DETECTION

and recommend that it be accepted as fulfilling the dissertation requirement for the

DEGREE OF DOCTOR OF PHILOSOPHY

____ Date: June 11, 2007
JAY F. NUNAMAKER, JR.

____ Date: June 11, 2007
JUDEE K. BURGOON

____ Date: June 11, 2007
MOHAN TANNIRU

____ Date: June 11, 2007
AMNON RAPOPORT

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date: June 11, 2007
Dissertation Director: JAY F. NUNAMAKER, JR.

____ Date: June 11, 2007
Dissertation Director: JUDEE K. BURGOON

STATEMENT BY AUTHOR

ACKNOWLEDGEMENTS

Portions of this dissertation have been previously published in [17, 20, 59-61, 76, 82-84]. This dissertation was supported by the Defense Academy for Credibility Assessment under the direction of Counterintelligence Field Activity through a Personnel Dissertation Research Award Program. In addition, portions of this research were supported by funding from the U. S. Air Force Office of Scientific Research under the U. S. Department of Defense University Research Initiative (Grant #F49620-01-1-0394) and Department of Homeland Security – Science and Technology Directorate under cooperative agreement NBC2030003. The views, opinions, and/or findings in this report are those of the author and should not be construed as an official U.S. Government position, policy, or decision.

I would like to thank my committee, Jay F. Nunamaker, Jr., Judee K. Burgoon, Mohan Tanniru, and Amnon Rapoport for their consistent support and assistance. Their constructive feedback has been most appreciated through all levels of this research.

I would like to thank my colleagues James Lee, John Kruse, Mark Adkins, Douglas P. Twitchell, Thomas O. Meservy, Joel Helquist, Christopher Diller, Karl Wiers and the remainder of those in the Center for the Management of Information, the MIS faculty.

I would also like to thank Timothy Levine for allowing me to use the videos captured during his experiments as stimuli for my experiments and Roberto Mejias for assisting in recruiting novice and professional participants for my experiments.

Finally, I wish to thank my parents, Michael and Jean Jensen, for their support.

DEDICATION

To Carolina, Benjamin, Daniel, and Christian for their sacrifice, love, and unending support.

Table of Contents

Table of Contents - *Continued*

Table of Contents - *Continued*

LIST OF FIGURES

List of Figures - *Continued*

LIST OF TABLES

List of Tables - *Continued*

ABSTRACT

One effective way for organizations to capture expert knowledge and experience is to encapsulate it within an expert system (ES) and make that system available to others. While ES users have access to the system's knowledge, they shoulder the difficult task of appropriately incorporating the ES recommendations into the decision-making process.

One proposed application of an ES is in the realm of deception detection. Humans are inherently poor at recognizing deception when it occurs and their confidence in their judgments is poorly calibrated to their performance. An ES has the potential to significantly improve deception detection; however, joining an ES and a human decision maker creates many important questions that must be addressed before such a system will be useful in a field environment. These questions concern changes in decision outcomes, decision processes, and the decision maker that result from ES use.

To examine these questions, a prototype system was created that implements new and unobtrusive methods of deception detection. Kinesic analysis examines the body movement of a potential deceiver and linguistic analysis reviews the structure of utterances from a potential deceiver. This prototype, complete with explanations, was utilized in two experiments that examined the effects of access to the prototype, accuracy level of the prototype, user training in deception detection, and novice or professional lie-catcher status of the users.

Use of the prototype system was found to significantly improve professional and novice accuracy rates and confidence alignment. Training was found to have no effect on

novice accuracy rates. Accuracy level of the prototype significantly elevated accuracy rates and confidence alignment among novices; however, this improvement was imperceptible to the novices. Novices using the prototype performed on a level equivalent to professionals using the prototype. Neither professional nor novice users of the prototype exceeded the performance of the prototype system alone. Implications of these findings include emphasizing the development of computer-based tools to detect deception and defining a new role for human users of such tools.

CHAPTER 1

INTRODUCTION

Organizations increasingly seek to capture expertise and experience that is valuable in decision making. One effective way to capture expert knowledge is to encapsulate it within an expert system (ES) and make that system available to others who may be less knowledgeable and less experienced. While past attempts at capturing and reusing expert knowledge have had mixed results, organizations still believe and invest in ESs in order to improve organizational performance [1].

As computing power has increased, the potential for ESs to be incorporated in an organization's decision making has also increased. ESs now assist organizations to make sound investing decisions, effective financial forecasts, and accurate medical diagnoses. Indeed, some tasks that humans undertake would be extremely difficult or impossible without ESs.

One proposed application of an ES is in the realm of deception detection. Successful deception detection is at the heart of effective communication. It allows humans to gauge the validity of the information they receive and thus make correct decisions. Despite the importance of accurately separating truth from deceit, both laypersons and professional lie-catchers perform rather poorly in deception detection [9, 69, 139]. Computer-based tools are being developed that encapsulate expert knowledge and may enhance human ability to detect deception. However, how a human operator will use such tools remains an open question.

1.1 Statement of the Problem

Deception[1] is commonplace in human communication. One study that examined the frequency of deception estimates that people typically lie on average one to two times a day [28]. People most often lie about their feelings, preferences, attitudes, and opinions. Less frequently, people lie about their actions, plans and whereabouts [26]. While most lies are innocent in nature, some lies have enormous consequences. History is replete with men and women who have caused a great deal of harm through skillful deception. Recent efforts to increase national and international security have brought deception and its effects to the forefront of public attention.

Although deception has garnered much recent attention, deception is not a new problem. Researchers have been fascinated with deception and deception detection for centuries. Perhaps one reason is that humans are inherently poor at recognizing deception when it occurs. Numerous studies have noted that people typically identify deception with accuracy only slightly better than chance (approximately 54%) [9, 69, 139].

Coupled with poor accuracy is poor calibration of confidence in deception judgments [25]. Confidence in judgments is important in deception detection as it may affect the attentiveness of the lie-catcher, the lie-catcher's verification efforts, and misallocation of time and resources as erroneous judgments are made [139].

Computer-based decision aids have been forwarded as a method of remedying the problem of low detection accuracy and poorly calibrated confidence in judgments. While

[1] A complete explanation of deception appears in section 3.1

many methods exist to differentiate falsehood from truth, all methods are tied together by one trait: they rely on a human decision maker to make the final judgment. As Burgoon and colleagues [17] stated, deception detection "defies a fully automated solution. A more promising approach is to integrate improved human efforts with automated tools…, the end goal being a system that singles out individuals for further scrutiny in a manner that reduces false positives and false negatives." The centrality of the human decision maker in any deception detection task makes the context of deception detection ideal for studying ESs. ESs have the potential of being extremely beneficial in deception detection by improving detection accuracy and curbing over-confidence; however, the human decision maker must properly integrate the decision aid in the decision-making process in order to attain these benefits.

Joining an ES and a human decision maker to form a system of deception detection opens the door to a host of questions that must be addressed if such a deception detection method is to be useful in a field environment. One of the most critical sets of questions pertains to the impact of the ES on the decision outcomes and the human user: *What effects does a deception detection ES have on the human decision maker, human decision-making strategies, and decision outcomes in a deception detection task?*

1.2 Purpose of Research

Decision aids have a number of characteristics that are important in deception detection. They can be ever-vigilant, handle repetitive and complex tasks [42], and are easily calibrated. However, decision aids require a human user to control what is analyzed, determine what is relevant, and interpret the results. Humans have abilities which are important in deception detection that are extremely difficult for decision aids to reproduce. Humans excel at improvisation and flexibility, recall relevant facts at appropriate times, can synthesize information, and are proficient in judgment [42]. However, humans suffer from decreasing vigilance over time, cognitive laziness [41], and a bias-based deception detection approach [75]. Thus, a marriage of human and computer abilities may provide one solution to the deception detection problem [84, 148].

The purpose of this study is to explore the effects an ES has on deception detection accuracy and user confidence (*decision outcomes*). The effects of the deception detection tool will be explored by presenting findings from repeated detection tasks. Findings from two different populations (novice and professional lie-catchers) will be shared. Further, the effects of training in human-only deception detection will also be discussed as training was thought to improve judgment accuracy and affect how system recommendations were incorporated into final judgments.

Decision strategies that ES users utilize in formulating a final judgment of deception are of particular interest in this study. Each user may, at his or her discretion, incorporate the knowledge, suggestions, and justifications that are provided by an ES. However, the strategies that the user applies are poorly understood. This work will attempt to partially uncover general strategies that trained and untrained novices and professionals use in incorporating ES output in decisions. Such decision strategies are bounded on one hand by a complete abdication of decision making in favor of the ES output and on the other hand by complete disuse in favor of human-only decision making.

Additionally, this study will expand knowledge concerning humans and ES filling complementary roles. The task of deception detection requires numerous capabilities, some of which the ES has and some of which a human user has. The complementary roles work in harmony when each independently arrives at the same conclusion. However, when the roles are in conflict, the resolution of the conflict is poorly understood. This conflict resolution may come about in different ways between experts and novices and between trained and untrained users. Further, the explanation facilities of the ES may be used differently between the groups during the resolution of the conflict.

Finally, this study will address perceptions of the *decision maker* utilizing the ES. Central in this investigation are analysis of user trust of the system, willingness to use the system as a decision aid, and willingness to allow the system to operate without supervision.

1.3 Chapter Organization

To answer questions about decision outcomes, decision strategies, and decision makers in computer-aided deception detection, a multi-methodological research approach has been adopted. Chapter 2 discusses the research approach. Computer-aided deception detection draws upon two critical research areas: deception detection and computer-aided decision making. Chapter 3 reviews these topics and uses them to create a set of hypotheses. Chapter 4 presents a prototype ES that utilizes linguistic and kinesic analysis for unobtrusive deception detection. The system, called the Behavioral Analysis Prototype, is a semi-functional ES that provides explanations and recommendations about the level of deception that is present in a face-to-face conversation it has monitored.

Chapter 5 presents an experiment designed for novice users of the system. The experiment tested the effects of system accuracy and training in deception detection. Chapter 6 presents an additional experiment examining the differences between novice system users and professional system users. Finally, conclusions and future steps are discussed in Chapter 7.

CHAPTER 2

RESEARCH APPROACH

No single research methodology is sufficient for a complete examination of a phenomenon of interest. Every methodology has strengths and weaknesses that must be carefully weighed in deciding on a research approach. Research in computer-aided deception detection is largely new, so much of this research is exploratory in nature. However, this research benefits heavily from decades of research in computer-aided decision making in other domains. Therefore, a balanced, multi-methodological approach has been selected.

This research approach mitigates weaknesses of certain methodologies by combining them with the strengths of others. Further, utilizing multiple research methods maintains necessary flexibility of exploratory research while building upon successful research of the past.

2.1 Multi-methodological Approach

This research adopts the research framework originally outlined by Nunamaker et al. [93], wherein they specify four broad categories of research methodologies necessary for MIS research: theory building, observation, experimentation, and system development. Each methodological category is a necessary component of a complete research approach, and example methodologies for each category are shown in Figure 2-1. Notable in the Nunamaker et al. framework is the inclusion of systems development as a viable method of research. This work considers the IT artifact to be a critical piece of the investigation [93, 96].

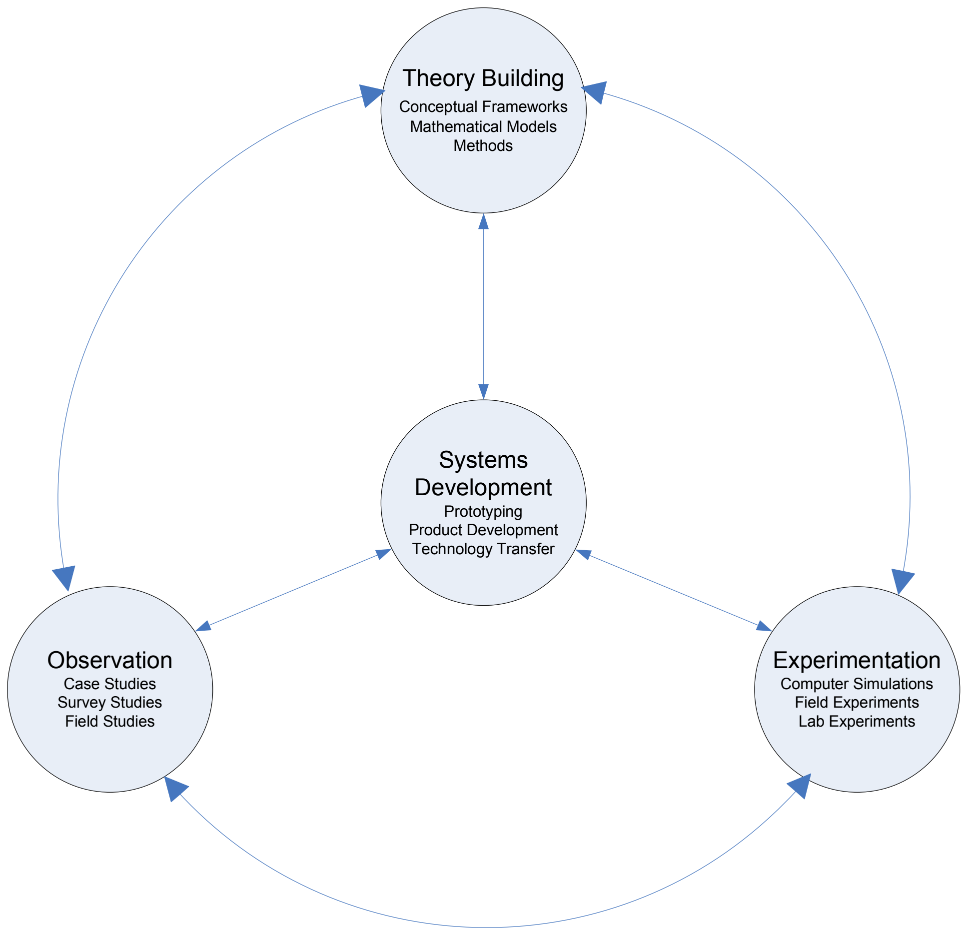


Figure 2-1 Nunamaker et al.'s multi-methodological research approach

2.2 Criteria for Methodologies

Methodologies for this study were selected from the Nunamaker et al. framework based on several factors. First, proper alignment must exist between the research questions and the methodology. Second, the methodology must be able to provide adequate validity and reliability so that conclusions may be drawn from the findings. By building on past research and utilizing methodologies which have provided valid and reliable findings, this can be accomplished. Third, the methodologies must accommodate the study of both actions (as demonstrated by decisions) and perceptions (as provided by self-reports). The study of both perceptions and actions requires methodologies from the experimentation and observation categories. Finally, the methodologies must conform to constraints in resources and support.

2.3 Selected Methodologies

A single methodology was selected from each of the systems development, observation, and experimentation methodological categories. The selected methodologies are: develop a prototype system, administer surveys, and carry out laboratory experiments. These methodologies will be guided by theoretical underpinnings described in Chapter 3 and satisfy all of the criteria suggested in section 2.2 These methodologies and their respective categories and characteristics are shown in Table 2-1.

Table 2-1 Selected methodologies in research approach

Methodology	Category	Characteristics
Prototype	Systems Development	Demonstration of feasibility for technology Provides foundation and context for other methodologies Not fully operational system
Surveys	Observation	Acquisition of a general understanding of core concepts Inexpensive and unobtrusive Provides insight into subject perceptions and opinions
Laboratory Experiment	Experimentation	Controlled test of specific treatments Rigorous examination of theory and observations Provides insight into subject actions and decisions

2.3.1 Laboratory Experiment

The primary method of inquiry in this research is experimentation. Laboratory experiments are approximations of reality when researchers can explore specific items of interest while controlling factors that may cloud the relationships between the items of interest. The experiments described here were carefully controlled to provide valid results concerning the use and influence of a computer aid in a deception detection task. The use of laboratory experiments allows the researcher to examine the actions a decision maker takes (not just the perceptions or expected actions). Additionally, experimentation allows rigorous testing of the relationships stipulated in a theoretical model through control and manipulation of the experimental conditions.

Although tight control is a benefit in experimental methodologies, one may criticize control as limiting the ecological validity of the findings. No experiment can fully replicate the factors present in a real scenario, and experimental research in decision making has been singled out as unrepresentative of actual decision making [95]. However, to analyze the effects of selected factors of interest, the remaining factors must be held constant for purposes of comparison. The experimental decision-making environment described here represents the most realistic scenario possible given the constraints described in section 2.2 .

2.3.2 Surveys

Accompanying the laboratory experiment are surveys designed to capture the opinions and perceptions of the computer aid users. Surveys are an accepted method to gather information about attitudes, perceptions, and opinions from individuals in a population of interest. Surveys are largely unobtrusive and fairly inexpensive to administer. Further, they can be easily adapted to explore new phenomena of interest.

Criticism of survey use in IS research has arisen surrounding validity and reliability that has questioned the amount of rigor in survey-based research [122]. To address criticisms in survey research, this work follows steps outlined by Grover et al. [52]. These steps include: 1. Report approach to select samples; 2. Report a profile of the target population; 3. Report characteristics of respondents; 4. Append questionnaire; 5. Provide validity and reliability measures; and 6. Pretest instrument.

In this research a series of surveys were created, validated, and administered to users of the deception detection computer aid. Two populations (novice and professional lie-catchers) were targeted and recruited to participate in the experiment and surveys. The surveys were utilized to gain a broad understanding of the attitudes of both populations and are meant to be used in conjunction with the experimental results to provide a more complete understanding of the computer-aided deception detection.

2.3.3 Prototype

At the foundation of much of IS research is the IT artifact or system [5, 93, 96]. A great deal of knowledge can be gained from analysis, design, and construction of a system that fulfills a current need [57]. Among the most valuable knowledge that is provided by system development is feasibility of the proposed system that amounts to proof by demonstration [93].

Prototyping is the creation of a novel, partially functional system that provides the developer and future user a chance to examine how the system may perform when complete. With the system in the prototype stage, the developer and future user may explore the functionality, interface, and, perhaps most importantly, potential usability of the proposed system. For this research, a prototype system was constructed with limited functionality. Chapter 4 describes the instantiation and implementation of the system.

While the prototype system is not intended to be a central contribution of this research, it is a demonstration of capabilities of a future, fully functioning system to detect deception automatically and unobtrusively. The prototype system also provides a large portion of the context in which the decision-making task occurred. Therefore, this work gives a detailed description of the development and capabilities of the system.

CHAPTER 3

LITERATURE REVIEW

Many streams of research are relevant in computer-aided deception detection. This chapter reviews the two most critical streams of research: deception detection and computer-aided decision making. Each of these research areas is reviewed separately, then concepts common to both research areas are reviewed, as shown in Figure 3-1.

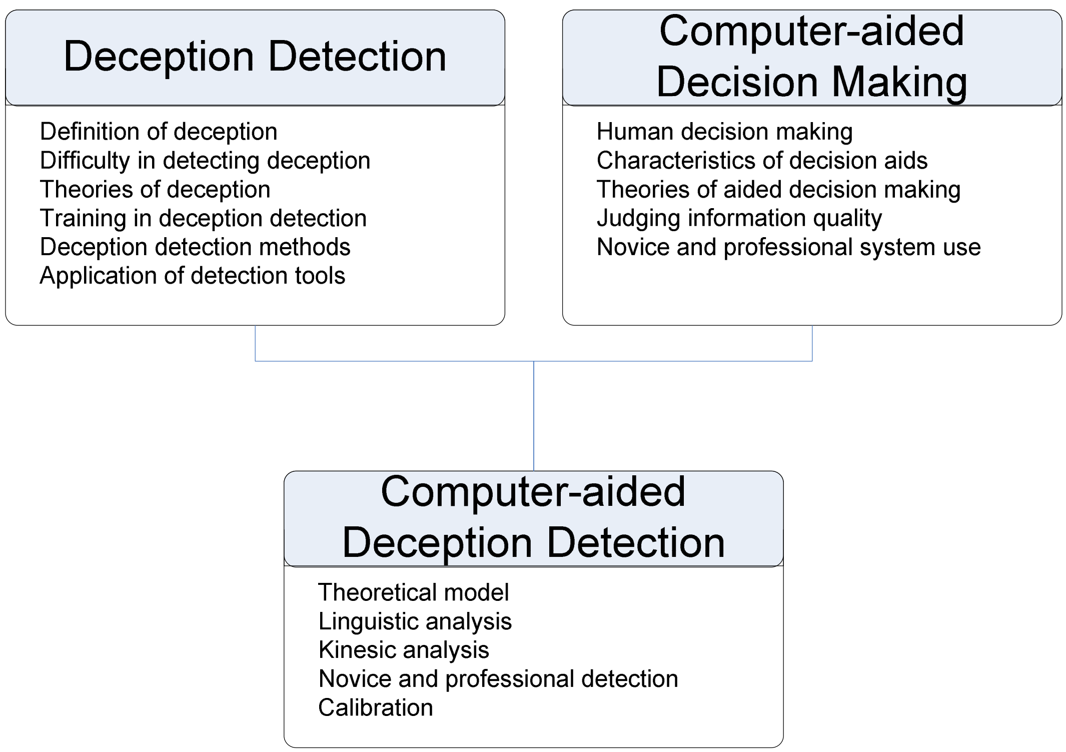


Figure 3-1 Relevant literature map

The deception detection review begins by exploring extant knowledge about deception and the difficulties humans face in detecting it. The review concerning deception detection continues with summaries of relevant theories of deception, effects of training in deception detection, detection methods and their usefulness in a field environment.

The review regarding computer-aided decision making presents relevant work concerning human decision making, characteristics and effects of existing decision aids, theories of computer-aided decision making, judging information quality, and novice and professional ES use.

Finally, findings from both streams of literature are synthesized to highlight relevant concepts pertaining to computer-aided deception detection. Such topics include a theoretical model of computer-aided deception detection, linguistic and kinesic analysis, professional and novice deception detection, and proper calibration of judgments. Hypotheses will then be presented which build on these concepts.

3.1 Deception Detection

3.1.1 Definition of Deception

Deception is commonplace in human communication. One study examining the frequency of deception estimated that people typically lie on average one to two times a day [28]. People most often lie about their feelings, preferences, attitudes and opinions. Less frequently, people lie about their actions, plans and whereabouts [26].

Before any investigation of deception detection can occur, a firm definition of deception must be in place. One widely-held view of deception is that it is synonymous with outright falsification. While falsification can certainly be classified as deception, people may also mislead others by utilizing alternative strategies such as concealment and ambiguity [13, 18]. Therefore, this definition is unnecessarily stringent and does not represent the breadth of ways people deceive [19]. By using other strategies to mislead, the deceiver effectively mixes truth with fabrication or ambiguity in a manner that resembles truthful communication.

The definition of deception used in this work is “a message knowingly transmitted with the intent to foster false beliefs or conclusions” [14]. This relaxed definition may include such deceptive strategies as half-truths, equivocation, concealments, and omissions. Other methods of deceiving can be found in Table 3-1. These methods of deception are more representative of how people actually deceive.

Table 3-1 List of methods of deception

Methods of Deception	
Lies	Evasions
Fabrications	Equivocation
Concealments	Exaggerations
Omissions	Camouflage
Misdirection	Strategic ambiguity
Bluffs	Hoaxes
Tall tales	Charades
White lies	Sophistry

The methods listed in Table 3-1 are not mutually exclusive so a deceiver may adapt his or her method of deception during an interaction depending on the observed success of the deception [19].

The deceiver’s intention is also part of deception’s definition. Deception is a purposeful act that must be undertaken consciously. Misinformation perpetuated by mistake or with innocent intentions is not deception. Conversely, truthful messages that are arranged to form a false belief in the receiver are classified as deception [26].

3.1.2 Difficulty in Detecting Deception

Researchers have been fascinated with deception and with deception detection for centuries. Perhaps one reason that deception and deception detection have attracted so much attention is that humans are inherently poor at recognizing deception when it occurs. Numerous studies note that people typically identify deception with accuracy only slightly better than chance [9, 69, 139]. Most recent estimates of human deception detection ability indicate that humans correctly identify truth or deception approximately 54% of the time (61% correct classification of truth and 47% correct classification of deception) [9]. This poor performance is not limited to laypersons, but is also found in professional lie-catchers such as police officers and federal law enforcement officers [29, 77, 139].

Several possible explanations exist for this consistent inaccuracy in judging deception. One is the well-documented phenomenon called the *truth bias* where persons engaged in communication will systematically assume that what they are receiving is truthful [75, 80, 86]. Some researchers believe that the truth bias is a manifestation of various decision heuristics such as the *availability heuristic* [9, 94, 132]. In daily interactions, the vast majority of human communication is truthful. Thus to save cognitive effort, all communication is deemed truthful. However, in situations when humans are exposed to large numbers of lies a contrasting bias has been observed. This bias has been termed the *lie bias* or *Othello bias* [33].

Another explanation for poor judgment accuracy is mistaken reliance on behaviors or cues that do not distinguish between truth and falsehood [139]. Most people believe that behaviors can differentiate truth-tellers from deceivers; however, people rely on behaviors such as gaze aversion and persistent self-touching, which have been shown to be unreliable when separating truth from falsehood [125, 139]. Although certain behaviors statistically differ between truth-tellers and deceivers, these behaviors are largely unknown or ignored [139].

Finally, it has been argued that humans do not detect lies from behavior observed during an interaction. Rather, humans rely on information from third parties or from physical evidence in judging truth or deception [99]. Under such conditions, the process of validating information received from a sender would not occur at the time of the interaction, but would occur in retrospect days, weeks, or months later [99].

Not only do humans demonstrate poor accuracy in detecting deception, they also seem to overestimate their ability to detect deception [25]. Overconfidence is frequently encountered when studying human decision making under uncertainty [109, 110]. However, the level of confidence is of considerable importance in deception detection as it directly affects the attentiveness of the lie-catcher, the lie-catcher's verification efforts, and misallocation of time and resources as erroneous judgments are made. Confidence levels vary a great deal among untrained, non-professional lie-catchers (e.g., student experimental subjects) [25]. The confidence level increases dramatically for those with some training in deception detection or who are employed to identify deceit; however, this increase is not accompanied with a corresponding increase in detection accuracy [29, 62, 77, 139]. In complex decision making other than deception detection, extensive experience and training calibrated confidence in performance [70]; however, the exact relationship between deception detection accuracy and confidence is still under examination. Research has shown that there is very little correlation between deception detection ability and confidence in that ability [25, 29].

3.1.3 Theories of Deception

With a broad definition of deception established, relevant theories regarding the act of deceiving can be effectively reviewed. Such theories attempt to explain the internal and external forces at work when deceivers engage in deception and, perhaps more importantly, how these forces manifest themselves in the form of behavior. These theories provide the guiding foundation for methods of deception detection. Theorizing about deception has migrated from a non-strategic view to a more strategic view. As both perspectives are critical to understanding deception detection, they will both be addressed. Relevant theories that will be reviewed include: Leakage Theory [34], Zuckerman's four factor model of deception [152], the Self-Presentational view [26], and Interpersonal Deception Theory [14]. All of these theories have non-strategic and strategic components; however, they may be productively categorized on a non-strategic-strategic continuum as shown in Figure 3-2.

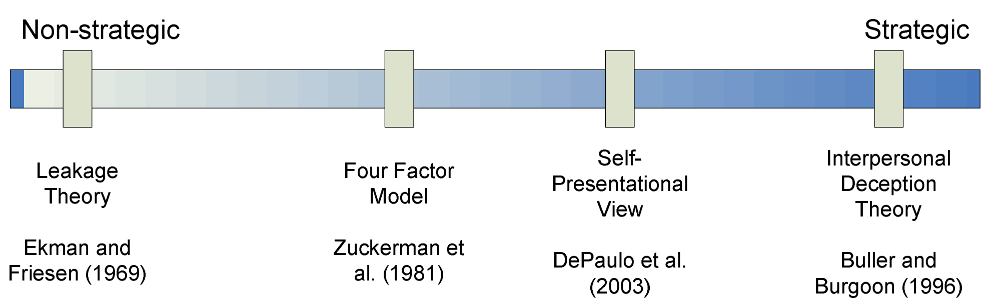


Figure 3-2 Four theories of deception on a non-strategic-strategic continuum

Sigmund 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” [44]. This characterization of deception effectively encapsulates the key tenants of Leakage Theory: deceivers cannot help “leaking” behavioral cues that unmask the deception. Under this view of deception, researchers scoured deceptive discourse searching for tell-tale signs of deceit.

Among those who first attempted to systematically analyze deceptive interactions were Ekman and Friesen [34]. They theorized that deception manifests itself through two categories of behavioral cues: leakage cues and deception cues. They suggested that leakage cues occurred when deceivers tried to conceal true emotions and reactions during deception. This concept laid the foundation for the investigation of micromomentary facial expressions as a sign of deceit. For example, when questioned, a deceiver may leak a fleeting facial expression showing fear before composure is regained. Deceptive cues concern behaviors that deception may be going on. For example, the “lack of the usual illustrative hand movements” was forwarded as a sign of deception [34].

In a departure from Leakage Theory, Zuckerman and colleagues [152] refute the notion that deception is associated with a single reliable, diagnostic cue. Later research substantiated this finding [33]. In their four factors model, they argue that the act of deceiving causes observable, behavioral changes in four broad areas: arousal, emotion, cognitive effort, and attempted control of behavior. Zuckerman et al. suggest that elevated arousal is positively correlated with deception; deceivers display signs of fear or guilt; deceivers show greater cognitive load; and deceivers’ efforts to maintain an honest demeanor squelch other behavior that is normally seen in honest communication. Such attempted control may appear “planned, rehearsed, and lacking spontaneity” [152].

The Self-Presentational view [26] of deception utilizes the “deception discrepancy” as its foundation: what liars believe differs from what liars claim. Despite the deception discrepancy, the deceiver must present a convincing impression on the receiver. From this discrepancy, five categories of cues are predicted. First, liars are predicted to be less forthcoming than truth tellers; second, accounts from liars are predicted to be less compelling than truthful accounts; third, deceivers will be less positive and pleasant than truth tellers; fourth, deceivers will be more tense than truth tellers; fifth, deceptive accounts will contain fewer ordinary details and unusual contents than truthful accounts [26]. Under the Self-Presentational view special significance is given to the motivation of the lie as a critical moderator of behavioral displays associated with deception.

The most strategic view of deception comes from Buller and Burgoon [14], who consider deception through the lens of interpersonal communication. Interpersonal Deception Theory (IDT) highlights the dynamics of the interaction and underscores the deceiver’s multiple roles as he or she must monitor, interpret and adapt deceptive messages according to their perceived effectiveness. A key tenet of IDT is that the differences between truthful and deceptive displays will fade over time as the deceiver is able to monitor the deception’s effectiveness and tailor the message to the context and receiver. IDT predicts that deceivers will initially display decreased involvement, immediacy, pleasantness, elevated cognitive effort and over control. However, these disparities shrink through the interaction. Also critical in IDT are the goals and motivations behind the deception [14].

3.1.4 Training to Detect Deception

Many researchers are cautiously optimistic about improving deception detection through training [27, 43]. However, doubt currently surrounds the amount of improvement that is possible through training [12]. There are numerous explanations for the ambiguity surrounding the effects of training in deception detection. First, there are no tell-tale signs of deceit [152] and training individuals to look for a small set of deceptive cues may be counter-productive. Behavioral cues appear probabilistically with deception and must not unduly influence the judgment when they are observed [27]. Second, much of the uncertainty about the effects of training can be attributed to inconsistent and poorly constructed studies [43]. Frank and Feeley [43] specify six criteria that need to be fulfilled for a valid investigation of training in a credibility assessment task. They note that few of the past studies in deception detection training have satisfied these criteria. These six criteria are shown Table 3-2.

Table 3-2 Criteria of effective training

Relevance – Training must be relevant to what subjects usually encounter
High Stakes – Lies must be consequential
Proper Training – Material must be scientifically based and effectively taught
Proper Testing – Proper assessment of training effects
Generalizable across situations – Training must be applicable to other situations
Generalizable across time – Training must be applicable to other times

More recent research indicates that training has a positive, significant effect on deception detection ability [43, 139]. In addition to live lecture, computer-based training in deception detection is also effective in improving detection skills [47, 48].

Finally, some researchers also believe that current research in lie-detection actually underestimates the ability of human lie-detectors [43]. Research has highlighted the abilities of some humans to detect deception with the possible implication being that the techniques these humans use may be replicated by others [35].

3.1.5 Deception Detection Methods[2]

Many methods of deception detection currently exist and some are regularly utilized by professional lie-catchers such as police and security professionals. Additionally, many new methods of deception detection have been proposed, and their levels of effectiveness and reliability are currently being studied. The methods of deception detection can be divided into two broad categories: physiological and behavioral. Methods belonging to each category appear in Table 3-3 and brief descriptions of each method appear below.

Table 3-3 Methods of deception detection

Physiological Methods	Behavioral Methods
Polygraph	Statement Validity Assessment
Brain Activity Analysis	Linguistic Analysis
Thermal Analysis	Behavioral Analysis
Voice Stress Analysis	

[2] An extended version of section 3.1.5 was originally published in J. K. Burgoon, M. L. Jensen, J. Kruse, T. O. Meservy, and J. Jay F. Nunamaker, "Deception and intention detection," in Handbooks in Information Systems, vol. 2, H. Chen, T. S. Raghu, R. Ramesh, A. Vinze, and D. Zeng, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2007.

Physiological methods are premised on the assumption that arousal, emotions, and cognitive changes associated with deception generate systematic physiological changes in blood flow, hemo-oxygenation, neuronal activity and the like [139]. Due to the tight coupling of these systems, technologies that tap into one or more physiological processes may accurately discriminate truthful from deceptive communication. Out of all the approaches to deception detection, perhaps the most recognized is the polygraph or “lie-detector.” The basic assumption behind the polygraph is that deception causes an increase in arousal stemming from feelings of guilt or fear of the consequences of being caught in a lie. The polygraph detects the arousal via sensors attached to the body that measure heart rate, palmar sweat, and respiratory features [139]. There are two main methods of interviewing which use the polygraph: the Control Question Test (CQT) and the Guilty Knowledge Test (GKT)[3]. In the CQT, the level of arousal from control and neutral questions is used in a comparison to crime related questions [139]. Field-tested accuracy rates for the CQT range from 84-92% in the guilty condition and 30-78% in the innocent condition [139]. Although the CQT test is commonly used in the United States, it has often been criticized as subjective, non-scientific, and unreliable [2, 37]. The GKT determines whether an interviewee has knowledge about a crime that would only be known to the perpetrator. Accuracy rates based on field tests for the GKT range from 71-98% for the guilty condition and 73-91% for the innocent condition [139]. The GKT

[3] The Guilty Knowledge Test has also been termed the Concealed Information Test.

enjoys a more objective, scientific footing [2], however specific and confidential details about a crime must be obtained for its use.

Other emergent methods intended to augment the polygraph rely on analysis of brain activity. One of the first methods utilizes an electroencephalogram (EEG) to measure event related brain potentials (ERPs). Experimentally and in criminal applications, this method has yielded accuracy approaching the polygraph [38]. In addition to the EEG, functional magnetic resonance imaging (fMRI) has been used to differentiate between real and imagined events [45]. Currently the reliability and accuracy of deception detection based on the fMRI is being debated. A noninvasive alternative that is now being investigated is near infrared spectroscopy (NIRS) [137]. NIRS does not require people to remain stationary and uses optical technology to measure neuronal, metabolic, and hemodynamic changes that may indicate arousal associated with deception.

Another alternative that measures peripheral cardiac activity is high-definition thermal imaging [101, 102]. This method is also based on the premise that deception increases the level of arousal. Proponents of the thermal imaging method suggest that arousal is shown by an instantaneous warming pattern around the eyes.

Voice stress analysis (VSA) detects the indicators of psychological stress in the voice and from these indicators infers a level of deception. Numerous studies have concluded that deceivers exhibit an elevation in voice pitch when deceiving [139]. Accuracy of VSA is comparable to the polygraph [126].

The purpose of behavioral methods is similar to the physiological methods: search for arousal, emotions, and cognitive changes that may accompany deception. In addition, behavioral methods of deception detection also search for strategies deceivers may use to mislead. First among behavioral methods is statement validity assessment (SVA). SVA is a general category of deception detection which focuses on verbal content. Two common methods of SVA are Criteria-Based Content Analysis (CBCA) and Reality Monitoring (RM). CBCA is based on the Undeutsch-Hypothesis which states that 'a statement derived from a memory of an actual experience differs in content and quality from a statement based on invention or fantasy' [136]. CBCA takes place during a structured interview where the interviewer scores responses according to predefined criteria such as general characteristics, specific contents, motivation related contents, and offence related elements. CBCA has been used successfully in judging the validity of statements given by children and it has been used in criminal cases where children are involved [139].

RM also uses a scoring mechanism to judge potential deception. However, it is based on the hypothesis that verbal recall of actual events will contain more perceptual, contextual, and affective information than recall of fabricated events. Reality monitoring requires the interviewer to judge levels of clarity, perceptual information, spatial information, temporal information, affect, reconstructionability of the story, realism, and cognitive operations [119].

In contrast to CBCA and RM, linguistic analysis operates independently of message meaning. It is based on a similar premise: that fabricated messages structurally differ from truthful messages. Two areas in linguistic analysis are message feature mining and speech act profiling. Message feature mining uses features extracted from text to determine the likelihood of the truthfulness of the message. Categories of features include quantity, complexity, uncertainty, nonimmediacy, expressivity, diversity, specificity, and affect [148-150]. Speech act profiling is a method of conversation classification and visualization. This method compares the communicative style of the interactants so that deceptive strategies such as equivocation and indecisiveness can be seen clearly [134].

Finally, observation of behavioral cues is also used as a method of deception detection. Numerous studies show that deceivers behave differently than truth tellers [13, 15, 18, 26, 139, 152, 153]. Behavioral cues that may indicate deception may be categorized into groups of non-strategic and strategic cues [14, 19]. The strategic category includes behavior such as information management and image management. The non-strategic category includes behavioral cues concerning noninvolvement and arousal [19]. Cues dealing with micro-momentary facial expressions [33], motion and gesture, and posture all fall into behavioral analysis.

3.1.6 Application of Deception Detection Methods

The applications of a deception detection tool or system are many and varied. For a method of deception detection to be feasible in screening it must be usable in a natural setting. This necessitates unobtrusive instrumentation, a prompt judgment of suspicion, and robustness under varying conditions [20]. For example, the sole use of a polygraph in every credibility assessment would be infeasible as it requires each interviewee to be attached to body sensors for an extended amount of time. Table 3-4 displays a characterization of each deception detection method according to requirements on the environment [20].

Table 3-4 Characterization of environmental constraints of detection methods

Controlled	Semi-Controlled	Natural
Polygraph	Statement Validity Assessment	Thermal Analysis
Brain Activity Analysis	Near Infrared Spectroscopy	Vocal Analysis
	Micro-Momentary Expressions	Linguistic Analysis
		Behavioral Analysis

3.2 Computer-aided Decision Making

The promise of computer aids has captured the imagination of human inventors through numerous decades. Computer aids may have the potential to free humans from mundane and tedious tasks, increase human productivity, and minimize human error. Computer aids appear in innumerable applications as varied as automobile manufacturing, airplane piloting, car navigation, corporate payroll management, and nuclear reactor management [4, 56, 117]. In fact, some researchers suggest that intelligent agents may be employed in almost any environment imaginable [108 pg. 32].

To examine current knowledge related to computer-aided decision making, key concepts in human-only decision making are first reviewed. Next, characteristics of decision aids are explored along with mechanisms by which they assist in decision making. Relevant theories of computer-aided decision making are then be shared followed by a discussion on judging data quality. Finally, past research involving novice and professional computer-aided decision making is reviewed.

3.2.1 Human Decision Making

Decision makers in natural settings are challenged on many fronts. Orasanu and Connelly [95] suggest several difficulties that humans must navigate during consequential decision making (reproduced in Table 3-5).

Table 3-5 Factors complicating decisions

Ill-structured problems
Uncertain dynamic environments
Shifting, ill-defined, competing goals
Action/feedback loops
Time stress
High stakes
Multiple players
Organizational goals and norms

In considering each of these challenges, the decision maker must effectively weigh the consequences and trade-offs in some rational way to decide on courses of action. However, as Newell and Simon [91] note, there are individual limits to the processing capabilities of every decision maker. Thus, decision makers operate within the confines of *bounded rationality* [115]. Simon further notes that humans tend to "make do" rather than seek the best solution [114]. Simon terms this tendency *satisficing.* At the core, the design and intent behind the majority of computer-based decision aids are to extend the limits of the decision maker's rationality and reasoning power [129].

3.2.2 Characteristics of Decision Aids

Much of the current knowledge concerning computer aids has roots in the work done by Simon, Newell and their colleagues. Their Logic Theorist [89] and, later, their General Problem Solver (GPS) [90, 91] represent some of the first attempts at developing computer aids. Using their GPS, Newell and Simon examine not only the outcomes of the aid (quality of the results), but also compare the decision processes of the aid with that of acknowledged experts of their time [91, 116]. Other important contributions to early work in computer-aided decision making are Gelernter's Geometry Theorem Prover [46] and Samuel's checker-playing program [111].

Since Simon and Newell's GPS, the field of decision aids has expanded to include various types of computer-based tools. Most decision aids can be classified into three groups: automation, decision support systems, and expert systems. Each category draws upon distinct philosophies and theories and has a distinct purpose; however, all decision aids share the goal of "improving the effectiveness of the decision maker and the quality of decisions that the decision maker makes" [129]. To understand these categories, a framework first developed by Klein and Methlie [67] has been adapted and expanded. This framework appears in Figure 3-3.

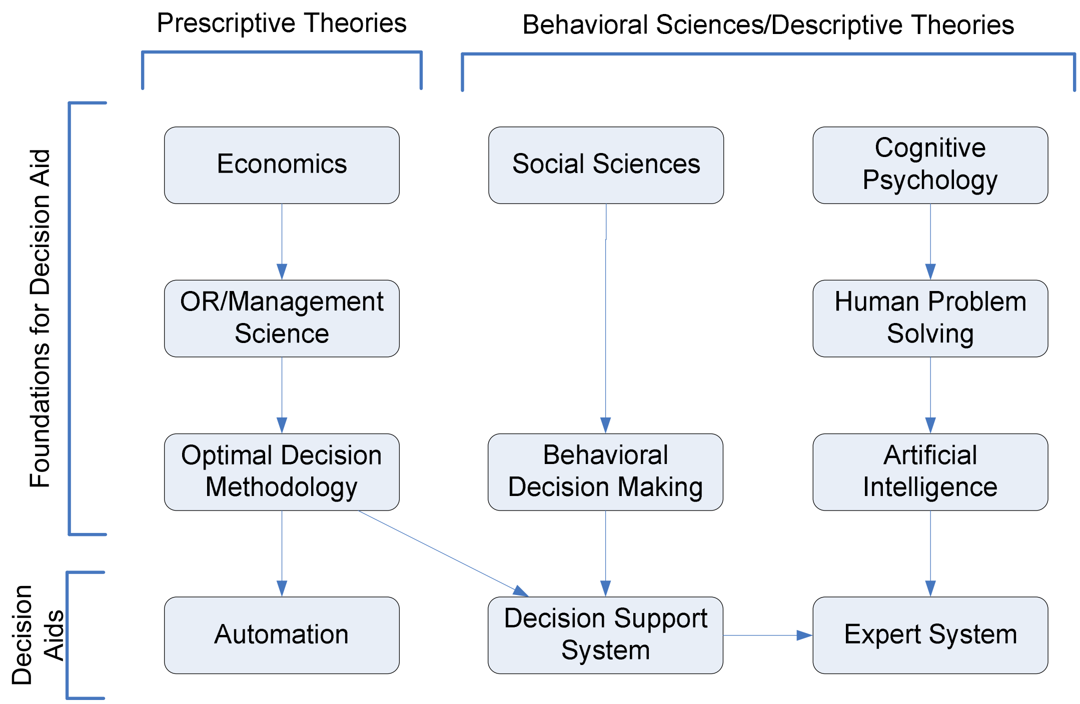


Figure 3-3 A framework of decision aids and their foundations

The framework organizes decision aids based on their foundational theories and decision approaches. The first method of decision support is pure automation. This type of decision aid is based entirely on normative decision making and its goal is to remove the human decision maker. In its purest state, automation performs all steps of decision making including the implementation of the decision. This may be appropriate for tasks that require a great deal of speed in execution or that are well defined and repetitive. Automation may not be a binary classification where a task is fully automated or completely manual. Sheridan and colleagues [98, 112] propose a classification continuum to judge levels of automation where complete automation and complete manual operation are opposing endpoints on a scale of automation. Sheridan at al.'s levels of automation are reproduced in Figure 3-4.

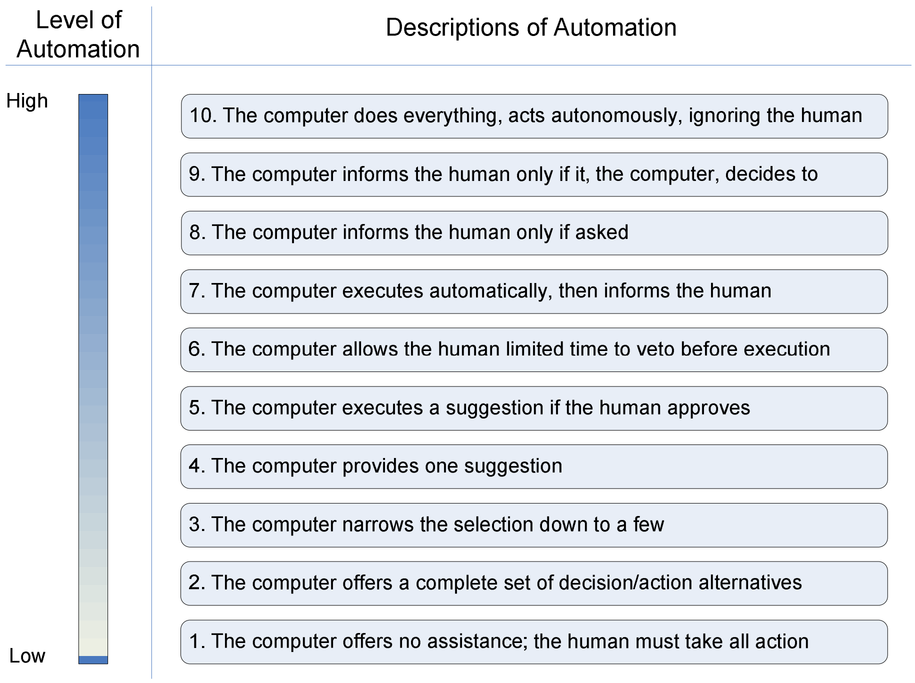


Figure 3-4 Sheridan at al.'s levels of automation

A decision support system (DSS) rests on descriptive theories of decision making and seeks to engage the user in a dynamic supportive environment [8, 120, 121]. There are diverging views concerning the role of a DSS. One view is that a DSS does not provide any recommendations but supports individual decision making by providing a malleable environment in which the human decision maker can investigate various alternatives and the effects those alternatives may have [67]. Such a view favors a non-directing DSS role and provides the environment, data, and tools to solve a problem [6, 63]. A significant benefit of this view is the decoupling of the DSS from the problem

[121]. Such decoupling allows users to adapt to the problems they face. In contrast, a directing DSS role requires closer association between the problem and DSS. With tighter association, the DSS can provide more guidance when investigating a problem and may also restrict the types of analyses to a relevant subset [113]. Restrictions may come in the form of access and sequencing of analysis components. Guidance may come in informative and suggestive direction [113].

Finally, the expert system (ES) combines analytic aspects of a DSS with that of knowledge base to produce a recommendation to a specific problem. Some consider an ES to be a special case of a DSS [130]. The most critical component in the ES is the knowledge base which is captured from individuals with acknowledged expertise in solving a particular problem.

The purpose of the ES is to mimic the reasoning abilities of trained and experienced human experts and make this capability available to other, less-skilled individuals [56]. This reasoning ability is typically performed using a rule-based approach (heuristics) (e.g., [55]) or inferential learning techniques from artificial intelligence where the thought processes of the human expert are captured for reuse (e.g., [21]). While some have noted the failure of ESs to achieve widespread adoption [49], many organizations have seen significant improvements in complex task accomplishment through ES use [56].

As the functionality of the computer-based decision aids vary, so do the types of problems for which they are suited (see Table 3-6). Automation is suited for well-defined problems that are static and structured. Conversely, DSS and ES tools are designed for poorly specified problems that are dynamic and context sensitive [3, 67, 120, 121].

The primary responsibilities of the human user also vary with the functionality of the systems. These responsibilities are summarized in Table 3-6. The role of the human user when working with high levels of automation changes from that of participant in decision making to that of supervisor of the decision maker [98]. This is in contrast with the participatory roles of the human user with the DSS and ES. These two types of systems support human decision making and do not supplant it.

Table 3-6 Problems suitable for decision aids and responsibilities of users

	Automation	Decision Support Systems	Expert Systems
Problems addressed	Structured problems	Structureable and Unstructured Problems	Structureable and Unstructured Problems
Role of Human Decision Maker	Supervise info. collection Supervise info. processing Supervise implementation	Design problem model Identify relevant info. Collect info. Generate alternatives Select alternative Implement alternative	Supervise info. collection Review system explanations Reconcile inconsistencies Select alternative Implement alternative

3.2.3 Theories of Computer-aided Decision Making

By focusing attention on participatory decision aids (DSS and ES), one can begin to explore the essential elements that contribute to successful computer-aided decision making. A traditional view of the effects of DSS and ES may suggest a direct link between the system capabilities and system performance [129]. Such a view considers the functionality of the system and its important role in improving decision making performance. However, a direct link neglects many significant factors that also influence decision performance (such as task-technology fit) [50, 151].

Todd and Benbasat [129] present a model of key factors that influence the impact of DSS on decision performance. In addition to the system's capabilities, they posit examining the effects of the decision task, decision maker, desired effort, desired accuracy, and decision strategy. A reproduction of their model is presented in Figure 3-5.

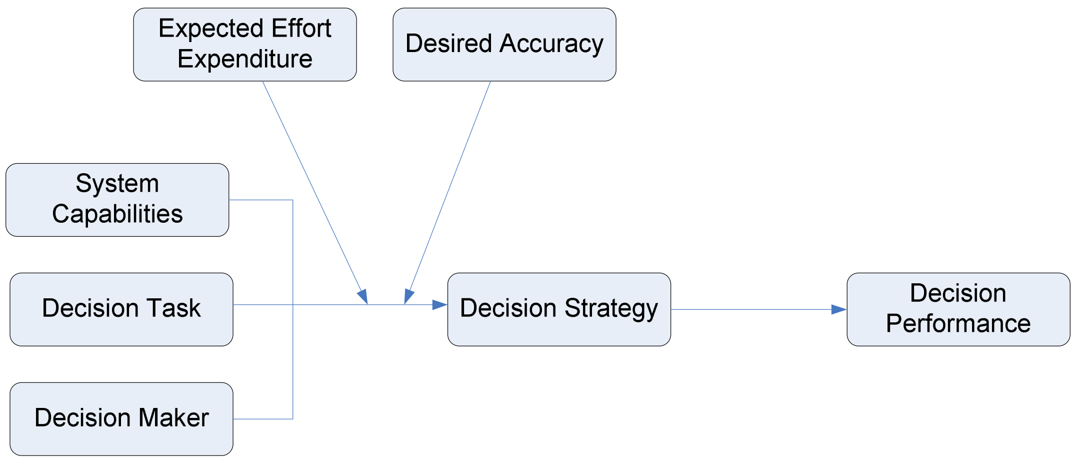


Figure 3-5 Todd and Benbasat's model of factors in DSS

From the early days of DSS, qualities of the decision task and the alignment between the technological capabilities of the system and requirements of the task were suggested as contributing factors to improving computer-aided decision performance [129]. Later, researchers introduced what is now known as *task-technology fit*, the level or degree to which the technology or decision aid is suited to the task [50]. If there is greater alignment between the technology and the task, decision performance is enhanced.

Qualities of the decision maker may also be influential on decision performance. Researchers have suggested that individual differences (such as cognitive style) may have an effect on decision performance [23, 120]; however, this suggestion was questioned on the grounds of relevance for implementing DSSs and ESs in organizations [58]. Despite this challenge, expertise of the decision maker has remained an important factor in effective computer-aided decision making as experts have more sophisticated mental models of problems and can find root causes more easily than novices [95].

According to Todd and Benbasat [129], decision strategy fully mediates the relationship between decision outcomes and the system capabilities, decision task, and decision maker. There are numerous decision strategies that may be employed to form a decision [103], and humans can and do alternate between decision strategies. In multiple criteria decision making, sample strategies include the equal weight heuristic where all criteria are equally weighted [32], weighted additive rule where criteria are weighted based on subjective importance, and elimination by aspect where alternatives are eliminated from consideration based on a single criterion [131].

Todd and Benbasat also claim that desired accuracy and desired expended effort moderate the relationships between decision strategy, system capabilities, decision task, and decision maker. Expected effort and desired accuracy are anticipatory measures formed by the user ahead of any decision making and heavily influence the types of decision processes that are used in arriving at a final decision [129]. Past research has shown that there is an inverse relationship between these cognitive constructs (i.e., most desire low effort and high accuracy) and that desired effort seems to be the dominant factor at the expense of accuracy [103]. Todd and Benbasat argue that decision makers use decision aids as a means to reduce effort expenditure and not sacrifice accuracy or alter their decision strategy [127-129].

As ESs utilize knowledge garnered from human experts, they have the unique ability to provide explanations about their conclusions and recommendations. Explanations typically involve the reasoning (trace line), justification, and strategic implications of a conclusion [1, 51, 147]. Researchers have suggested that the provision of context-relevant explanations has a positive effect on perceptions of the ES, ES use, learning, and accurate decision making [30]. In an attempt to specify the relevant factors involved with ES use, Gregor and Benbasat [51] have developed a theoretical model that incorporates explanations. A reproduction of this model is shown in Figure 3-6.

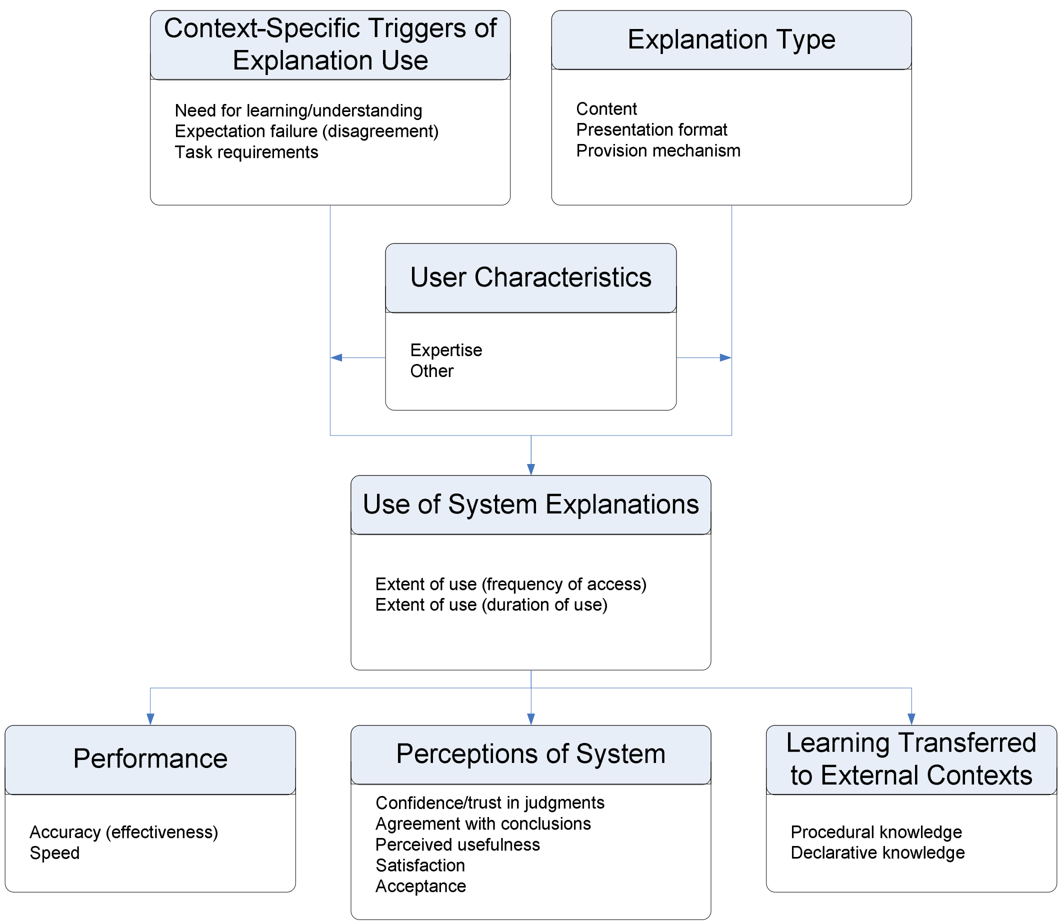


Figure 3-6 Gregor and Benbasat’s model of ES explanation use

Although the central phenomenon of both the Todd and Benbasat and the Gregor and Benbasat models is performance, there are significant differences. Both incorporate user characteristics such as expertise. However, the Gregor and Benbasat model focuses on explanation use as one of the key drivers of performance. In addition, they posit that explanation use may also lead to other beneficial outcomes such as learning and positive perceptions of the ES.

The model highlights the importance of context-specific triggers that need explanation. Such triggers may involve disagreement between the user's expectations and ES results [51], perceived anomalies in the ES [78], and a recognized need for learning [51]. The context-specific nature of the triggers emphasizes the tight coupling between ES and context; therefore, the task-technology fit that is explicit in the Todd and Benbasat model is also implied in this model.

Empirical evidence supports the link between explanation use, ES use, and performance improvements [1, 51, 78, 147]. Explanations benefit both novices and professionals in improving their decision making [1]. The provision of justification-type explanations demonstrate particular value [147].

3.2.4 Judging Information Quality

Context-specific triggers for explanation use are opportunities for the users to gauge the quality of the information the ES provides. While information quality can be partially judged by ES-generated explanations, other indications of information quality also exist. Understanding how humans judge the quality of the information they receive via a computer-based tool is a necessary prerequisite to improving decision outcomes [40].

Nelson et al. [88], suggest a two-dimensional categorization of antecedents to human judgments of information quality that concerns not only perceptions of the information itself, but also perceptions of the quality of the system providing the information. A reproduction of their categorization is shown in Table 3-7.

Table 3-7 Nelson et al.'s antecedents of information quality

Information Quality		System Quality	
Dimension	Definition	Dimension	Definition
Accuracy	Info. is correct, unambiguous, meaningful, believable	Accessibility	System accessible with low effort
Completeness	All possible relevant states are represented	Reliability	System dependable, available over time
Currency	Info. is up-to-date; Info. reflects current state of world	Response Time	System offers timely responses to requests for information
Format	Info. is presented in understandable, interpretable manner	Flexibility	System can adapt to variety of user needs and changing conditions
		Integration	System facilitates combination of info. from various sources.

Nelson et al. tested their proposed model within the data warehousing domain and noted that among the information quality dimensions, accuracy is most dominant, followed by completeness and format. Among the system quality dimensions, reliability is most dominant, followed closely by accessibility [88].

There is strong evidence that humans have difficulty identifying errors in data [66]. To address this issue, researchers have explored ways of providing meta-data called data quality information (DQI) that provides insight to the user about the quality of that data [40]. DQI differs from explanations provided by an ES in that DQI provides evaluations of the data being examined. Explanations typically concern the process by which the data is being examined [51].

DQI is often expensive to collect and maintain but can be very useful to individuals who are not intimately familiar with the applicable domain [22, 40]. DQI is typically presented in a symbolic manner to represent the credibility of the data being analyzed. For example, in a study conducted by Chengalur-Smith et al., DQI was represented by categorical (above or below average) and interval (0-100) indicators [22].

3.2.5 Novices, Professionals, and Expert System Use

Past research has examined the differentiated use of ESs by novice and experts in diverse settings and domains. Traditionally, expertise was established by employment in a domain or particular employment-related responsibilities (e.g., [1]). While professional responsibilities may be evidence of expertise, such responsibilities do not ensure expertise. Therefore, the term *professional* will be used as it is more accurate in describing experimental subjects and does not carry the same connotation that the term *expert* carries.

Professionals have been shown to use ESs differently than novices. They request more feedback and procedural knowledge explanations than novices and when they request explanations, professionals are more likely to adhere to the ES recommendation [1]. Professionals tend to verify their own models of decision making in light of the information that the ES is providing while the main goal of novices is to understand the problem they are facing [78]. Professionals also tend to use the ES explanations more to resolve anomalies that they perceive in the ES [51]. Novices are more likely to request strategic and definitional explanations [1]. Also, as users transition from novice to professional an increase is observed in the amount of attention given to DQI during the course of decision making [40].

3.3 Computer-aided Deception Detection

With reviews of computer-aided decision making and deception detection complete, pertinent concepts from both domains join to lay the foundation for computer-aided deception detection. The most important concept in computer-aided deception detection comes from maximizing the capabilities of both a human user and computer-aide. When improved human efforts join with technology, the potential exists for more diagnostic judgments of truth and lie [17]. More behavioral channels can be monitored and more cues that may indicate deception or truth can be captured. This concept is illustrated in Figure 3-7.

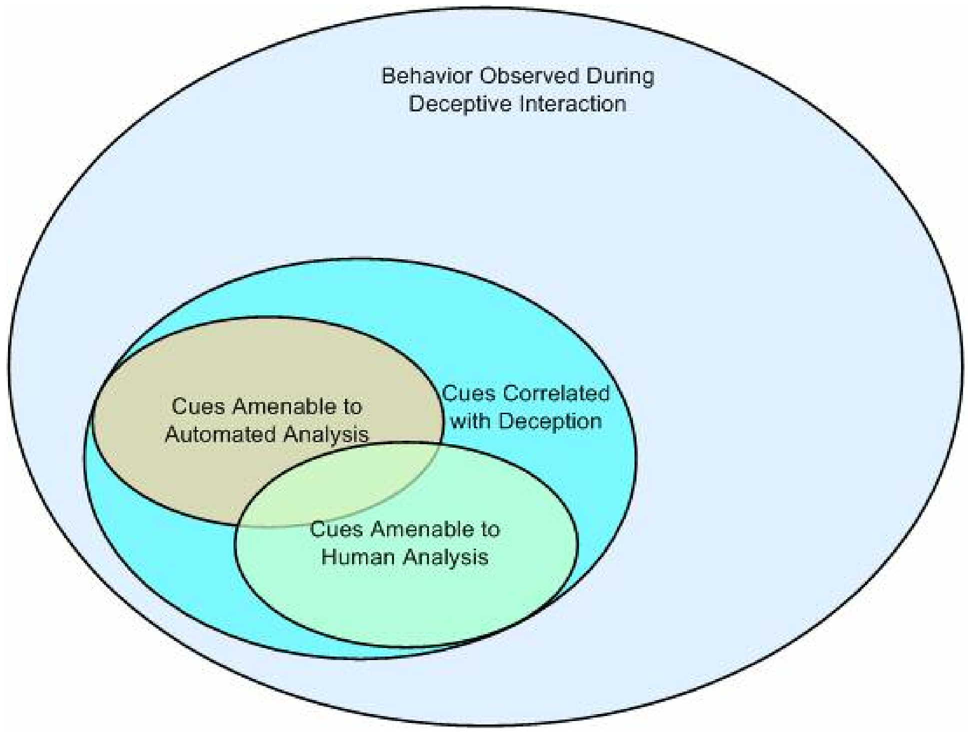


Figure 3-7 Cues to deception that are amenable to human and automated monitoring

Division of responsibilities between the human user and system is a nontrivial task that must be carefully examined to judge the effectiveness of joining improved human-only efforts with the capabilities of an ES. To guide the examination, a theoretical model of computer-aided deception detection will be presented that builds on existing work in DSS and ES. Capabilities of computer-based kinesic and linguistic analysis will be reviewed. Novice and professional deception detection abilities will be discussed and the topic of calibration will be addressed.

3.3.1 Theoretical Model for Computer-aided Deception Detection

A theoretical model suited to address computer-aided deception detection was formed by combining relevant factors from the Todd and Benbasat model [129] and

Gregor and Benbasat model [51]. This model appears in Figure 3-8. As the theoretical model deals only with deception detection, the task component of the model has been omitted; however, this factor could easily be included in a more generalized theoretical model for use outside of deception detection.

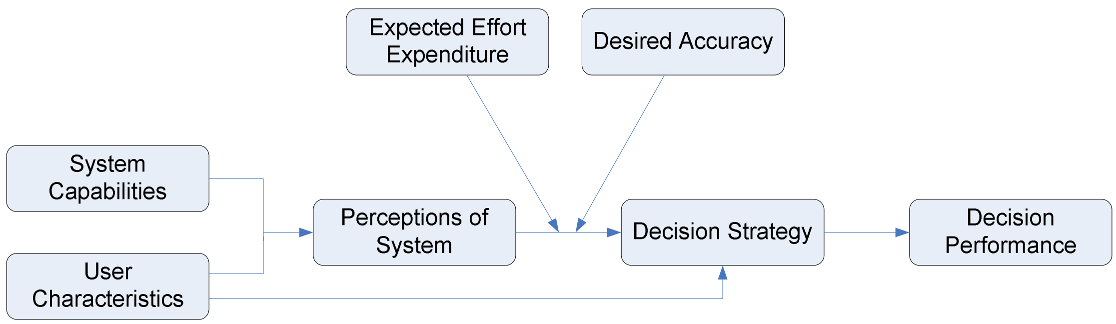


Figure 3-8 Theoretical model of computer-aided deception detection

Key independent factors in the model include system capabilities and characteristics of the decision maker. System capabilities may include qualities such as the provision of explanations, type of explanations, types of analyses the system performs (e.g., kinesic analysis, linguistic analysis, etc.), and accuracy and reliability of recommendations. User characteristics may include level of experience with deception detection, past experience with ESs or DSSs, and the user's general tendency to trust system recommendations.

The system capabilities and user characteristics directly affect the user's perceptions of the system. Such perceptions may include opinions concerning the accuracy of the data the system is analyzing, accuracy of the models or approach the system is using to analyze the data, and the user's willingness to use the system as a decision tool or delegated agent [68].

Next, perceptions of the system are thought to directly affect the strategies that users utilize in coming to a final decision. Within the context of deception detection, decision strategies can be extremely complex and difficult to capture objectively. Most deception detection decision strategies seem to be rooted in loosely defined heuristics. In fact, many professional lie-catchers are not fully aware of how they make judgments [138]. In contrast, the decision strategies reviewed previously (e.g., weighted additive rule, elimination by aspect, etc.) imply a structured approach to decision making. Therefore, previously reviewed strategies will not be utilized in this study (although further research could benefit from such investigation). Rather, decision strategy signifies how the user incorporates the knowledge and recommendation into the final decision or judgment. Decision strategies may include human anchoring and system anchoring. User characteristics may affect how the user incorporates system knowledge and recommendations.

As with the Todd and Benbasat model, expected effort and desired accuracy moderate the link with decision strategy. Effort and accuracy are interpreted similarly to the definitions outlined in [129] as described above.

Finally, decision strategy feeds into decision performance. Of primary importance in decision performance is accuracy (forming a correct judgment). Past work has forwarded the notion that a normative judgment should be the standard by which all decisions are measured. This notion ignores decision correctness, which is a critical component of deception detection. In addition to accuracy, alignment between decision confidence and accuracy is also a key component of decision performance as it directly affects the decision maker's conviction to the course of action.

3.3.2 Linguistic Analysis

Key capabilities provided by an ES in deception detection involve analysis of behavior that is difficult for humans to do but relatively easy for a computer to do. Among the deception detection methods that are useful in a natural environment, linguistic analysis can be automated to capture cues that are correlated with deception which are difficult for humans to monitor.

This discussion of linguistic analysis will only addresses message feature mining as this is the technique that was implemented and tested in the prototype (discussed in Chapter 4). Speech act profiling may at some future point be included in the prototype. However, given the current state of the prototype, speech act profiling is outside of the scope of this dissertation.

Empirical evidence suggests that deceivers may use language differently than truth-tellers [15, 16, 92]. Deceivers may have shorter talk time, share fewer details, and display elevated uncertainty [26]. Deceivers may also demonstrate less language diversity, less complexity, and use more negative words [92]. Numerous linguistic features may be used to gauge the credibility of messages. For example, both word counts and verb counts may be used to judge message lengths and average sentence lengths, and average word lengths may be used to judge language complexity. Zhou and colleagues [148, 149] present a listing of linguistic features that may be useful in deception detection. This list of features is summarized in Table 3-8.

Table 3-8 Zhou et al.'s linguistic features useful in deception detection

Category	Features	Description
Quantity	Words, verbs, noun phrases, sentences	Raw counts of words, verbs, noun phrases, and sentences
Complexity	Clauses, sentence length	Clauses / sentences; Words / sentences
	Word length	Characters / words
	Noun phrase length	Words in noun phrases / noun phrases
	Pausality	Punctuation marks / sentences
Uncertainty	Modifiers, modal verbs, other references	Raw counts of modifiers, modal verbs, and third person pronouns
	Uncertainty	Count of words denoting uncertainty
Non-immediacy	Passive voice	Passive verbs / total verbs
	You, self, and other reference	Raw counts of second-person pronouns, first-person singular, and first-person plural pronouns
Expressivity	Emotiveness	(Adjectives + adverbs) / (nouns + verbs)
Diversity	Lexical diversity	Unique words / total words
	Content word diversity	Unique content words / total content words
	Redundancy	Function words / sentences
Specificity	Temporal-immediate, temporal-nonimmediate	Temporal-immediate words / total words; Temporal-nonimmediate words / total words
	Spatial-close, spatial-far	Close words / total words; Far words / total words
	Positive imagery, negative imagery	Mean imagery score; Positive sensorial words / total words; Negative sensorial words / total words
Affect	Positive pleasantness, negative pleasantness	Mean pleasantness score; Pleasant words / total words; Unpleasant words / total words
	Positive activation, negative activation	Mean activation score; Positive activation words / total words; Negative activation words / total words

Although the features developed by Zhou et al., were originally intended to be used with synchronous and asynchronous persistent conversations (eg., chat, email), they are founded upon research that has taken place in face-to-face interactions [148].

The accuracy rate of linguistic analysis varies across contexts and environments; however, the method seems to provide consistent discrimination between liars and truth-tellers at a rate higher than typical humans (54%). Using numerous classification methods (including logistic regression, discriminant analysis, decision trees, and neural networks), the features described by Zhou et al. achieve accuracy rates between 60-90% in both experimental and field settings [20, 133, 150].

3.3.3 Kinesic Analysis

In addition to linguistic analysis, another unobtrusive deception detection method is kinesic analysis. Empirical evidence suggests that deceivers move their hands and heads differently than truth-tellers. Two meta-analyses conclude that there is a significant decrease in the amount of illustrating deceivers do in comparison to truth-tellers [26, 139]. Illustrating gestures are those gestures which normally accompany speech. They can include iconics, metaphorics, beats, and cohesives in the McNeill [81] classification. Illustrating gestures can represent semantic content in speech, can emphasize certain points, or can designate a relationship between ideas in speech. For example, an illustrator may be seen when one individual is giving directions to another and demonstrates a left turn with her hand. This gesture would represent the semantic content of the phrase “go left.”

DePaulo et al. [26] also point out that deceivers display significantly more chin raises than truth-tellers. They also observed less but non-significant undifferentiated head movement in deceivers; however, Buller et al. [15] found that deceivers show significantly less total head movement than truth-tellers. Meservy and colleagues [82] have developed a set of features that can be extracted from 2D image (frame) sequences. These features have been used productively in past research for identifying body movements that are correlated with deception. These features appear in Table 3-9.

Table 3-9 Meservy et al.'s kinesic features useful in deception detection

Body Part Measured	Features	Description
Head, RH, LH	x, y position	Position of the center point
Head	Angle	Angle of declination
Head	Angle difference	Difference in angles between previous frame and current frame
Head, RH, LH	Difference	Distance between x, y position between previous and current frame
Head, RH, LH	Triangle area	Area of triangle formed by connecting right, left hands and head
Head, RH, LH	Distance	Distance between head, and one hand; Euclidean distance between hands
Head, RH, LH	Tri-center distance	Distance between the center of the triangle connecting the head and hands and each of the hands and head
RH, LH	Top Quadrant	Dichotomous flag showing the presence of the hand above the chin
RH, LH	Right Quadrant	Dichotomous flag showing the presence of the hand to the right of the right shoulder
RH, LH	Center Quadrant	Dichotomous flag showing the presence of the hand between the left and right shoulders
RH, LH	Left Quadrant	Dichotomous flag showing the presence of the hand to the left of the left shoulder

As with linguistic analysis, the accuracy rates of kinesic analysis also vary across contexts and environments. However, the accuracy rate seems to be consistently higher than the rate exhibited by most people. Using numerous classification methods (including logistic regression, discriminant analysis, decision trees, and neural networks), the features described by Meservy et al. achieve accuracy rates between 60-88% in both experimental and field studies [17, 20, 59, 61, 82, 84].

3.3.4 Novice and Professional Deception Detection

There is significant evidence that professional lie-catchers perform only slightly better than novices (see [139] for a complete review). The lone exception to this finding may be United States Secret Service officers who, as a group, produced an accuracy rate significantly better than novice lie-catchers [35]. Other professional lie-catchers (including officers of state police departments, Federal Bureau of Investigation, Central Intelligence Agency, Drug Enforcement Agency, National Security Agency, state judges, and psychiatrists) performed no better than regular college students [35]. One explanation for this poor performance may be the relatively low stakes of the interviews the professionals viewed. However, Vrij and Mann [141] found that when viewing murder suspects' interviews, police offers were only 64% accurate. Vrij and Mann further state that professionals who held simple decision heuristics for deception detection performed the poorest in the detection task.

3.3.5 Calibration

As the accuracy of linguistic and kinesic analyses is less than 100% but greater than the performance of a typical human, users of decision aids must exert considerable effort to remain properly calibrated. On one hand, the operators may observe the less-than-perfect performance of the automated aid and summarily dismiss any conclusions formed by the computer aid. On the other hand, the human tendencies to satisfice, create decision-making heuristics, and ignore conflicting information provide fertile ground for over-reliance on the decision aid [87]. In the extreme, over-reliance on the decision aid may be classified as an *automation bias* [97, 118]. This bias has been seen repeatedly in contexts outside deception detection. The bias is divided into errors of omission, where the decision aid neglects pertinent information, and errors of commission, where the decision aid directs an incorrect course of action [87, 118]. Both errors of omission and commission result in hampered judgments.

The difficulty in calibration has been categorized into three major areas: disuse, misuse, and abuse [97]. Disuse encompasses problems associated with operators avoiding or ignoring the results of the automated aid. Misuse may include undue reliance on automation, increased complacency, and reduced situation awareness. Abuse concerns misapplication of automation (e.g., automation that increases workload, automation does more harm than good, etc.). While automation abuse may involve management and designers, automation disuse and misuse deal fundamentally with the user of the automated tool. Disuse and misuse can be effectively modeled by examining them along continua of actual reliability of the decision aid and perceived decision aid reliability, as shown in Figure 3-9. Disuse may occur when perceived reliability is lower than actual reliability and misuse may occur when perceived reliability is higher than actual reliability. One of the primary goals of the user should be proper calibration between actual reliability and perceived reliability. When expectations of the decision aid are too high or too low, the decision making ability of the human-computer system may suffer.

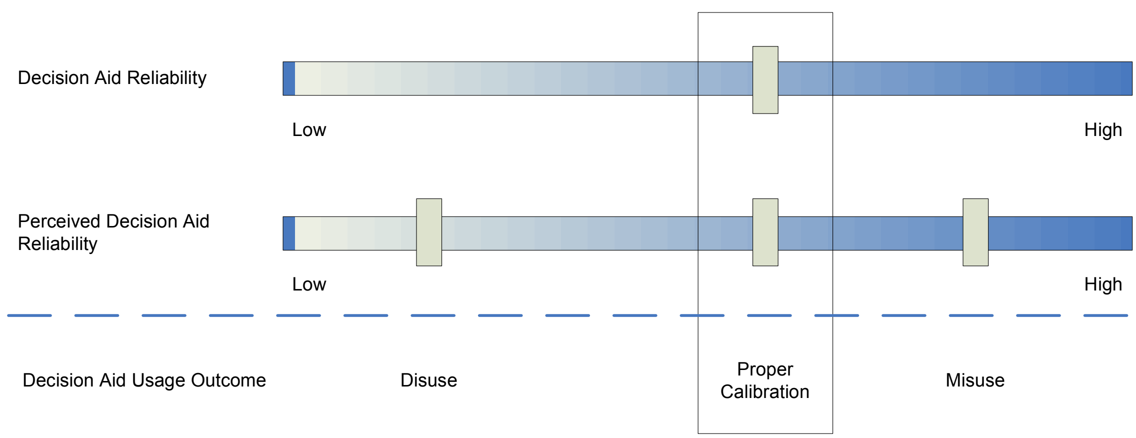


Figure 3-9 Calibration based on real and perceived reliability

Numerous methods have been proposed to counter the difficulty of calibrating computer-aided decision making. Parasuraman and Riley [97] suggest a guide to countering drawbacks of human-computer decision making. They address problems associated with automation use, misuse, disuse, and abuse. The suggestions are summarized in Figure 3-10.

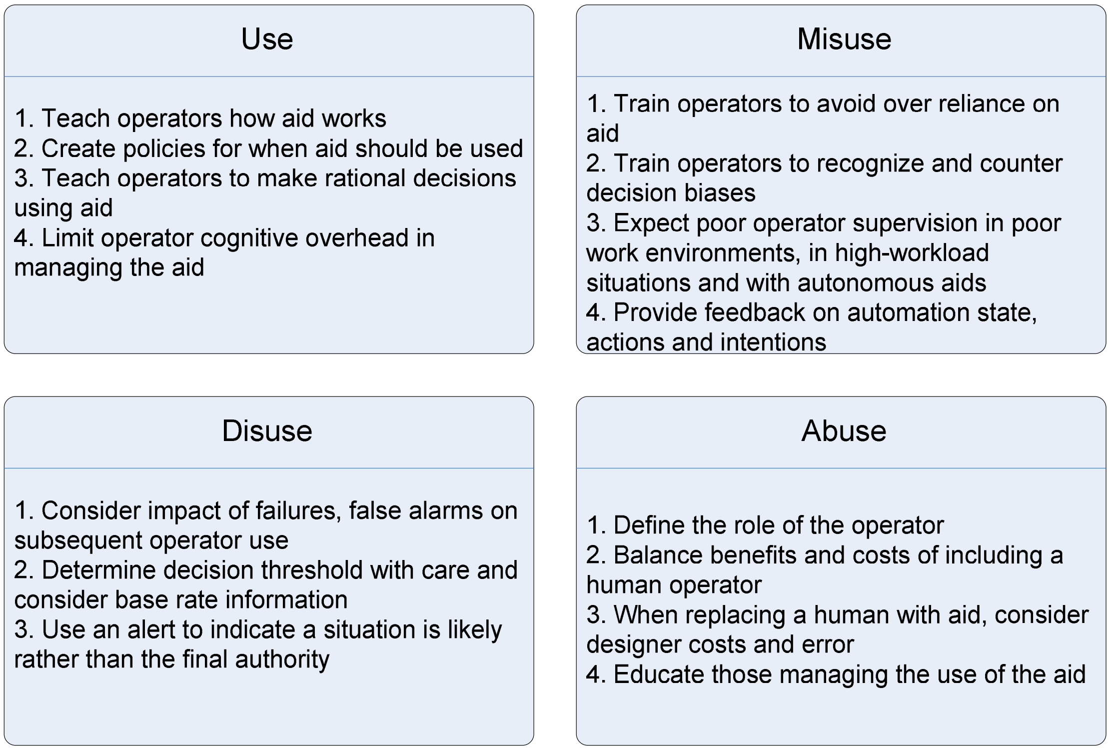


Figure 3-10 Suggestions for misuse, disuse, and abuse of decision aids

3.4 Research Questions and Hypotheses

To explore the key research question concerning the effects of a deception detection tool on the human decision maker, decision strategies, and outcomes, a series of hypotheses were formed that can be tested experimentally. The hypotheses were developed based on the existing literature and are divided into two groups. The first set of hypotheses deals solely with novice lie-catchers; the second set compares novices and professional lie-catchers.

3.4.1 Hypotheses Concerning Novices

Although validation studies are currently on-going, the kinesic and linguistic analyses have demonstrated accuracy rates from 60%-90%. This range is higher than the accuracy typically observed in unaided humans (54%) [9]. The kinesic and linguistic cues that these analyses examine may be difficult for humans to monitor consistently and vigilantly [17]. Therefore, use of a system that implements kinesic and linguistic analyses should extend and improve the abilities of the user.

In acknowledgment of the variation in accuracy rate observed in kinesic and linguistic analysis, two different accuracy rates are examined[4]. It is anticipated that accuracy of the system will also influence the accuracy of the user's judgment [129].

H1a: Novice users of the lower-accuracy system will demonstrate higher judgment accuracy than unaided novices.

H1b: Novice users of the higher-accuracy system will demonstrate higher judgment accuracy than novice users of the lower-accuracy system.

These two hypotheses taken together test a main effect for system use on accuracy.

A computer-based tool is not subject to many of the biases that hamper the ability of humans in detecting deception (e.g., reliance on a single or small set of deceptive cues [139]). The tool presents empirically-based deception judgments and the rationale behind those judgments. The user is able to review the reasons behind the assessments and then reconcile the reasoning with his or her observations. The reconciliation should induce a

[4] Discussion of the higher and lower accuracy levels of the system will be deferred until section 5.4.2.

critical consideration of the basis for each judgment [1]. This process of reconciliation is predicted to align the level of confidence with the accuracy level.

H2a: Novice users of the lower-accuracy system will demonstrate more alignment between judgment accuracy and confidence in judgment than unaided novices.

H2b: Novice users of the higher-accuracy system will demonstrate more alignment between judgment accuracy and confidence in judgment than novice users of the lower-accuracy system.

These two hypotheses taken together test a main effect for system use on confidence alignment.

To directly measure the influence of the system on the users' judgments, users are requested to provide judgments before observing the recommendations of the system (initial judgments). These initial judgments can then be examined and productively compared with the final judgments after the users have had the opportunity to review the system suggestions.

In past research, bogus manipulations significantly improved experimental participants' deception detection ability [73]. One explanation for this finding was that the participants subjected to bogus manipulations were conditioned to process messages more critically [73]. To examine this possibility, initial judgments from novice system users are compared with the judgments of unaided novices.

RQ1: What is the difference between the accuracy and alignment of the unaided novice users' judgments and the accuracy and alignment of the aided novice users' initial judgments?

Once the novice users gain access to and use the system, their judgment performance improves. As hypothesized above, the improvement is manifested in both judgment accuracy and alignment between judgment confidence and judgment accuracy. Further, it is anticipated that performance improvement will be greater for the users of the higher-accuracy system.

H3: Novice users will see improvement in judgment accuracy and alignment between initial and final judgments.

H4: Novice users of the higher-accuracy system will see greater improvement in judgment accuracy and alignment between initial and final judgments than novice users of the lower-accuracy system.

To further improve accuracy in deception detection, training of the human user will be introduced. As stated previously, training in deception detection has produced mixed results; however, more recent findings indicate that training improves detection ability by a small, but statistically significant amount [43].

H5: Trained novices will demonstrate higher judgment accuracy than untrained novices.

H6: Trained novices will demonstrate more alignment between judgment accuracy and confidence in judgment than untrained novices.

Researchers propose that users may treat the decision aid as a source of authority which is less likely to be questioned [97]. Most users are aware of the huge processing capability and memory of a computer and may believe that the computer's ability exceeds their own [87]. This may result in novices misusing (over-relying on) the system and system anchoring.

H7: When novices' initial judgments conflict with the system recommendations, the novices will anchor on the system recommendations.

One of the key factors in understanding ES use is perceived anomalies [51, 147]. Perceived anomalies serve as the trigger for explanation use, which in turn affects perceptions of the ES and ES use [51]. Novice users may not have the experience required to notice anomalies in the system operation and may not request clarification by way of system explanation use [147]. This lack of experience may also manifest itself by the lack of changes in user perceptions across higher-accuracy and lower-accuracy conditions.

RQ2: What is the difference in the number of explanations viewed by the novice users of the higher-accuracy system and the novice users of the lower-accuracy system?

RQ3: What is the difference in the emotional trust in the system, willingness to use the system as a decision aid, and willingness to use the system as an autonomous agent among novice users of the higher-accuracy system and the novice users of the lower-accuracy system?

These hypotheses are summarized along with directionality in Figure 3-11. The independent measures are on the left of the figure, while the dependent measures are on the right of the figure.

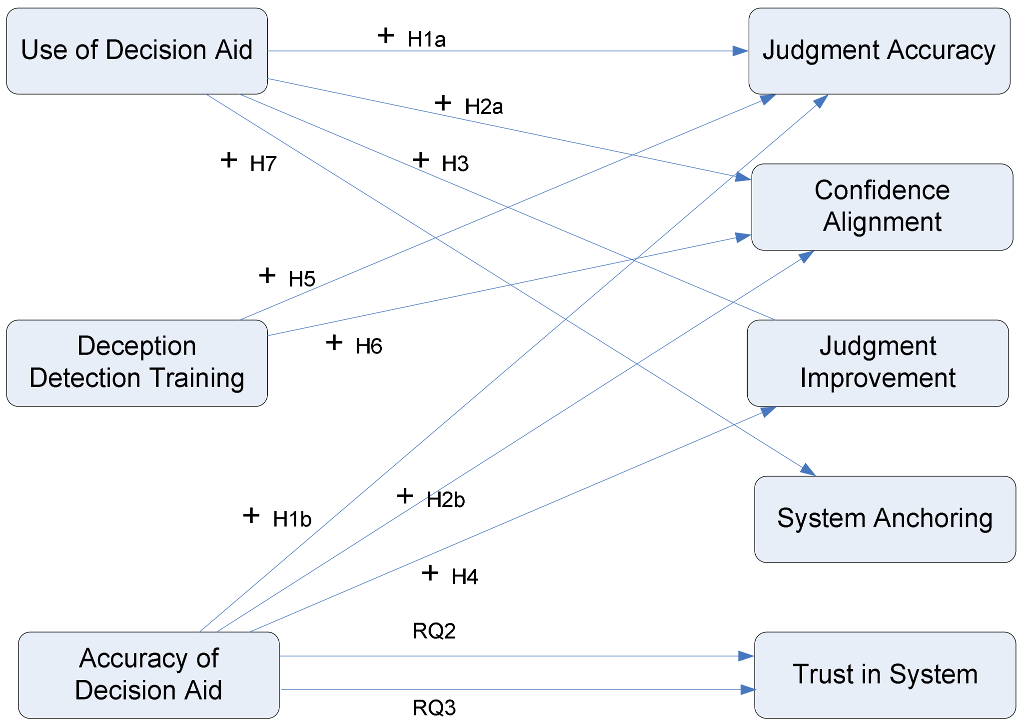


Figure 3-11 Hypotheses for novice computer-aided deception detection

3.4.2 Hypotheses Concerning Novices and Professionals

As with the hypotheses involving only novices, initial (unaided) and final (computer-aided) judgments can be productively compared to examine the effect of the computer aid. Consistent with prior findings (e.g., [35, 141]), it is proposed that professional lie-catchers may not be able to exceed the ability of novices in their initial judgments.

RQ4: What is the difference in the accuracy and alignment of the initial judgments of novices and professionals?

As soon as the novices and professionals are exposed to the functionality of the computer-aid, the professionals will more fully engage the system by utilizing its capabilities [1]. Such engaging will lead to explanation use and careful examination of the recommendations [51]. This, in turn, should lead to improved judgment accuracy and alignment between confidence and accuracy by the professionals [1, 51, 147].

H8: Aided professionals will see greater improvement in judgment accuracy and alignment between initial and final judgments than aided novices.

H9: Aided professionals will demonstrate higher accuracy in final judgments than aided novices.

H10: Aided professionals will demonstrate more alignment between judgment accuracy and confidence in final judgments than aided novices.

H11: Aided professionals will view more system explanations than aided novices.

As hypothesized above, professionals will be more likely to engage an ES and explore the explanations behind its recommendation. The utilization of system explanations has lead to greater acceptance of system recommendation in past research and favorable perceptions of the system [1, 51]. However, professional lie-catchers' abilities in deception detection do not approximate the level of expertise demonstrated in other contexts when ESs have been used. The accuracy of professional lie-catchers and laypersons in detecting deception is roughly equivalent [139]. Despite this equivalence, professionals manifest higher levels of confidence in their own judgment [26]. Higher confidence could encourage a reluctance to change judgment when countered by the system and could feed unfavorable impressions of the system.

Of these opposing forces, reliance on experience and confidence in initial judgments will be dominant. Thus, professionals will be less likely to alter their initial judgments and will have unfavorable perceptions of the system.

H12: When professionals' initial judgments conflict with the system recommendations, the professionals will anchor on their initial judgments.

H13: Aided professionals will demonstrate less emotional trust in the system, be less willing to use the system as a decision aid, and be less likely to use the system as an autonomous agent than aided novices.

These hypotheses are summarized along with directionality in Figure 3-12. The independent measures are on the left of the figure, while the dependent measures are on the right of the figure.

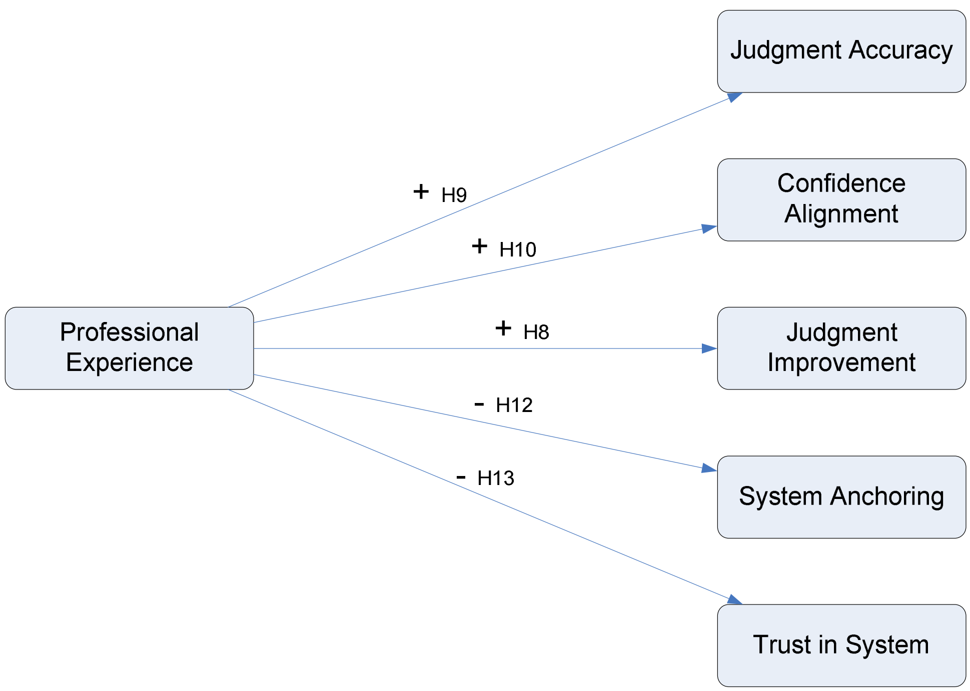


Figure 3-12 Hypotheses for expert and novice computer-aided deception detection

CHAPTER 4

EXPERIMENTAL PROTOTYPE FOR UNOBTRUSIVE, MULTI-MODAL DECEPTION DETECTION

The behavioral analysis prototype (BAP) is an interactive expert system that implements techniques from linguistic and kinesic analysis methods of deception detection. The methods used in the BAP build on past work in deception detection that has highlighted differences in the behavior of truth tellers and deceivers. Although the methods have been productively used in various communication modalities (e.g., email [150], synchronous chat [135], and voice-only conversations [60]), the methods in this dissertation will be applied only to face-to-face interaction. The BAP is comprised of four main components that appear in Figure 4-1.

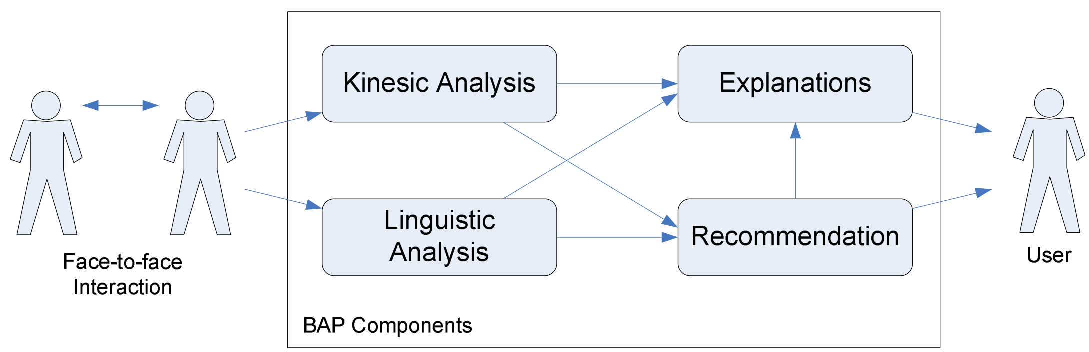


Figure 4-1 Components of the BAP

The BAP is not a fully functional, autonomous expert system; numerous steps in linguistic and kinesis analysis are currently manual. However, the BAP represents a reasonable proof-of-concept and is representative of the functionality that kinesic and linguistic analyses can provide. Further, the fully automated implementation of linguistic and kinesic analyses in a future, full-featured version is feasible.

Both implementations of the kinesic and linguistic analyses methods draw upon similar approaches. First, a recorded, face-to-face interaction is segmented into meaningful units. These units may cover a single question-answer exchange or may span an entire interaction. Next, low-level[5] features are created from image sequences and from transcripts of the interaction. Third, low-level features are used to calculate higher-level features which track particular behaviors that may indicate deception. The higher-level features were mentioned previously (see sections 3.3.2 and 3.3.3) and will again be revisited. The low- and higher-level features are then summarized across the segment and finally, the features are used to classify an interaction as truthful or deceptive. This process appears in Figure 4-2. Each step of kinesic and linguistic analyses will be addressed below.

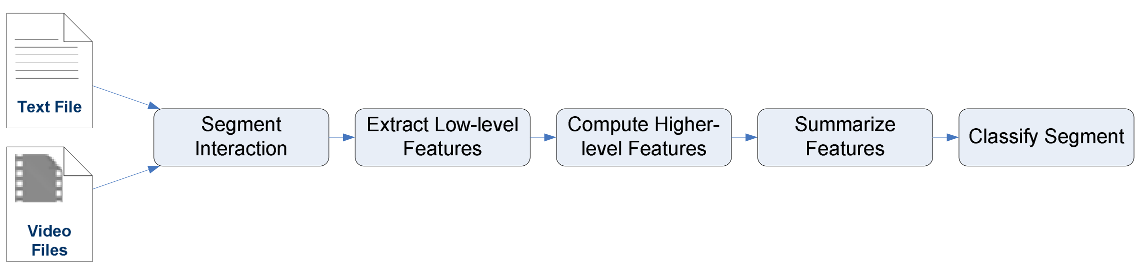


Figure 4-2 Steps of kinesic and linguistic analyses

[5] Low- and higher-level features can be ambiguous terms. In this context, features are higher-level if they are derived from other features (low-level features or other higher-level features). Such features may also be termed first- and second-order data.

4.1 Implementation of Linguistic Analysis[6]

The input for linguistic analysis is the text transcription of an interaction between two individuals. In this research, face-to-face interviews were manually transcribed, although automating the transcription process is technologically feasible. The first step in linguistic analysis is segmentation of an interaction. Segmentation allows linguistic analysis to focus on a particular section of an interview or other interaction that is of particular interest. For the purposes of this research, the segmentation was manually accomplished. However, methods do exist that can automate this step (e.g., [64]).

After the transcripts are segmented, a number of low-level cues are extracted from the segmented transcripts. These low-level features are generated by a part-of-speech parser and tagger and include counts and means of various linguistic constructs. The BAP utilizes the General Architecture for Text Extraction (GATE) [24] for text parsing. The results from GATE constitute the low-level linguistic features.

Additional features are calculated from the low-level features in an attempt to capture more abstract characteristics of speech. Examples of these higher-level features include level of emotiveness and passive voice ratio. Examples of low- and higher-level linguistic features are shown in Table 4-1.

---

[6] A summarized version of section 4.1 was submitted for publication in M. L. Jensen, T. O. Meservy, J. K. Burgoon, and J. F. Nunamaker, "Automatic, multimodal evaluation of human interaction." Tucson, AZ, 2007, pp. 1-19.

Table 4-1 Examples of low and higher-level linguistic features

Level	Feature Name	Calculation
Low	Verbs	Count of verbs
	Nouns	Count of nouns
	Adjectives	Count of adjectives
High	Passive verb ratio	Passive verbs / total verbs
	Lexical diversity	Unique words / total words
	Emotiveness	(Adjectives + adverbs) / (nouns + verbs)

Some higher-level features utilize the Whissell dictionary of over 7,000 words with scaled values for affect-related words and phrases [10, 142]. Features using the Whissell dictionary include levels pleasantness and activation. According to Zhou and colleagues [148, 149], low- and higher-level linguistic features can be divided into eight categories: quantity, uncertainty, non-immediacy, expressivity, diversity, specificity, and affect. Descriptions of the features appear in Table 3-8.

Following the extraction of low-level features and calculation of higher-level features, summarization of the features occurs. Summarization of the linguistic features is accomplished through counts, means and ratios. Using the summarized feature values, the interaction can then be classified as either deceptive or truthful. This can be accomplished through various classification techniques such as discriminant analysis, logistic regression, support vector machines, neural networks, and decision trees [150]. In the BAP, logistic regression was selected as the classification technique. Logistic regression is not limited by many of the assumptions that hamper discriminant analysis [105], and provides interpretable coefficients. Logistic regression provides a score (between 0 and 1) that indicates level of deception based on the linguistic features. Weka 3: Data Mining Software was used for logistic regression [144].

4.2 Implementation of Kinesic Analysis[7]

The preferred input for kinesic analysis is high quality digital video of a single subject who is in a sitting position away from any objects, such as a table, that might hide the hands or head. Higher quality video allows for more reliable position estimation of the head and hands. However, the method has also been used successfully with converted analog video (e.g., [59]).

As with linguistic analysis, the first step in kinesic analysis is to segment the video into meaningful units. Also in line with linguistic analysis, segmentation was done manually; however, methods do exist which can automate this step. Once segmentation has occurred, image sequences (frames) are extracted from the video. Any high-end, video editing software is capable of extracting image sequences and in this case, Avid Xpress Pro version 4.6 was used. A sample frame taken from an image sequence is shown in Figure 4-3.

[7] A summarized version of section 4.2 was published in T. O. Meservy, M. L. Jensen, J. Kruse, D. P. Twitchell, G. Tsechpenakis, J. K. Burgoon, D. N. Metaxas, and J. F. Nunamaker, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," IEEE Intelligent Systems, vol. 20, 2005.

Figure 4-3 Sample video frame

Next, low-level features are identified by estimating hand and face positions. Numerous techniques exist for automatic tracking of human head and hands. Notable among these techniques are Pfinder, developed at MIT [145], and Vector Coherence Mapping, developed at Wright State University [106]. The features used in kinesic analysis 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 appear in Figure 4-4.

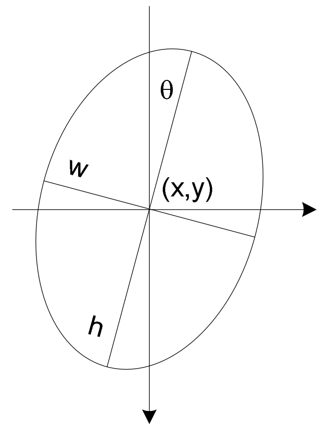


Figure 4-4 Necessary measurements for feature use [82]

The BAP utilizes a tracking algorithm developed by Computational Biomedicine Imaging and Modeling Center (CBIM) at Rutgers University [76]. 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. The 3-D LUT is based on the red, green, and blue components of individual pixels and is created in advance of any tracking using skin color samples.

After extracting the hand and face regions from an image, the tracking algorithm computes elliptical "blobs" that identify 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 tracking algorithm is beyond the scope of this dissertation; however the interested reader is directed to [76, 84]. A video frame that has been subjected to CBIM's tracking algorithm appears in Figure 4-5.

Figure 4-5 Single frame subjected to tracking algorithm [82]

From low-level features (shown in Figure 4-4), approximately 150 additional single-frame and multiple-frame higher-level features can be calculated automatically [82]. A limited subset of these features has been shown to differentiate between truthful and deceptive interactions and these features appear in Table 3-9. Implementations of sample features listed in Table 3-9 are illustrated in Figure 4-6 and Figure 4-7. Single-frame features include the distance between two blobs (see Figure 4-6) and the quadrant features (see Figure 4-7). Distance features are calculated by using the Euclidean distance formula on the center points of the blobs. The quadrant features are calculated using the measurements from the head blob. The base of the head blob delineates the boundary between quadrant 1 and the other quadrants. The other quadrants are specified using the width of the head blob. Two head widths comprise the width of quadrant 3 and quadrants 2 and 4 take up what remains [17].

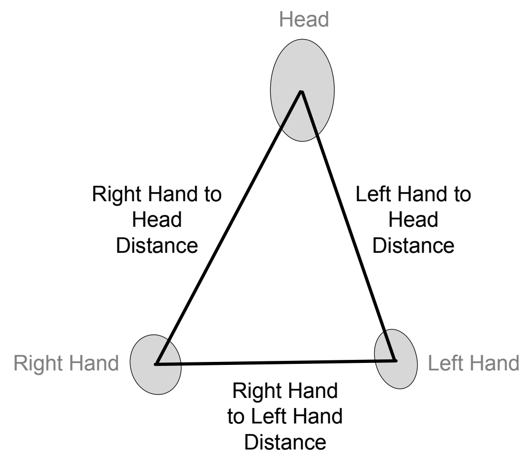


Figure 4-6 Distance features [82]

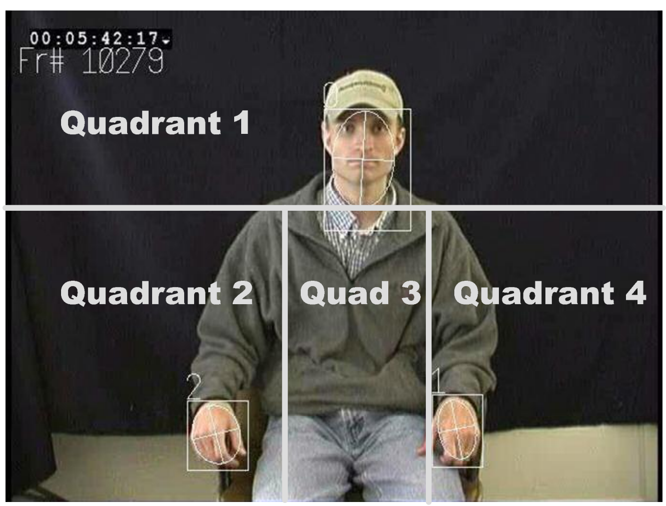


Figure 4-7 Quadrant features [82]

How far a blob has moved between frames is an example of a multiple frame feature. This feature is termed the difference feature in Table 3-9 and is calculated using the

Euclidean distance formula on the center-points of a single blob taken from two sequential frames. This feature is demonstrated in Figure 4-8.

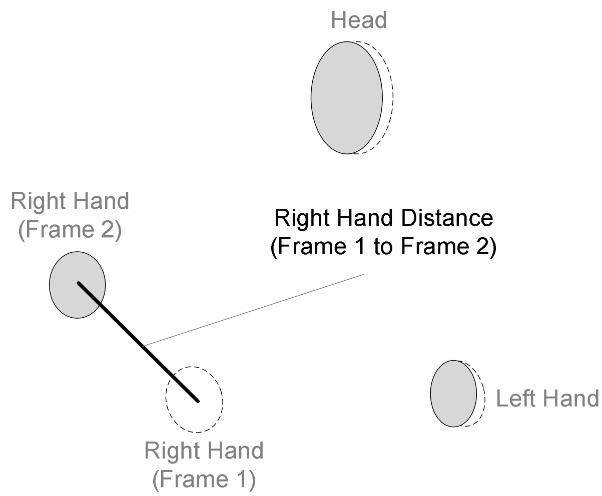


Figure 4-8 Difference feature [82]

After the extraction of lower-level features and the calculation of higher-level features, summarization of the features takes place. A value for each feature is recorded for every frame within an image sequence. These values are summarized via means and variances over the length of the segment to provide comparable measures.

Using the summarized feature values, the interaction can then be classified as either deceptive or truthful. As with linguistic analysis, this can be accomplished through various classification techniques such as discriminant analysis, logistic regression, support vector machines, neural networks, and decision trees [84]. With the BAP, logistic regression was selected as the classification technique for kinesic analysis. Logistic regression returns a score (between 0 and 1) that indicates a level of deception based on kinesic features. As in linguistic analysis, Weka 3: Data Mining Software was used for logistic regression [144].

4.3 Deception Levels, System Recommendation, and Confidence Level

The BAP reports a number of calculations that are based on the raw results of the logistic regressions from kinesic and linguistic analyses. Figure 4-9 shows an enlargement of sample BAP results. Additional samples of BAP results and recommendations are shown in Appendix A.

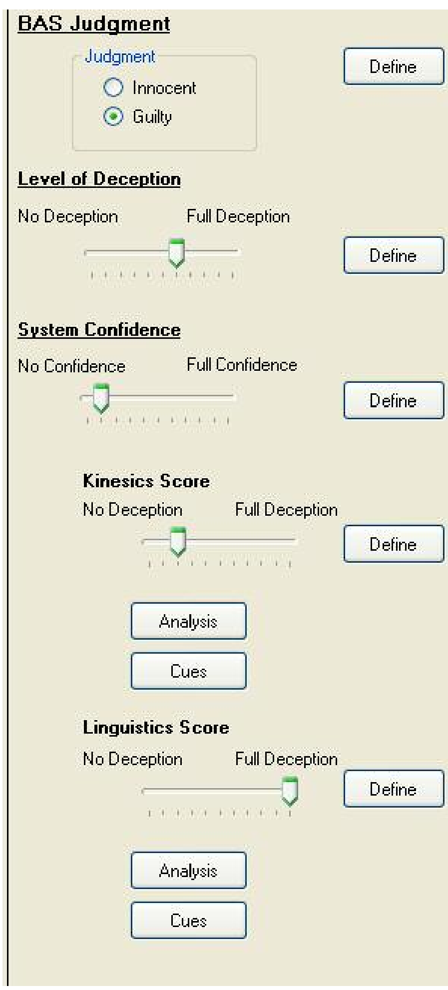


Figure 4-9 Sample BAP results

For both kinesic and linguistic analyses, the BAP offers a level of deception. These levels are termed linguistics and kinesics scores. As described previously, each score is taken from the raw output of logistic regression[8] and is transformed to a measure between 0 and 10.

The linguistic and kinesic scores are combined to produce a level of deception that is scaled from 0 to 10 (no deception to full deception). The combination of the scores takes place through a simple average[9]. Building on the level of deception, the judgment is a classification of guilt or innocence[10]. The BAP deems an individual guilty if the level of deception meets or exceeds a threshold of 5.

In addition to the judgment and level of deception, the BAP also provides a level of confidence in the judgment. The confidence level is intended to indicate strength of classification between guilty and innocent and is also scaled between 0 and 10. The confidence level is calculated by examining the absolute difference between the level of deception and the threshold of classification (deception level meets or exceeds 5). Thus, deception levels of 0 (indicating complete honesty) or 10 (indicating complete deception) will both have confidence levels of 10, indicating high strength of classification. The

[8] The raw output of logistic regression is a real number between 0 and 1 that can be interpreted as probability of class membership.
[9] The combination of linguistic and kinesic scores may take place via numerous methods (e.g., score weighting based on historical accuracy rates). Proper fusion of these scores was not addressed in the current dissertation. However this is a critical issue that must be explored for multi-modal deception detection to be possible.
[10] The BAP was used in a context where direct questioning about cheating occurred. Thus, a judgment concerning the guilt or innocence of an individual is appropriate. However, a judgment of guilt or innocence may not be appropriate in cases where direct questioning about wrongdoing does not occur. In such a case, the binary classification would indicate only that deception has occurred, not that the individual was guilty or innocent.

calculation used for confidence level is shown in Equation 4-1 where c is the level of confidence, d is the raw score of logistic regression (deception level), t is the threshold, and w is a term intended to transform the score to a value between 0 and 10.

Equation 4-1 $$c = | d - t | w$$

4.4 Explanations

A critical component of BAP is the explanatory capability that it provides. The explanations that the BAP provides inform the user of the reasoning behind the recommendations. As shown in Figure 4-9, BAP explanations come in three categories: definitions, cues, and analysis. These categories of explanations were designed to fill similar roles of explanations from other ESs, namely: definition, rule-trace, and justification explanations [1, 30]. In contrast to past ESs [1], the explanations provided by the BAP are provided solely in the feedback channel. There are no feedforward explanations that are available to the user before the system provides a recommendation.

In line with past research on ESs, the explanations were created with expert knowledge and reviewed by other experts. The explanations in the BAP were developed using knowledge and feature descriptions from peer-reviewed, published literature. In addition, an expert in deception detection reviewed the text of the explanations and provision of the explanations to ensure correctness and accessibility. Each type of explanation is discussed below and sample BAP explanations appear in Appendix A.

4.4.1 Definitions

The BAP provides definitions of five critical components: linguistic analysis, kinesic analysis, level of confidence, level of deception, and system judgment. The explanations share procedural and declarative knowledge about each of these components. The definitions are available whenever the BAP is accessible; however, the contents of these definitions do not change with the analysis of different individuals.

4.4.2 Cues

Explanations concerning the cues deal primarily with the results of blob analysis and message feature mining. Relative scores for individual features are presented in the cues explanations. For example, one portion of the cues explanation for linguistic analysis may read “This person demonstrates a HIGH amount of lexical diversity.”

Scores for individual features are reported as they relate to other scores. Raw feature values are categorized into levels of LOW, MODERATE, and HIGH based on standard deviations from the mean of the training set. This method of explanation provision effectively notes any deviation from observed norms and at the same time does not burden the user with needless numeric feature values. Feature values more than one standard deviation below the mean are termed LOW; values within one standard deviation (+/-) of the mean are classified as MODERATE; values greater that one standard deviation above the mean are classified as HIGH.

4.4.3 Analysis

The analysis explanations provide natural language reasoning behind a level of deception as provided by linguistic or kinesic analyses. The analysis explanations provide context for the features that are described in the cues and definition explanations and link the features to concepts that are familiar to the user such as arousal and tenseness. An example analysis explanation from the kinesic score is: “Analysis of this interview indicates that this person demonstrates moderate levels of body movement and activation of the head. This person does not demonstrate excessive over control or tenseness. Further, this person does not display agitation or nervousness.”

The analysis explanations also provide a probability of deception that mirrors what is shown to the user in the deception levels. For example, if the deception level for kinesic analysis shows 0-3, 4-6, or 7-10, then the probability of deception stated in the analysis explanation will be LOW, MODERATE, and HIGH, respectively.

4.5 Experimental Requirements and Modifications

Several additional components were added to the BAP for purposes of experimentation. First, a video viewer was added to the BAP to enable a user to view recorded interactions during the experiments. Further, a volume bar was also added to ensure that conversations observed during the experiments were audible. These additions appear in the complete BAP interface shown in Figure 4-10.

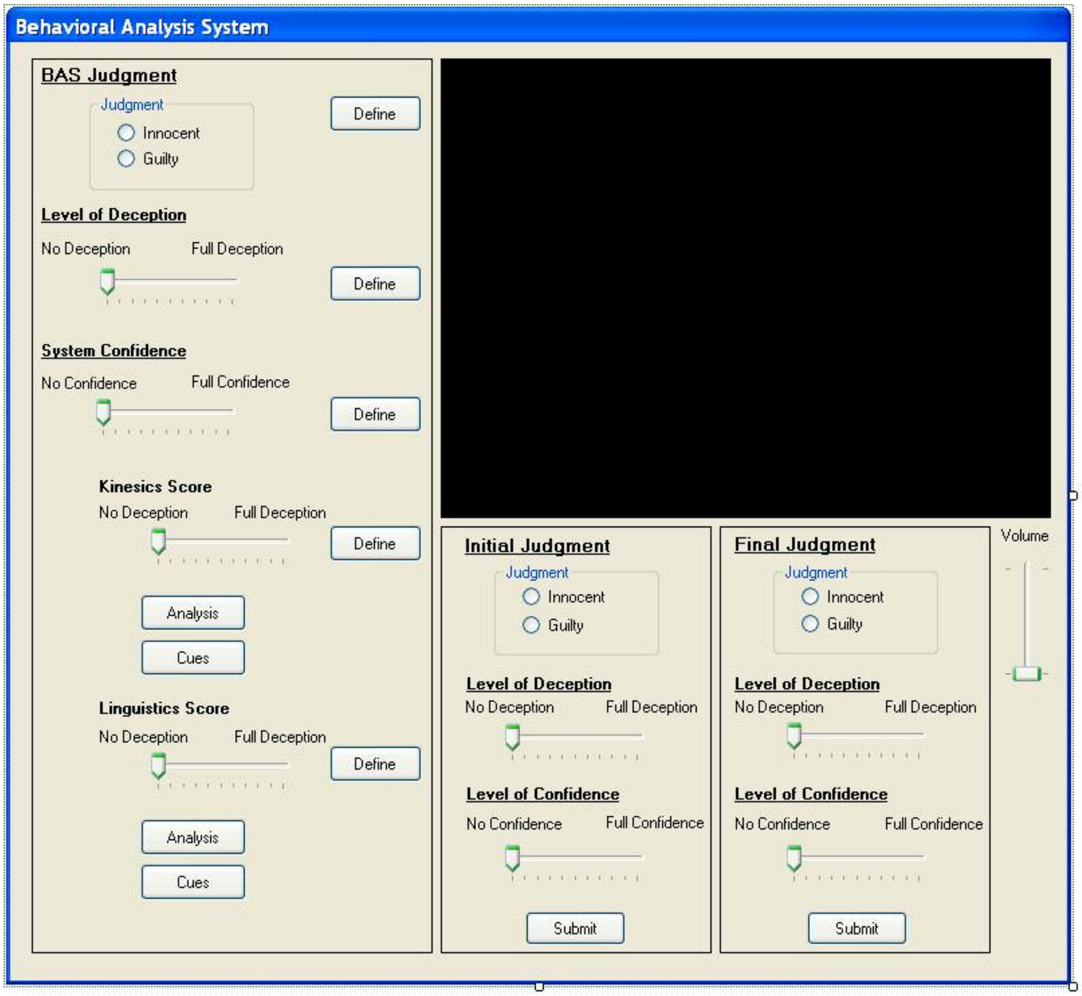


Figure 4-10 Complete BAP user interface

4.5.1 Interactions with the BAP Interface

As shown in Figure 4-10, the experimental BAP requires two judgments from the user. Directly after viewing a video, the user is required to enter an initial judgment. This judgment consists of a judgment (guilty or innocent), level of deception, and level of confidence. Following the completion of the initial judgment, the user is allowed access to the BAP's recommendations and explanations. After the user has examined the results from the BAP, the user records a final judgment that again includes a judgment, level of deception and level of confidence. The order of interaction with the BAP is demonstrated in Figure 4-11.

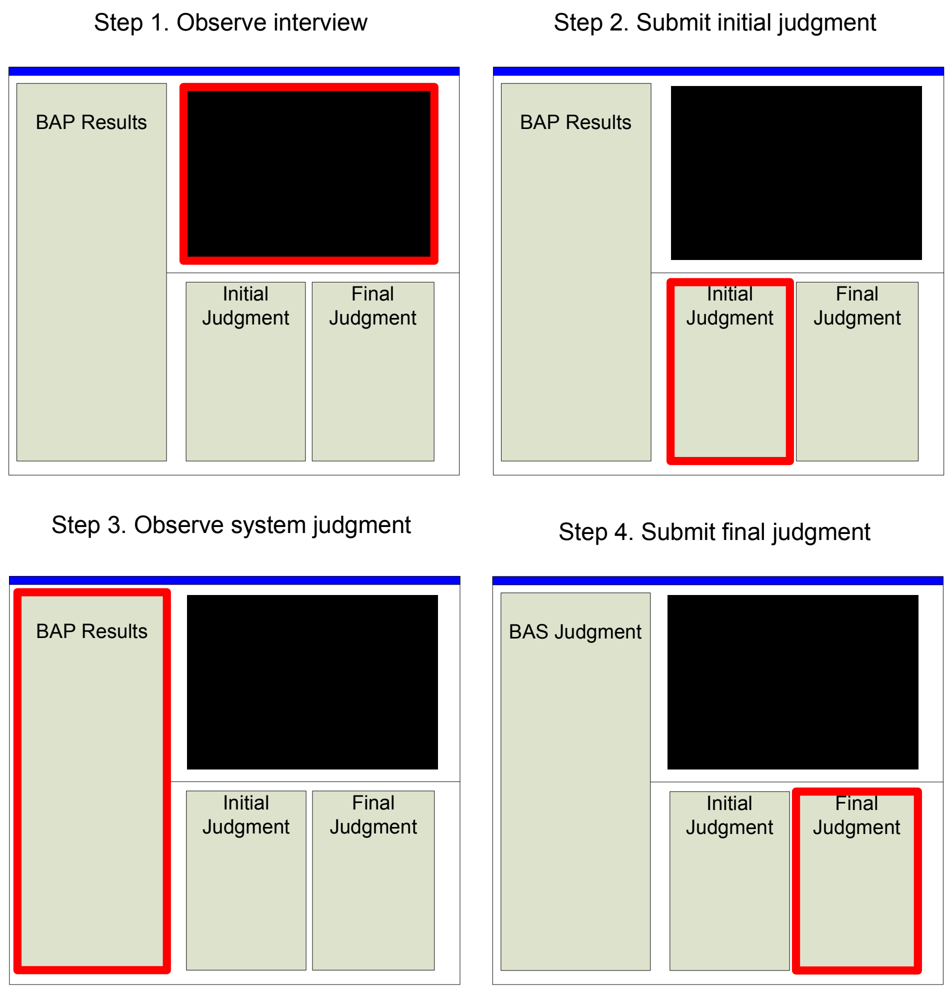


Figure 4-11 Interaction with the experimental BAP

For practical use, the initial and final judgments are unnecessary and can be easily excluded. However, in the experimental setting, the BAP interface provides an effective means for recording responses from users that can be examined later.

4.5.2 Training the BAP

Logistic regression is a classification method that requires training data to create a classification function. Training data are necessary for both linguistic and kinesic analyses. The experimental BAP was trained by following methods from signal detection theory (SDT) [123, 124], whereby diagnostic features were included in a classification function designed for the sole purpose of analyzing the interactions used in the experiments. The diagnostic features were selected from the kinesic and linguistic features outlined in sections 3.3.2 and 3.3.3. The experimental BAP calculated the weights using an iteratively-reweighted least squares algorithm contained in the Weka data mining toolbox [71, 144]. This algorithm uses ridge estimators to improve weight estimates and to diminish the error in predictions that are made using the weights [71].

4.5.3 Creation of Explanations

As mentioned in section 4.4 , the explanations were manually created based on expert knowledge. However, in the experimental BAP the explanations are tightly coupled to the recorded interaction the user observes. In other words, the explanations (in particular the analysis explanations) are specifically tied to the recorded interactions the users viewed during the experiments. This method of explanation provision is infeasible in a fully operational system. However, this method was adopted in the prototype system as it allowed rapid development and gave the users the impression that the system was fully functional.

Automatic provision of explanations is feasible in future systems. The definition explanations are static and can be used repeatedly in analyzing numerous interactions.

The cues explanations can be automatically extracted by examining the results of blob analysis and message feature mining and comparing the feature values to historical means. Finally, the analysis explanations can be developed which rely on the coefficients in the logistic regression function. Natural language explanations can be developed which link certain weighted combinations of cues to general interactional dimensions such as involvement, arousal, and tenseness (for example, see [60])

CHAPTER 5

NOVICE COMPUTER-AIDED DECEPTION DETECTION

An experiment was designed and carried out to address the hypotheses outlined in section 3.4.1, and to examine the effects of the BAP system on novice judgments. The description of the experiment and results are discussed below.

5.1 Participants

The experiment involved 185 participants recruited from two sections of an upper-division Management Information Systems (MIS) course at the University of Arizona. The MIS course is required for all business majors (accounting, business economics, business management, entrepreneurship, finance, management information systems, marketing, and operations management) and public administration majors (criminal justice, human and health services, public management and policy). The mean age of the participants was 21.9, mean years of secondary education was 3.8, and of all the participants, 50.4% were female.

5.1.1 Motivation of Participants

The participants were recruited to participate in the experiment by public announcements during class. The participants were awarded extra credit toward their final grades for participation (approximately 2% of their final grade). To further motivate the participants, those who scored in the top 10% in each experimental condition were offered a cash reward in the amount of $10.

5.1.2 Justification of Sample

The novice participants selected for this study represent untrained, non-professional lie-catchers that could benefit from a decision aid in deception detection. They have been exposed to some forms of deception but do not bear the professional responsibility of detecting deception on a regular basis. They have not had extensive training in deception detection so effects of training and BAP use can be captured and analyzed accurately.

5.2 Stimulus Materials

The stimulus materials for this study were provided by a previous experiment that was conducted at Michigan State University (MSU) in 2004 by Timothy Levine and colleagues [72]. In that study, high-stakes, unsanctioned deceptive and truthful interactions were digitally recorded. The original purpose of the study was to examine the behavior of deceivers and truth-tellers under high stakes and to understand observers' accuracy rates across varying ratios of truth-tellers and deceivers [74].

The interviews were approved by the Institutional Review Board at MSU for use in further deception detection studies (e.g., observer studies) at MSU and other institutions. The data set contains 22 interviews which were used in the experiment (7 deceptive, 15 truthful). All of the interviews were used for training the BAP and 10 interviews, 5 truthful and 5 deceptive, were selected for stimulus materials during the experiment.

5.2.1 Method of Collection of the Stimulus Materials[11]

In a variation of the Exline procedure [36], students from an introductory communication course at MSU were invited to participate in an experiment. The participants were informed that the study concerned effective teamwork, and deception was not mentioned in any experimental instructions. The participants were informed that they would be working in pairs to answer difficult trivia questions and they were promised a cash reward ($20 each) if they performed well on the trivia questions. Each participant was paired with a confederate, and an experimenter entered the room and asked a number of obscure trivia questions.

After a few questions, the experimenter was called out of the room and left the set of trivia questions and answers in the room with the participant and confederate. The confederate then encouraged the participant to cheat and look at the answers. The participant then either observed the answers or refused to observe the answers. The participants self-selected their treatment group.

After a few minutes, the experimenter returned and finished asking the trivia questions. After all the trivia questions were complete, the confederate and participant were separated and the participant was brought to an interview area where he or she was interviewed. The participants were previously told that the interview would concern the role of teamwork in responding to the trivia questions. Instead, the interviewer confronted

[11] A complete description of all the conditions in the MSU cheating study can be found in T. R. Levine, R. K. Kim, H. S. Park, and M. Hughes, "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, vol. 73, pp. 243 - 260, 2006.

the participants with a structured interview to find out if they had cheated. All participants were paired with the same confederate and all participants were interviewed by the same interviewer. The interviewer posed the same questions to all participants. A full script of the interview questions appear in Table 5-1.

Table 5-1 Questions in MSU cheating study interviews

Did you find the trivia questions difficult?
Was teamwork much of a factor? How so? Please explain.
In looking at your score, you did better than other groups. Does this surprise you?
How would you explain your success?
Did you cheat when the experimenter left the room?
Why should I believe you?

5.2.2 Advantages and Disadvantages of the Stimulus Materials

Using these interviews has many advantages and some drawbacks. First, the stakes in these interviews were higher than most experimentally generated interviews and the lies were unsanctioned. Typically, high stakes are assumed to be present if there are severe consequences for the deceiver if the deception is revealed. In this case, each deceptive interviewee was seeking to preserve his or her reputation and character through deception. Such self-preservation is thought to be a strong motivation and is commonly guarded through deception [28]. Additionally, follow-up conversations with the participants showed that they feared disciplinary action for cheating during the experiment.

In each interview, ground truth is certain. The confederate knew which participant cheated and which did not. Also, the questions were worded such that deceivers were forced into outright falsification and justification. There was little room in the interview for ambiguous forms of deception such as equivocation and exaggeration.

Additionally, the scenario is accessible to most observers and the interview is consistent across all cases. Cheating is familiar to many college students as it is discussed frequently in class and campus communities. Thus, student observers would avoid confounding effects from unfamiliarity with interview material. Further, there would be no effect resulting from differing questioning techniques across cases as the questions in the interviews are all consistent. Other deception studies have involved probing as part of the interview questioning. While lack of probing departs slightly from reality, it does provide for comparability across cases.

Finally, the interviews are relatively short with an approximate time of two and a half minutes for each interview. This benefit allows for many interviews to be observed by one person in a reasonable amount of time.

While the stakes in the interviews are increased, they still may not approach the high stakes present in a criminal investigation; interviews from an actual law enforcement scenario would be ideal. Such interviews are available but would be extremely difficult to use in a study where others are observing the interview because of privacy issues and human subject protections.

Also, the participants self-selected their own groups (deceptive, truthful). This caused unequal numbers of truthful and deceptive interviews. Although humans are faced with unequal amounts of truth and deception in normal interactions with others, unequal numbers of truths and deceptions may skew the training of the BAP.

5.3 Experiment Design Matrices

As described in section 3.4.1, there are three independent variables that were manipulated during this experiment: use of the BAP; accuracy of the BAP; and access to training. Use of the BAP and accuracy of the BAP are split among three treatments: No BAP access, access to higher-accuracy BAP, and access to lower-accuracy BAP. Thus, the experiment utilizes a mixed (between- and within-subjects) model with repeated judgments as the within-subjects factor. The experimental design matrix and sample size in each cell is illustrated in Table 5-2.

Table 5-2 Experimental design matrix for novices

	No Training	Training
No BAP Access	$N = 31$	$N = 31$
Access to Lower-accuracy BAP	$N = 31$	$N = 31$
Access to Higher-accuracy BAP	$N = 30$	$N = 31$

5.4 Independent Variables

5.4.1 Use of the BAP

Use was operationalized by availability and access to the BAP. The access group was free to use the BAP as much or as little as they wished in forming their judgments. The no-access group was not able to use the BAP. In the no-access group, the computer interface excluded all components of the BAP and only required a final judgment, as shown in Figure 5-1.

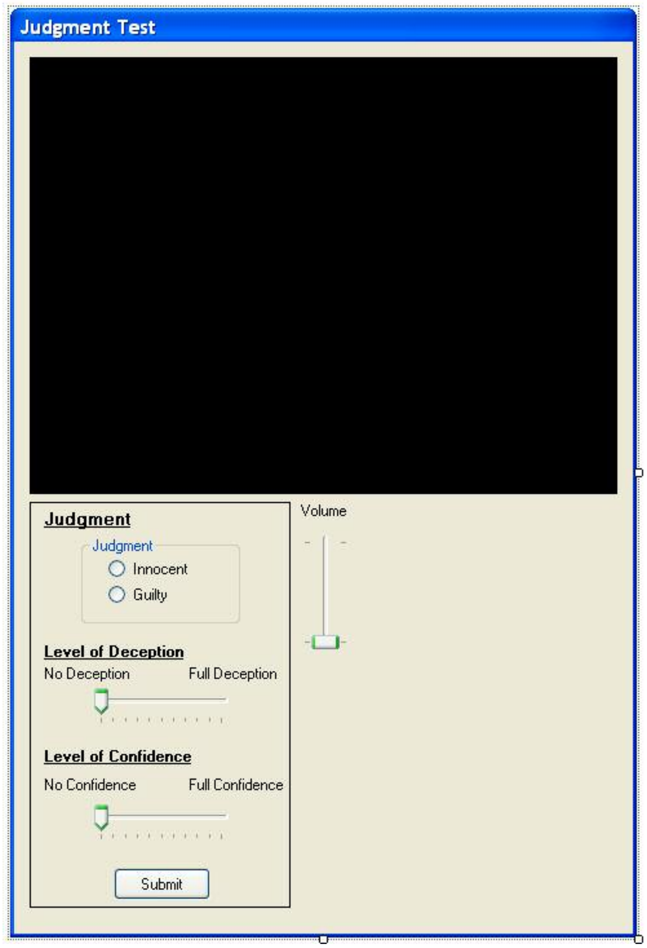


Figure 5-1 BAP interface for the no-access condition

The difference between availability of the decision aid and use of the decision aid is readily acknowledged. Usage may vary by degree through time and thus is not a suitable factor. On the other hand, availability is an observable treatment that may be administered. Actual BAP use is addressed in section 5.6.3 when the improvement between initial and final judgments of BAP users is discussed.

5.4.2 Accuracy of the BAP

Accuracy of the BAP was manipulated to form two conditions: higher accuracy (80% correct classification) and lower accuracy (60% correct classification). These accuracy rates are within the performance range of past implementations of kinesic and linguistic analysis (see sections 3.3.2 and 3.3.3), and therefore represent reasonable performance expectations for future systems similar to the BAP. Further, research investigating automated decision aids with varying reliability rates suggests that there seems to be a cutoff at 70% accuracy when the decision aid ceases to be useful in making decisions [143]. This experiment examined accuracy rates above and below this suggested threshold.

To attain an accuracy rate of 80% correct classification across truthful and deceptive interactions, linguistic and kinesic analyses were initialized to focus on a small, manually-selected subset of behavioral features. The features that were utilized appear in Table 5-3.

Table 5-3 Features used in higher-accuracy BAP

Linguistic Features		Kinesic Features	
Feature	Description	Features	Description
Lexical Diversity	Unique words / total words	Avg Head difference	Average of the difference in head x,y position between previous and current frame
Verb Quantity	Count of verbs	Avg RH Q3	Average number of frames the right hand is in the torso area
Affect Ratio	Number of affective words / total words	Avg LH Q3	Average number of frames the left hand is in the torso area
Group References	Count of first-person plural pronouns	Var Head difference	Variance in the difference in head x,y position between previous and current frame
Activation	Number of active words / total words	Var Distance between LH and RH	Variance in the distance between the left and right hands

By focusing on these features, the BAP in the 80% condition misclassified a single deceptive interview and a single truthful interaction. The BAP results and recommendations for all 10 interviews seen in the higher-accuracy (80%) condition appear in Table 5-4. Higher linguistics scores, kinesics scores, and deception levels correspond to higher levels of deception.

Table 5-4 Results of the BAP in the higher-accuracy condition

Interview No.	Linguistics Score	Kinesics Score	Judgment	Deception Level	Confidence Level	Actual
1	9	2	Guilty	6	1	Guilty
2	2	1	Innocent	1	8	Innocent
3	2	4	Innocent	3	5	Innocent
4*	3	2	Innocent	2	6	Guilty
5*	5	6	Guilty	6	1	Innocent
6	0	2	Innocent	1	8	Innocent
7	3	7	Guilty	5	0	Guilty
8	8	1	Innocent	5	1	Innocent
9	7	10	Guilty	9	7	Guilty
10	10	8	Guilty	9	7	Guilty

(*misclassified)

While the features were manually selected, the features that were used have all demonstrated utility in differentiating truthful from deceitful interactions. All of the included features were taken from the Zhou et al. [148] and Meservy et al. [82] feature sets that have been successfully used in past deception detection.

Once the features were established in the higher-accuracy condition, the features for the lower-accuracy condition were designated. This was accomplished following implications suggested by SDT. SDT recommends that higher accuracy can be attained by including more diagnostic features in the decision process [124]. Thus, by inverse, lower accuracy rates can be attained by excluding diagnostic features in the decision process. Therefore, the lower-accuracy condition was created by manually excluding features that were included in the higher-accuracy condition until 60% accuracy was achieved. The features utilized by the BAP in the 60% accuracy condition appear in Table 5-5.

Table 5-5 Features used in lower-accuracy BAP

Linguistic Features		Kinesic Features	
Feature	Description	Features	Description
Verb Quantity	Count of verbs	Avg Head difference	Average of the difference in head x,y position between previous and current frame
Affect Ratio	Number of affective words / total words	Var Head difference	Variance in the difference in head x,y position between previous and current frame
Group References	Count of first-person plural pronouns	Var Distance between LH and RH	Variance in the distance between the left and right hands
Activation	Number of active words / total words		

By focusing on these features, the BAP in the 60% condition misclassified two deceptive interviews and two truthful interactions. The BAP results and recommendations for all 10 interviews seen in the lower-accuracy (60%) condition appear in Table 5-6.

Table 5-6 Results of the BAP in the lower-accuracy condition

Interview No.	Linguistics Score	Kinesics Score	Judgment	Deception Level	Confidence Level	Actual
1	9	1	Guilty	5	0	Guilty
2	1	1	Innocent	1	7	Innocent
3	1	2	Innocent	1	7	Innocent
4*	4	2	Innocent	3	4	Guilty
5*	8	5	Guilty	6	3	Innocent
6	1	2	Innocent	2	7	Innocent
7	6	7	Guilty	7	3	Guilty
8*	7	4	Guilty	6	1	Innocent
9	3	10	Guilty	7	3	Guilty
10*	4	4	Innocent	4	2	Guilty

(*misclassified)

5.4.3 Training

The training manipulation occurred in two conditions: training and no-training. The training material used in this experiment was adapted from existing training materials produced by the Center for the Management of Information at the University of Arizona and was taken from a 14-lecture course on nonverbal communication [31]. The training consisted of a video-taped lecture with embedded examples, a knowledge quiz, and a judgment quiz that were all delivered via computer. The training focused on deceptive cues that humans are qualified to identify and addressed key topics shown in Table 5-7. A summary of the training and quiz materials appear in Appendix B.

Table 5-7 Key topics presented in deception detection training

Definition of deception
Forms deception can take
Theoretical connections between deception and observable behavior
Deception cues associated with arousal
Deception cues associated with emotion
Deception cues associated with cognitive effort
Deception cues associated with memory processes
Deception cues associated with communicator tactics

Those in the training condition were shown the training video prior to participating in the deception detection task. Those in the no-training condition were also shown the training video but after the conclusion of the deception detection task. The video was promoted as a benefit to the experimental participants, and showing the video to all participants balanced the time required for different treatment groups.

The video-taped lecture lasted approximately 23 minutes. Following the lecture, the participants were required to answer a knowledge quiz consisting of 10 questions that were designed to cement important information about deception detection. Following the completion of the quiz, the correct answers were revealed so the participants could check their responses. Following the knowledge quiz, the participants completed a judgment quiz where they viewed five video-taped interviews and determined if the interviewee they observed was being deceptive or truthful. Following each response, the correct answer was shown and explanations were provided that highlighted behavioral cues the interviewees displayed that could have tipped off an observer.

5.5 Experimental Methods

Upon arriving to the lab, the subjects were pseudo-randomly[12] assigned to conditions. The participants were seated at a computer and each viewed an orientation video. For those with access to the BAP, the orientation video provided a brief description of the BAP and also reported accuracy rates of past validation efforts (although some studies have suggested higher performance, 60%-80% accuracy was reported to the participants). The contents of the orientation can be found in Appendix C. The participants filled out an initial survey that questioned them about topics like their level of comfort with a computer, expectations concerning the deception detection task,

[12] All conditions were randomly assigned from the beginning of the experiment except the lower-accuracy, no-training condition. The number of potential participants was thought to be limited; therefore participants were assigned other conditions first. When the number of participants was deemed to be sufficient, the lower-accuracy, no-training condition received participants to raise its sample size to the level of the other conditions. Assignments were randomly made thereafter.

and confidence in their decision making ability. All survey instruments appear in Appendix D.

Following the survey, those in the training conditions viewed the training materials. Those without access to the training began the deception detection task. The order of the 10 interviews each participant saw was randomized. Following each of the 10 judgments, each participant was asked to provide their reasoning behind their judgment and those with access to the BAP were asked if the system was of benefit to them. Following the deception detection task, all participants were questioned in a final survey. Those in the no-training condition then commenced with the training. The entire experiment lasted approximately 1.5 hours.

5.6 Analysis

Analysis addressed five dependent measures (judgment accuracy, confidence alignment, judgment improvement, system anchoring, and trust in the system).

5.6.1 Judgment Accuracy

Multiple judgments from each participant made available the use of a within-subjects design in addition to the traditional between-subjects design. A within-subjects design has the benefit of greater power than normal between-subjects designs as the error is reduced by the inclusion of multiple measurements from the same participant [65]. The within-subjects design also provided insight into any changes that may have occurred through time as the participants made repeated judgments. Therefore, analysis of accuracy took place by means of a three-way ($2 \times 3 \times 10$) mixed-model ANOVA with training and access to BAP type as between-subjects factors and judgment number as the within-subjects factor.

In addition to the training and BAP access factors, the initial model included covariates that were considered likely to affect the participants' judgments. These covariates included level of comfort with a computer, age, and years of education after high school. However, multivariate tests (Wilks' λ) of these covariates did not return significant results (Level of comfort with a computer $F(9, 165) = .807$, $p = .610$; age $F(9, 165) = .669$, $p = .736$; years of education after high school $F(9, 165) = 1.043$, $p = .408$). Therefore, these covariates were excluded from analysis and the mixed-model ANOVA was performed again.

The mixed-model ANOVA conformed to assumptions of covariance equality of the dependent measure across groups (Box's $M = 314.3$, $F(275, 48162.2) = .981$, $p > .05$) and sphericity (Mauchly's $W = .763$, $\chi^2(1, N = 44) = 47.6$, $p > .05$). For interpreting between-subjects findings, homogeneity of variance was also examined using Levene's test. Levene's test of the within-subjects factor indicated that there were significant violations of equality of error variance between the judgments of each interview. Therefore, the conservative Tamhane's T^2 post hoc test ($p < .05$) was adopted to examine between-subjects effects [85].

Table 5-8 shows the mean accuracy rates (%) across the conditions. Consistent with past research in deception detection (e.g., [74]), judgment accuracy is reported not only in total, but also within the truthful and deceptive categories (see Table 5-9 and Table 5-10).

Table 5-8 Mean accuracy rates of all final judgments

	No Training	Training
No BAP Access	44.8	48.7
Access to Lower-accuracy BAP	55.2	53.2
Access to Higher-accuracy BAP	58.7	61.0

Table 5-9 Mean final judgment accuracy rates of truthful interviews

	No Training	Training
No BAP Access	40.0	40.0
Access to Lower-accuracy BAP	54.8	53.5
Access to Higher-accuracy BAP	61.3	60.0

Table 5-10 Mean final judgment accuracy rates of deceptive interviews

	No Training	Training
No BAP Access	49.7	57.4
Access to Lower-accuracy BAP	55.5	52.9
Access to Higher-accuracy BAP	56.0	61.9

Examination of the between-subject effects indicates that system access significantly affected judgment accuracy, $F(2, 179) = 12.344$, $p < .001$, partial $\eta 2 = .121$. In support of H1a, Tamhane post hoc tests showed that individuals with access to the lower-accuracy BAP ($M = 54.2$, $SD = 13.9$) demonstrated a significantly higher accuracy rate than unaided individuals ($M = 46.8$, $SD = 15.9$). In support of H1b, Tamhane post hoc tests showed that individuals with access to the higher-accuracy BAP ($M = 59.8$, $SD = 13.8$) demonstrated a significantly higher accuracy rate than individuals with access to the lower-accuracy BAP.

Examination of the between-subjects effects do not indicate that training significantly affected judgment accuracy, $F(1, 179) = .432$, $p = .512$. This finding fails to support H5. Further, the between-subjects interaction effect (Training × Access) was also not significant, $F(2, 179) = .665$, $p = .521$.

As a within-subject design was used to analyze judgments of individual interviews, mean accuracy rates are reported for each interview. While the accuracy rates are reported according to interview number, the interviews were randomly ordered during the experiment. Mean accuracy rates for each interview along with recommendations of the BAP in the training, higher- and lower-accuracy conditions are shown in Table 5-11.

Table 5-11 Mean accuracy rates for each interview

	Interview No.										
Treatment	1	2	3	4	5	6	7	8	9	10	Mean
No BAP No training	32.3	29.0	58.1	83.9	19.4	32.3	19.4	61.3	54.8	58.1	44.8
No BAP Training	54.8	19.4	51.6	74.2	12.9	51.6	29.0	64.5	61.3	67.7	48.7
High acc BAP No training	56.7	73.3	80.0	56.7	16.7	86.7	16.7	50.0	76.7	73.3	58.7
High acc BAP Training	58.1	54.8	77.4	54.8	9.7	80.6	38.7	77.4	71.0	87.1	61.0
Low acc BAP No training	64.5	77.4	74.2	38.7	16.1	64.5	61.3	41.9	77.4	35.5	55.2
Low acc BAP Training	41.9	58.1	71.0	54.8	12.9	64.5	54.8	61.3	74.2	38.7	53.2
BAP Results											
High acc BAP	G	I	I	I	G	I	G	I	G	G	80.0
Low acc BAP	G	I	I	I	G	I	G	G	G	I	60.0
Actual	G	I	I	G	I	I	G	I	G	G	

Examination of the within-subjects factor reveals that judgment number exhibits a significant effect on judgment accuracy, $F(9, 1611) = 2.517$, $p = .007$, partial $\eta 2 = .014$. Figure 5-2 displays mean accuracy rates for each judgment number. The mean accuracy rates for each judgment hover between 50% and 60%, with notable exceptions at judgment number 5 and judgment number 9.

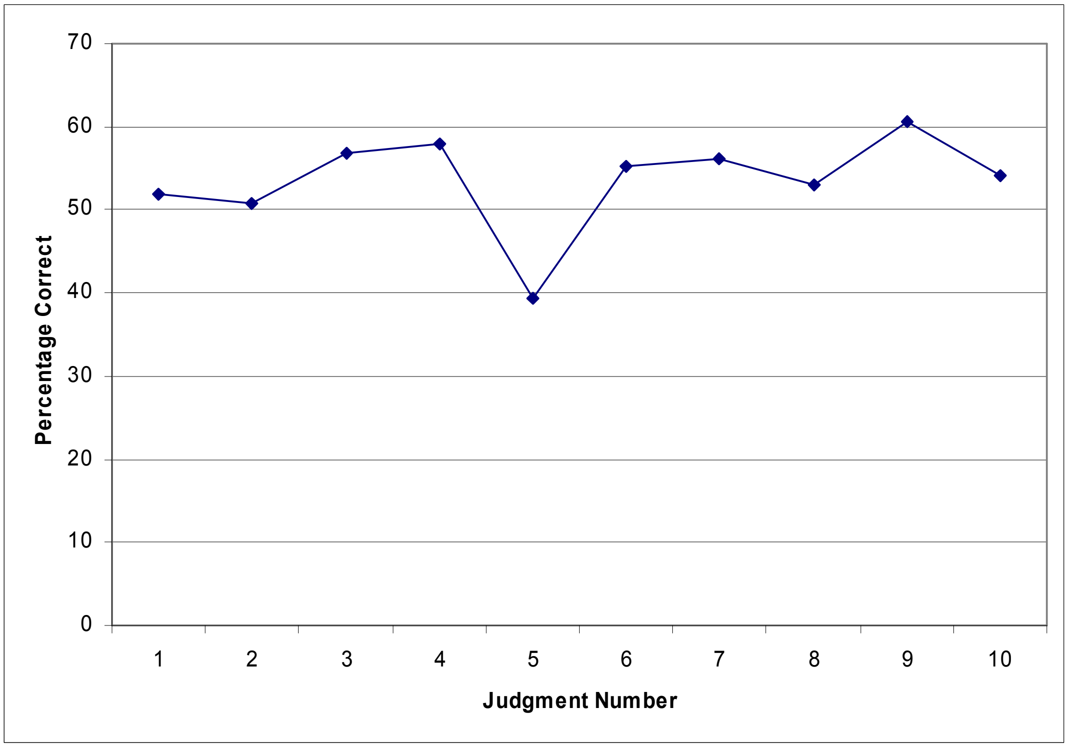


Figure 5-2 Change in judgment accuracy over detection task.

The nature of this effect was examined using a Bonferroni adjusted multiple comparison test ($p < .05$). Results indicated that judgment 5 (M = 39.5, SD = 49.0) was significantly lower than judgment 3 (M = 56.8, SD = 49.7), judgment 4 (M = 57.8, SD = 49.5), judgment 7 (M = 56.2, SD = 49.7), and judgment 9 (M = 60.5, SD = 49.0). A comparison between judgment 1 (M = 51.9, SD = 50.1) and judgment 10 (M = 54.0, SD = 50.0) did not yield significant results, suggesting that the amount of improvement that occurred during the detection task was negligible. The lack of improvement was underscored by an examination of the orthogonal polynomials which showed no significant linear trend.

Consistent with the between-subjects findings, the within-subjects interactions Judgment Number × Access (F(18, 1611) = 2.657, p < .001, partial η2 = .029) and Judgment Number × Access × Training (F(18, 1611) = 1.702, p = .033, partial η2 = .019) were found to be significant. The interaction Judgment Number × Training (F(9, 1611) = .792, p = .624) was not found to be significant.

5.6.2 Confidence Alignment

To measure the level of alignment between judgment accuracy and judgment confidence, a measure called the Mean Probability Score (MPS) was adapted and used. Brier [11] proposed the Probability Score (PS) as a measure that evaluates individuals' probability estimates of an event actually occurring. The PS is calculated in hindsight (i.e., after the event has occurred or has not occurred) and compares the binary outcome d (1 = the event occurred; 0 = the event did not occur) with the probability estimate to arrive at a score between 0 (perfect prediction) and 1 (counter-perfect accuracy). This calculation is shown in Equation 5-1 where $P'(A)$ is the probability estimate of the occurrence of event A.

Equation 5-1 $$PS = (P'(A) - d)^2$$

The MPS is the average of PSs taken over numerous judgments (see Equation 5-2) and is a gauge of probabilistic judgment consistency and quality [146]. It also effectively captures individuals' levels of confidence about their judgments. As individuals are more certain about their predictions, they will assert higher or lower probabilities based on whether or not they think an event will or will not occur. The degree to which these probabilities match actual outcomes is neatly summarized by the MPS.

Equation 5-2 $$MPS = \overline{PS}$$

In the deception detection task, probabilities of guilt or innocence were not explicitly requested from the novices. In place of probabilities, a judgment (guilty or innocent) and level of confidence (0-10) were requested. The judgment and confidence level were therefore combined to a single measurement that approximates a probability judgment. For example, a judgment of guilty with complete confidence would score 1; a judgment of innocent with full confidence would score a 0; and judgments of guilty or innocent with no confidence would both score .5. This transformation occurred as shown in Table 5-12.

Table 5-12 Steps in creating an approximate PS

PS = .5
If judgment = guilty
PS = PS + (Confidence Level)(.05)
Else
PS = PS - (Confidence Level)(.05)

Following the calculation of each PS, the MPS was calculated according to Equation 5-2. Then, the approach adopted in exploring judgment accuracy was utilized in examining MPS. Analysis of MPS took place by means of a three-way 2 (Training) × 3 (Access to BAP type) × 10 (Judgment Number) mixed-model ANOVA with judgment number as the within-subjects factor.

As with judgment accuracy, level of comfort with a computer, age, and years of education after high school were included in the initial model as covariates because they were thought to affect confidence alignment. Multivariate tests (Pillai's Trace) of these covariates did not yield significant results (Level of comfort with a computer $F(9, 165) = .912$, $p = .516$; age $F(9, 165) = 1.892$, $p = .056$; years of education after high school $F(9, 165) = .852$, $p = .569$). Therefore, these covariates were excluded from analysis and the mixed-model ANOVA was performed again.

The resultant mixed-model ANOVA violated assumptions of covariance equality of the dependent measure across groups (Box's $M = 433.7$, $F(275, 48162.7) = 1.353$, $p < .001$). Therefore Pillai's Trace was used to analyze the multivariate tests [85]. The mixed-model ANOVA also violated assumptions of sphericity (Mauchly's $W = .690$, $\chi^2(1, N = 44) = 65.195$, $p = .021$). Therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. For interpreting between-subjects findings, homogeneity of variance was also examined using Levene's test. Levene's test of the within-subjects factor indicated that there were significant violations of equality of error variance between the judgments of each interview. Therefore, Tamhane's T^2 post hoc test ($p < .05$) was adopted to examine between-subjects effects [85]. Table 5-13 shows the MPSs across the conditions.

Table 5-13 MPSs of final judgments

	No Training	Training
No BAP Access	.408	.368
Access to Lower-accuracy BAP	.316	.317
Access to Higher-accuracy BAP	.288	.269

Examination of the between-subjects effects indicates that system access significantly affected the MPS, $F(2, 179) = 17.687$, $p < .001$, partial $\eta2 = .165$. In support of H2a, Tamhane post hoc tests showed that individuals with access to the lower-accuracy BAP ($M = .316$, $SD = .112$) demonstrated a significantly lower MPS than unaided individuals ($M = .388$, $SD = .114$). Tamhane post hoc tests showed that individuals with access to the higher-accuracy BAP ($M = .219$, $SD = .114$) did not demonstrate significantly lower MPSs than individuals with access to the lower-accuracy BAP. However, examination of the significance levels showed that while the MPSs for the higher-accuracy condition were not lower than the MPSs for the lower-accuracy condition at the $p = .05$ level, the difference was significant at the $p = .10$ level. This significant difference provides weak support for H2b.

Further examination of the between-subjects effects did not indicate that training significantly affected MPS, $F(1, 179) = 1.591$, $p = .209$. This finding fails to support H6. Further, the between-subjects interaction effect (Training × Access) was also not significant, $F(2, 179) = .627$, $p = .535$.

MPSs for each interview along with recommendations of the BAP in the training, higher- and lower-accuracy conditions appear in Table 5-14. While the MPSs reported here are organized by the interview number, the order of the interviews seen in the experiment was randomized.

Table 5-14 MPS for each interview

	Interview No.										
Treatment	1	2	3	4	5	6	7	8	9	10	Mean
No BAP No training	.501	.437	.294	.160	.573	.472	.640	.298	.372	.329	.408
No BAP Training	.356	.549	.325	.249	.571	.305	.545	.273	.292	.213	.368
High acc BAP No training	.334	.213	.175	.274	.587	.121	.462	.310	.193	.212	.288
High acc BAP Training	.311	.310	.154	.299	.543	.132	.429	.162	.235	.116	.269
Low acc BAP No training	.300	.205	.147	.386	.549	.241	.288	.406	.195	.440	.316
Low acc BAP Training	.370	.298	.215	.296	.553	.239	.345	.254	.206	.395	.317

Examination of the within-subjects factor reveals that judgment number exhibits a significant effect on MPS, $F(8.990, 1609.189) = 2.308$, $p = .014$, partial $\eta 2 = .013$. Figure 5-3 displays mean accuracy rates for each judgment number.

Figure 5-3 Change in MPS over detection task

The nature of this effect was examined using a Bonferroni adjusted multiple comparison test ($p < .05$). Results indicated that judgment 5 (*M* = .415, *SD* = .324) was significantly more than judgment 4 (*M* = .294, *SD* = .308), judgment 7 (*M* = .297, *SD* = .300), and judgment 9 (*M* = .294, *SD* = .313). A comparison between judgment 1 (*M* = .337, *SD* = .335) and judgment 10 (*M* = .326, *SD* = .317) did not yield significant results, suggesting that the amount of improvement that occurred during the detection task was negligible. The lack of improvement was underscored by an examination of the orthogonal polynomials which showed no significant linear trend.

Consistent with the between-subjects findings, the within-subjects interaction Judgment Number × Access ($F(17.98, 1609.189) = 2.204$, $p = .003$, partial $\eta2 = .024$)

was significant. However, the interaction Judgment Number × Access × Training ($F(17.98, 1609.189) = 1.323$, $p = .163$) was not significant. The interaction Judgment Number × Training ($F(8.99, 1609.189) = .842$, $p = .577$) was also not significant.

5.6.3 Judgment Improvement

Those who had access to the BAP were required to submit an initial judgment prior to viewing the BAP results. The accuracy rates of the initial judgments from the users of the higher- and lower-accuracy BAP are shown in Table 5-15. For comparison purposes, the accuracy rates of those without BAP access are also included in Table 5-15. Judgment accuracy rates of truthful and deceptive interviews are shown in Table 5-16 and Table 5-17, respectively.

Table 5-15 Mean accuracy rates of all initial judgments

	No Training	Training
No BAP Access	44.8	48.7
Access to Lower-accuracy BAP	48.7	46.5
Access to Higher-accuracy BAP	47.7	46.8

Table 5-16 Mean initial judgment accuracy rates of truthful interviews

	No Training	Training
No BAP Access	40.0	40.0
Access to Lower-accuracy BAP	44.5	46.5
Access to Higher-accuracy BAP	50.0	44.5

Table 5-17 Mean initial judgment accuracy rates of deceptive interviews

	No Training	Training
No BAP Access	49.7	57.4
Access to Lower-accuracy BAP	52.9	46.5
Access to Higher-accuracy BAP	45.3	49.0

The MPSs of the initial judgments from the users of the higher- and lower accuracy BAP are shown in Table 5-18. Again, for comparison purposes, the MPSs for those who did not have access to the BAP are also shown in Table 5-18.

Table 5-18 MPSs of initial judgments

	No Training	Training
No BAP Access	.408	.368
Access to Lower-accuracy BAP	.348	.385
Access to Higher-accuracy BAP	.370	.368

To examine whether the accuracy rates from initial judgments of BAP users differed from unaided judgments, a two sample *t*-test was performed (two-tailed). There was no significant difference between the initial judgment accuracy of the users and the judgment accuracy of the non-users, $t(183) = .292$, $p = .771$. Users' initial MPSs were also examined via a two sample *t*-test (two-tailed). There was no significant difference in MPS between initial users' judgments and non-users' judgments, $t(183) = 1.23$, $p = .217$. These findings answer RQ1.

Analysis of judgment improvement for the BAP users took place by means of two three-way (Training × Access to BAP type × Initial-final Judgment) mixed-model

ANOVAs with initial and final judgments as the within-subjects factor. The first ANOVA involved initial and final correct judgments (judgment accuracy improvement) and the second ANOVA involved initial and final MPSs (confidence alignment improvement).

In the first ANOVA, the analysis conformed to assumptions of covariance equality of the dependent measure across groups (Box's M = 5.577, F(9, 161734.2) = .601, p = .797). However, the ANOVA violated assumptions of sphericity; therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. Finally, Levene's test did not indicate a violation of equality of the error variance across groups.

Examination of the within-subjects indicated that there was a significant difference in judgment accuracy between users' initial (unaided) judgment and the final (aided) judgments, $F(1, 119) = 73.324$, $p < .001$, partial $\eta^2 = .381$. This test indicated that final, aided judgments (M = 57.0, SD = 14.1) were significantly more accurate than initial, unaided judgments (M = 47.4, SD = 12.5).

The interaction effect between initial-final judgment and access to BAP type was also significant $F(1, 119) = 7.115$, $p = .009$, partial $\eta^2 = .056$. Other interaction effects, Initial-final Judgment × Training ($F(1, 119) = .614$, $p = .435$) and Initial-final Judgment × Access to BAP type × Training ($F(1, 119) = .409$, $p = .523$), were not significant.

Examination of the between-subjects effects indicated that there were no significant effects for training, $F(1, 119) = .108$, $p = .743$, or access to BAP type $F(1, 119) = 1.539$, $p < .217$.

Second, the mixed-model ANOVA concerning the improvement in alignment between judgment accuracy and judgment confidence conformed to assumptions of covariance equality of the dependent measure across groups (Box's M = 11.935, F(9, 161734.2) = 1.286, p = 239). However, the alignment ANOVA also violated assumptions of sphericity; therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. Finally, Levene's test did not indicate a violation of equality of the error variance across groups.

Examination of the within-subjects indicated that there was a significant difference in confidence alignment between users' initial (unaided) judgment and the final (aided) judgment, $F(1, 119) = 116.731$, $p < .001$, partial $\eta^2 = .495$. This test indicated that confidence levels of final, aided judgments ($M = .298$, $SD = .099$) were significantly more aligned with accuracy than confidence levels of initial, unaided judgments ($M = .368$, $SD = .094$).

The interaction effect between initial-final judgment and access to BAP type was significant, $F(1, 119) = 10.016$, $p = .002$, partial $\eta^2 = .078$, as was the interaction effect between initial-final judgment and training, $F(1, 119) = 3.959$, $p = .049$, partial $\eta^2 = .032$. The final interaction effect, Initial-final Judgment × Access to BAP type × Training ($F(1, 119) = .567$, $p = .453$), was not significant.

Examination of the between-subjects effects indicated that there were no significant effects for training, $F(1, 119) = .067$, $p = .797$, or access to BAP type $F(1, 119) = 1.169$, $p < .282$.

The significant improvement in both judgment accuracy and confidence alignment between initial (unaided) and final (aided) judgments supports H3. However, the lack of significant improvement in judgment accuracy and confidence alignment between users of the higher- and lower-accuracy systems does not support H4.

5.6.4 System Anchoring

When users' initial judgments disagreed with the recommendations of the system, that conflict had to be resolved before a final judgment could be achieved. To examine the tendencies of novices in how they resolve this conflict, the number of judgments where initial judgments contradicted system recommendations were tallied. The final judgment was then reviewed to determine if the user retained his or her original judgment or deferred to the BAP. If, during the course of viewing the 10 interviews, the user resolved this conflict more often by reverting back to his or her judgments, then the user was classified as *human-anchored*. Conversely, if the user accepted the recommendations of the system more often, then the user was classified as *system-anchored*.

Figure 5-4 shows the percentages of novice users who were classified as human-anchored (72%) or system-anchored (19%). In addition to these two groups, there were a smaller number of novices who evenly split between accepting the BAP recommendations and retaining their initial judgments (9%). The number of individuals who demonstrate human-anchoring significantly outnumber those who demonstrate system-anchoring, $\chi^2(1, N = 123) = 48.203$, $p < .001$. This finding contradicts H7.

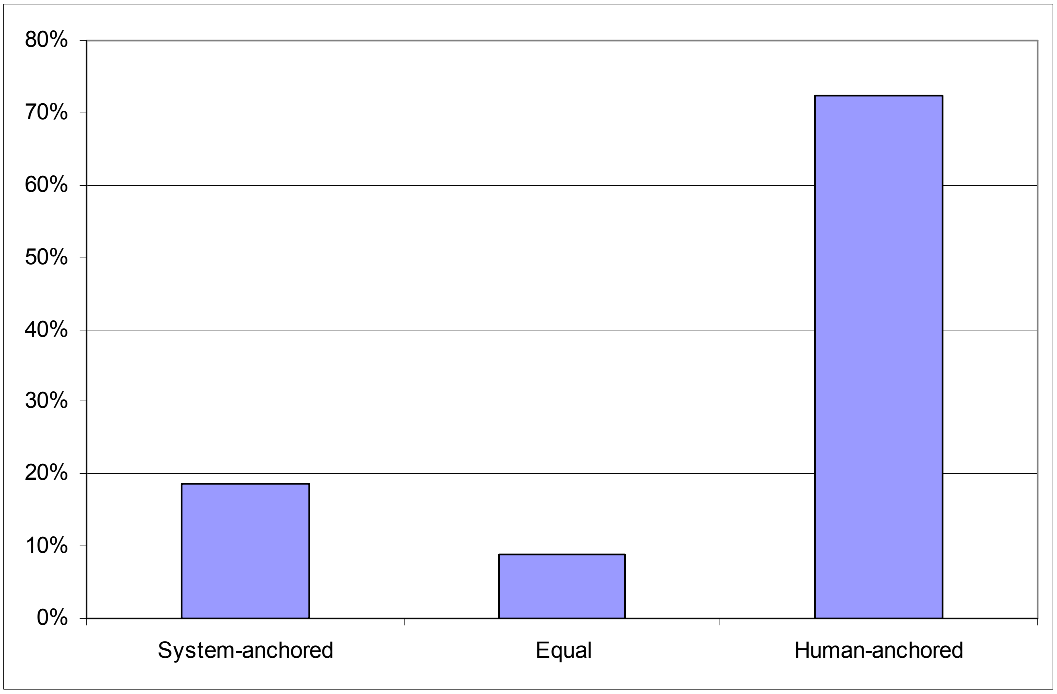


Figure 5-4 Human and system anchoring in novices

When users' initial judgments disagreed with the system recommendations, there was an opportunity to investigate the reasoning behind the recommendation. The reasoning behind the function of the BAP was revealed through explanations of three types: definitions, cues, and analysis (see section 4.4 for a more detailed description of the BAP explanations). The mean numbers of definitions, cues, and analysis explanations accessed in each condition appear in Figure 5-5.

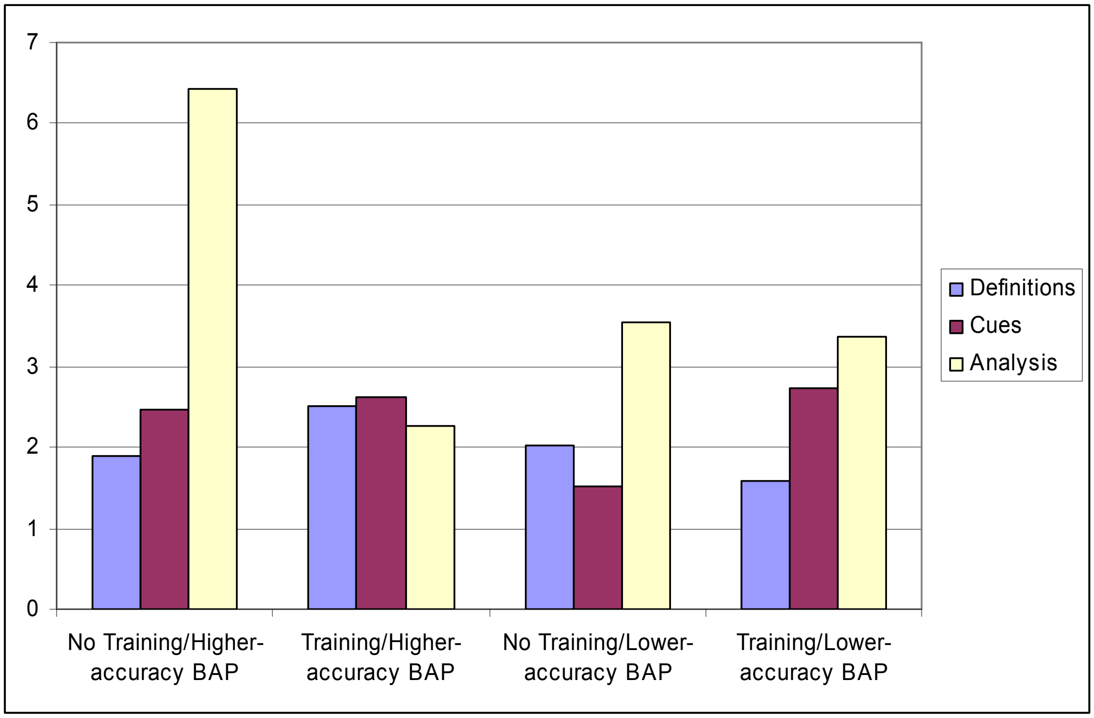


Figure 5-5 Means of BAP explanations accessed

Examination of the number of explanations accessed during the experiment was first attempted via a bivariate Pearson's correlation analysis. Results from this analysis indicated that the number of cues explanations were significantly correlated with the number of definitions explanations, $r(123) = .248$, $p = .006$, and with the number of analysis explanations, $r(123) = .237$, $p = .008$. As a result of possible related dependent measures, a two-way (training × access to BAP type) between-subjects MANOVA was utilized to examine explanation usage.

Review of the data showed that the number of explanations viewed in each category violated assumptions of normality by demonstrating heavy skewness. Adoption of a square-root transformation of the number of definitions, cues, and analysis explanations remedied this violation. Box's *M* test (p = .579), indicated variance-covariance equality across factors. Levene's test indicated inequality between the error variances of the number of analysis explanations across groups. However, Levene's test demonstrated equality of error variances for the number of other explanations (definitions and cues). Therefore, the Bonferroni procedure (α = .05) was adopted for pairwise comparisons.

Using Wilks' λ, the dependent variate was not significantly affected by access to system type, λ = .988, $F(3, 117)$ = .458, p = .712. Thus, there were no differences in the number of explanations viewed by novices across conditions. This finding answers RQ2.

Interestingly, there was a significant effect from training, λ = .935, $F(3, 117)$ = .2.703, p = .045. To explore the nature of this finding, tests of between-subjects effects revealed that training significantly affected the number of analysis explanations that were viewed, $F(1, 119)$ = 4.257, p = .041, partial η^2 = .035. Pairwise comparisons using the Bonferroni adjustment for multiple comparisons indicated that those who were untrained requested more analysis explanations[13] (M = 1.67, SD = 1.49) than those who were trained (M = 1.17, SD = 1.17).

5.6.5 Trust in the System

Trust can be a difficult and elusive construct to capture accurately. It has many definitions and varies depending on the context in which it appears. (See [79] for a

---

[13] Recall that the number of explanations was transformed because of non-normality. The means reported reflect the square-root transformation. This difference also clearly appears in Figure 5-5, which does not reflect the transformation.

comprehensive review of the trust construct.) The analysis and discussion of trust for the purposes of this dissertation are limited to the view and opinions of the BAP user. In this context, trust is an indication of a user's willingness to accept the BAP and all of the benefits and potential risks that this might entail [100].

Opinions about the performance and capability of the BAP were collected via a questionnaire that was administered following the conclusion of the judgment task. The questionnaire was adapted from an existing instrument that was first introduced to study recommender systems and intelligent agents [68] and contained 19 questions[14] scored on a 1 (strongly disagree) to 7 (strongly agree) scale. The wording of the instrument was changed to fit the context of computer-aided deception detection. The instrument included questions concerning four types of perceptions or intentions: emotional trust of the system, intentions to use system as a delegated agent, intentions to use the system as a decision aid, and competence of the system. Opinions about use of the BAP as a delegated agent or as a decision aid concern the users' perspective about the proper positioning of the BAP along the Sheridan and colleagues' levels of automation (refer to Figure 3-4). If a user is willing to use the BAP as a delegated agent, the user is willing to allow the BAP to operate in his or her place without significant supervision. Conversely, if a user is willing to use the BAP only as a decision aid, then the BAP may be consulted

[14] An additional four questions were included in the final novice questionnaire that concerned BAP user demographics.

during the judgment-making process but control over the judgment will be retained by the user.

Measures indicating the reliability and validity of the items included in the questionnaire appear in Appendix D. The constructs demonstrated acceptable reliability and convergent and discriminant validity with the exception of system competence. Therefore system competence was excluded from analysis.

Emotional trust, delegated agent intention, and decision aid intention were expected to be correlated [68]; therefore, examination of user trust commenced with a bivariate Pearson's correlation analysis. Results from this analysis indicated that emotional trust significantly correlated with intention to use the BAP as a decision aid, $r(123) = .405$, $p < .001$, and with intention to use the BAP as a delegated agent, $r(123) = .590$, $p < .001$. As a result of possible related dependent measures, a two-way (training × access to BAP type) between-subjects design MANOVA was utilized to examine users' levels of trust in the BAP. The mean factor scores for emotional trust, delegated agent intention, and decision aid intention in each condition are shown in Figure 5-6.

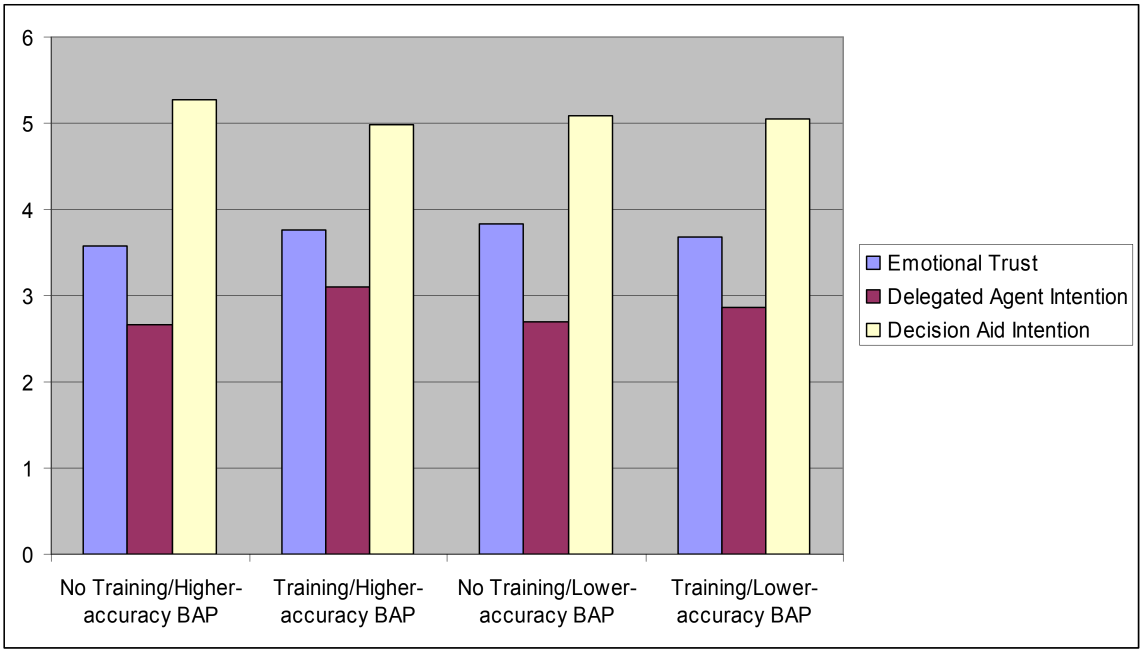


Figure 5-6 Mean factor scores for emotional trust, delegated agent intention, and decision aid intention

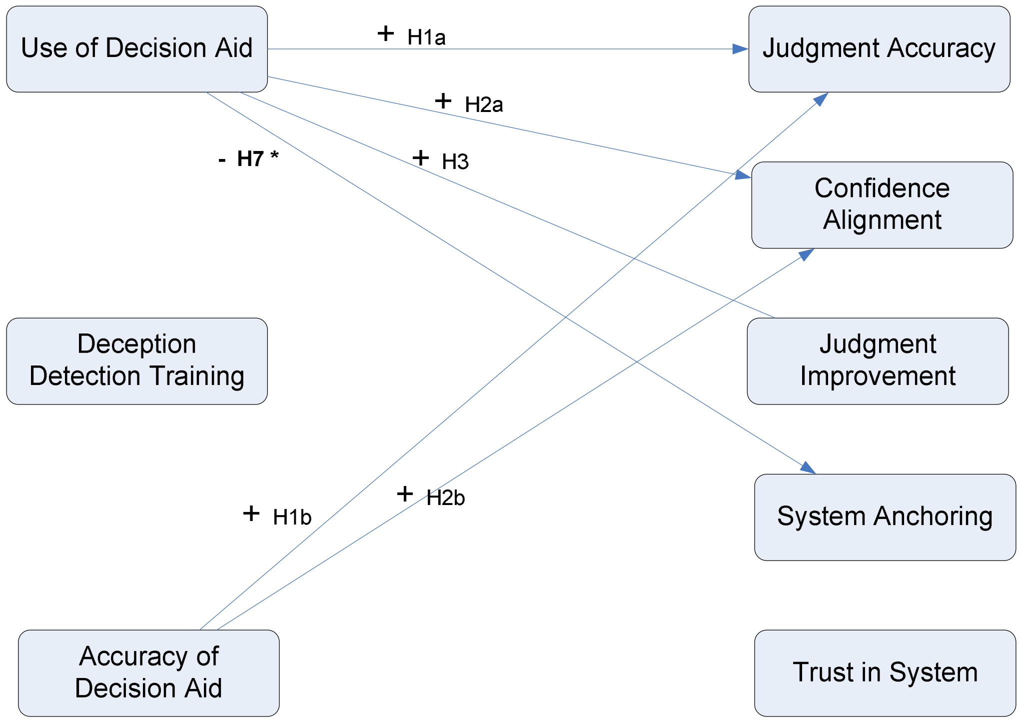
There were no violations of statistical assumptions in the data and Box's *M* test (p = .410) indicated variance-covariance equality across factors. Using Wilks' λ, the dependent variate was not significantly affected by access to system type, $\lambda = .990$, $F(3, 117) = .397$, $p = .755$ or training, $\lambda = .976$, $F(3, 117) = .968$, $p = .410$. Thus, there were no differences in levels of emotional trust, willingness to use the BAP as a delegated agent, and willingness to use the BAP as a decision aid across conditions. This finding answers RQ3.

5.7 Discussion

Findings concerning judgment accuracy, confidence alignment, improvement in judgment, system, anchoring, and trust in the system are summarized in Table 5-19. A graphical summary displaying the supported, hypothesized links between independent and dependent variables appear in Figure 5-7.

Table 5-19 Summary of findings for novice computer-aided deception detection

H1a: Novice users of the lower-accuracy system will demonstrate higher judgment accuracy than unaided novices.	Supported
H1b: Novice users of the higher-accuracy system will demonstrate higher judgment accuracy than novice users of the lower-accuracy system.	Supported
H2a: Novice users of the lower-accuracy system will demonstrate more alignment between judgment accuracy and confidence in judgment than unaided novices.	Supported
H2b: Novice users of the higher-accuracy system will demonstrate more alignment between judgment accuracy and confidence in judgment than novice users of the lower-accuracy system.	Weakly supported
RQ1: What is the difference between the accuracy and alignment of the unaided novice users' judgments and the accuracy and alignment of the aided novice users' initial judgments?	No difference
H3: Novice users will see improvement in judgment accuracy and alignment between initial and final judgments.	Supported
H4: Novice users of the higher-accuracy system will see greater improvement in judgment accuracy and alignment between initial and final judgments than novice users of the lower-accuracy system.	Not supported
H5: Trained novices will demonstrate higher judgment accuracy than untrained novices.	Not supported
H6: Trained novices will demonstrate more alignment between judgment accuracy and confidence in judgment than untrained novices.	Not supported
H7: When novices' initial judgments conflict with the system recommendations, the novices will anchor on the system recommendations.	Contradicted
RQ2: What is the difference in the number of explanations viewed by the novice users of the higher-accuracy system and the novice users of the lower-accuracy system?	No difference
RQ3: What is the difference in the emotional trust in the system, willingness to use the system as a decision aid, and willingness to use the system as an autonomous agent among novice users of the higher-accuracy system and the novice users of the lower-accuracy system?	No difference



* Opposite from hypothesized direction

Figure 5-7 Supported hypotheses for novice computer-aided deception detection

Novice lie-catchers without access to the BAP and without training in deception detection clearly demonstrated the problems that many who assess credibility face: their accuracy in detecting deception is low and their confidence is poorly calibrated to their performance. Within the context of this experiment, these experimental participants would have been better off flipping an unbiased coin to determine guilt or innocence, and their MPSs would be classified as poor at best [146].

This poor performance is also present in the initial judgments of participants who have access to the BAP but who have not had the opportunity to view the recommendations of the BAP. Accuracy and confidence alignment are no better for users' initial judgments than unaided novices' judgments. This finding discounts the possibility that users may have been preconditioned to more vigilantly monitor the behavior of the interviewees they were observing. Thus, the improvement in judgment performance of users of the BAP may more fully be attributed to use of the BAP.

The improvement in judgment performance was, perhaps, the most robust finding of this experiment. Users of the BAP demonstrated significant improvement in their ability to detect deception and align their confidence with their judgment accuracy. This finding was supported not only with a between-subjects examination, but also a within-subjects examination. Thus, the participants improved over their own unaided judgments. For example, those in the trained/higher-accuracy BAP condition demonstrated a 16.2% increase in accuracy over their untrained, unaided counterparts. Further, they demonstrated 14.2% increase over their own unaided judgments. Those with access to the lower-accuracy BAP also performed better than their unaided counterparts. However the improvement was less pronounced.

Interestingly, the BAP alone correctly classified 80% of the interviewees correctly in the higher-accuracy condition and 60% of the interviews in the lower-accuracy condition. Yet novices using the BAP correctly classified between 53-61% of the interviews correctly. This finding underscores the difficulty that humans face when detecting deception and draws into question the notion of a human-computer system for improved detection of deception (as illustrated by Figure 3-7). Further, this point may also invite thoughts of an automated detection system that excludes any interface with a human user.

While the experimental findings seem to support the exclusion of the human user, total exclusion is not condoned. Once paired with the BAP, the human user must also be allowed to grow accustomed to the performance and function of the BAP. It is possible that the benefits of coupling skilled human detectors with computer aids may only come as human responsibilities are clearly delineated and the functionality of the system is reliably understood (see Figure 3-10). The novice users of the BAP may not have had time to understand how they can work effectively with the BAP.

Further, there are many critical functions that a human user would perform in a field environment that were not incorporated in the experiment. For example, human responsibilities may be delineated in a manner similar to that of a polygraph operator, where the user would be responsible for the interrogation questions and the interpretation of the system results. In this scenario, the human user is aware of all the aspects of an interview or interrogation, but the BAP would only be useful when deemed appropriate by the human user. There are other potential divisions of responsibilities that may be possible and many of them are listed in Figure 3-4. However, before delineation of human and computer responsibilities are made, the accuracy and reliability of the BAP must be determined within that scenario. Only then can the capabilities of a human user be effectively paired with the BAP.

In the experiment, training did not improve judgment accuracy and did not improve alignment between confidence and judgment accuracy. However, training did affect the number of analysis explanations that users requested. This effect brings up the possibility that users thought they understood how the BAP operated as a result of the training.

Therefore, they did not seek analysis explanations as often. This seems likely as the BAP utilizes many of the concepts and findings discussed in the training, and one who was familiar with those concepts may have found the analysis explanations less informative.

The limited effect of training on judgment performance brings to light an important limitation of this study: the subjects had limited training in deception detection. The training session lasted 23 minutes with approximately 10 minutes of practice. Past research has called into question instruction on credibility assessments in brief sessions before detection tasks [43].

Clearly, training used in this experiment was not meant to be an exhaustive course on deception; however, the training contained valuable information that was useful in the deception task. The deceptive interviewees demonstrated many of the cues that were described in the training, but the novice lie-detectors did not catch these cues. Therefore, we do not discount the importance of training in deception detection, but rather acknowledge potential weakness in training methods. This conclusion is in line with conclusions reached by other researchers who investigate training in deception detection (e.g., [43]; also see Table 3-2).

Every effort was made to conform to the standards of effective training laid out in Table 3-2; however, one may argue that the stakes of the lies observed were low and that the training was not properly conducted. Possible improvements in the training may include live lecture (with question and answer opportunities), more examples of deceptive and truthful interactions under high stakes, and greater repetition of key concepts (perhaps with multiple lectures).

Examination of system anchoring among novice users produced findings that strongly contradicted what was hypothesized. The results indicated that when novices' initial judgments conflicted with the BAP's recommendations, the novices overwhelmingly retained their own judgments. This finding suggests that novice users, despite being told that the BAP's judgment was correct 60-80% of the time, opted to disregard the BAP's recommendation. For those who were trained, the anchoring on human judgment was even more puzzling as the users were explicitly and repeatedly told that humans perform poorly in detecting deception and that humans generally identify deception approximately 50 percent of the time.

One explanation for this finding may be that users demonstrate anchoring and adjustment biases resulting from the request for an initial judgment. Tversky and Kahneman [132] first hypothesized that adjustments to initial judgments are typically insufficient. However, this possibility was discounted by a pilot study described in Appendix E when novice subjects were required to provide only a final judgment.

Another explanation of this finding may be that user perceptions concerning the BAP did not reflect a high opinion of the system's operation. As seen in Figure 5-6, users' levels of trust in the BAP and their intention to the use the BAP as a delegated agent were relatively low. It appeared that the users were willing to consult the recommendation of the BAP but they did not think it competent enough to entrust it with autonomy.

CHAPTER 6

NOVICE AND PROFESSIONAL COMPUTER-AIDED DECEPTION DETECTION

To address the hypotheses outlined in section 3.4.2 and to examine the effects of the BAP system on novice and professional judgments, an experiment was designed and carried out. The experiment compared the performance of trained novices with access to the higher-accuracy BAP to professionals who also had access to training and the higher-accuracy BAP. The novice responses described in Chapter 5 (training/access to higher-accuracy BAP condition) were used and compared with professional responses. Thus, the experiment described in this chapter closely follows the structure and design of the experiment described in Chapter 5.

6.1 Participants

The participants in the novice condition were recruited from two sections of an upper-division Management Information Systems (MIS) courses at the University of Arizona. A total of 30 participants were randomly selected for the training and access to the higher-accuracy BAP condition in Chapter 5. The mean age of the novice participants was 21.5, mean years of secondary education was 3.5, and of all the participants, 45.0 percent were female.

The professional participants were recruited from federal and local law enforcement agencies in the Southwestern United States. A total of 16 participants were recruited from the United States Secret Service; a city police department; the Internal Revenue Service, Criminal Investigation Division; and the Bureau of Alcohol, Tobacco, Firearms, and

Explosives. The mean age of the professional participants was 38, mean years of secondary education was 5, and of all the participants, 12.5 percent were female. The professionals' average number of years of law enforcement experience was 11.4.

6.1.1 Motivation of Participants

The novice participants were recruited to participate in the experiment by public announcements during class. The participants were awarded extra credit toward their final grades for participation (approximately 2% of their final grade). To further motivate the participants, those who scored in the top 10% in each experimental condition were offered a cash reward in the amount of $10.

The professionals were recruited via announcements made during an inter-organizational meeting of local and federal law enforcement agencies. As an incentive for participation, professional participants were offered free training in deception detection following the conclusion of the experiment.

6.1.2 Justification of Sample

The novice participants selected for this study represent untrained, non-professional lie-catchers that could benefit from a decision aid in deception detection. They have been exposed to some forms of deception but do not bear the responsibility of detecting deception on a regular basis. They serve as an effect baseline with which one may contrast the capabilities of professional lie-catchers.

The professional participants selected for this study represent law-enforcement agents who shoulder the responsibility of sifting through numerous interactions in search of truthful accounts of events. Resulting from their experience, the professionals are acquainted with high-stakes deception and have successfully identified high-stakes deception in the past. They are capable of evaluating the BAP in light of their experience and may offer insights into how such a system may be incorporated into other professionals' deception detection efforts.

6.2 Stimulus Materials

The same stimulus materials that were used in the novice-only experiment (see section 5.2) were also used in this study. The stimulus materials were provided by a previous experiment that was conducted at Michigan State University (MSU) in 2004 by Timothy Levine and colleagues [72]. In a variation of the Exline procedure [36], high-stakes, unsanctioned deceptive and truthful interactions were elicited and digitally recorded. The data set contained 22 interviews which were used in the experiment (7 deceptive, 15 truthful). All of the interviews were used for training the BAP and 10 interviews, 5 truthful and 5 deceptive, were selected for stimulus materials during the experiment.

6.3 Experiment Design Matrices

As described in section 3.4.2, a single independent variable was manipulated during this experiment: professional experience in deception detection. Both professionals and novices had access to training in deception detection and also had access to the higher-accuracy BAP. Thus, the experiment utilizes a simple experimental design. The experimental design matrix and sample size in each cell appear in Table 6-1.

Table 6-1 Experimental design matrix for novices and professionals

	Novices	Professionals
Training/Access to Higher-accuracy BAP	$N = 30$	$N = 16$

6.4 Independent Variable

Both the professionals and novices had access to the higher-accuracy BAP. The higher-accuracy BAP in this experiment provided the same explanations and recommendations as were described in the novice-only experiment (refer to sections 4.5.3 and 5.4.2). The professionals and novices also had access to the training described in section 5.4.3 and in Appendix B. The sole independent variable in the experiment described in this chapter was the designation of professional or novice lie-catcher.

6.5 Experimental Methods

6.5.1 Experimental Methods for Novices

Since novice responses from the novice-only experiment were used in this experiment, the methods involving the novices exactly follow the description outlined in section 5.5 . Upon arriving at the lab, the participants were pseudo-randomly assigned to six conditions, one of which was training and access to higher-accuracy BAP condition. The students in this condition viewed an orientation video describing the experiment and the BAP and then completed an initial survey. Then the participants took part in the training which included a knowledge quiz and judgment quiz. Following the training, the participants judged the 10 randomly-ordered interviews. All novices had access to the higher-accuracy BAP. Following the detection task, the participants filled out a final questionnaire and were released. The experiment lasted approximately 1.5 hours for the novices.

6.5.2 Experimental Methods for Professionals

The experimental methods for the professionals closely followed that of the novices with only a few exceptions. Upon arriving at the experiment, the professionals were invited to introduce themselves and the experimenters also introduced themselves. A brief explanation of the experiment was provided orally along with a description of the components of the experiment (e.g., orientation video, computer-based deception detection training, detection task, etc.). Following the introductions and brief explanation, the professionals followed exactly the same procedure as the novices. A single additional question concerning experience level was asked of the professionals (see Appendix D).

Following the completion of the experiment, the professionals were debriefed and the experimenters conducted open discussions about the BAP, the professionals' perceptions about its functionality and usefulness, and the professionals' experiences in determining the veracity of suspects' statements. Following the debriefing and discussion portion, the professionals received additional training concerning human-based deception detection.

6.6 Analysis

Analysis concerning the five dependent measures (judgment accuracy, confidence alignment, judgment improvement, system anchoring, and trust in the system) will be presented sequentially.

6.6.1 Judgment Accuracy

To examine the effect of experience on accuracy in a repeated task, a two-way 2 (Professional/Novice) × 10 (Judgment Number) mixed model ANOVA was performed with judgment number as the within-subjects factor.

The mixed model ANOVA conformed to assumptions of covariance equality of the dependent measure across groups (Box's $M = 95.640$, $F(55, 3117.510) = 1.239$, $p > .05$) and sphericity (Mauchly's $W = .239$, $\chi^2(1, N = 44) = 59.835$, $p > .05$). For interpreting between-subjects findings, homogeneity of variance was also examined using Levene's test. Levene's test of the within-subjects factor indicated that there were significant violations of equality of error variance between the judgments of each interview.

Table 6-2 shows the final mean accuracy rates (%) across the conditions. Consistent with past research in deception detection (e.g., [74]), judgment accuracy is reported not only in total, but also within the truthful and deceptive categories (see
Table 6-3 andTable 6-4).

Table 6-2 Mean accuracy rates of final judgments

	Novices	Professionals
Training/Access to Higher-accuracy BAP	61.0	61.9

Table 6-3 Mean final judgment accuracy rates of truthful interviews

	Novices	Professionals
Training/Access to Higher-accuracy BAP	60.0	57.5

Table 6-4 Mean final judgment accuracy rates of deceptive interviews

	Novices	Professionals
Training/Access to Higher-accuracy BAP	61.9	66.25

Examination of the between-subjects effect indicates that being a professional lie-catcher did not significantly affect judgment accuracy, $F(1, 45) = .046$, $p = .861$. This finding fails to support H9.

As a within-subject design was used to analyze judgments of individual interviews, mean accuracy rates are reported for each interview. While the accuracy rates are reported according to interview number, the interviews were randomly ordered during the experiment. The randomized order of the interviews constitutes the judgment order. Mean accuracy rates for each interview along with recommendations of the BAP are shown in Table 6-5.

Table 6-5 Mean accuracy rates for each interview

	Interview No.										
Treatment	1	2	3	4	5	6	7	8	9	10	Mean
Novices	58.1	54.8	77.4	54.8	9.7	80.6	38.7	77.4	71.0	87.1	61.0
Professionals	75.0	62.5	87.5	75.0	12.5	56.3	25.0	68.8	68.8	87.5	61.9
BAP Results											
High acc BAP	G	I	I	I	G	I	G	I	G	G	80.0
Actual	G	I	I	G	I	I	G	I	G	G	

Examination of the within-subjects factor revealed that judgment number exhibited a significant effect on judgment accuracy, $F(9, 405) = 2.292$, $p = .016$, partial $\eta^2 = .048$. Figure 5-2 displays mean accuracy rates for each judgment number across novice and professional conditions. No interaction effects were significant.

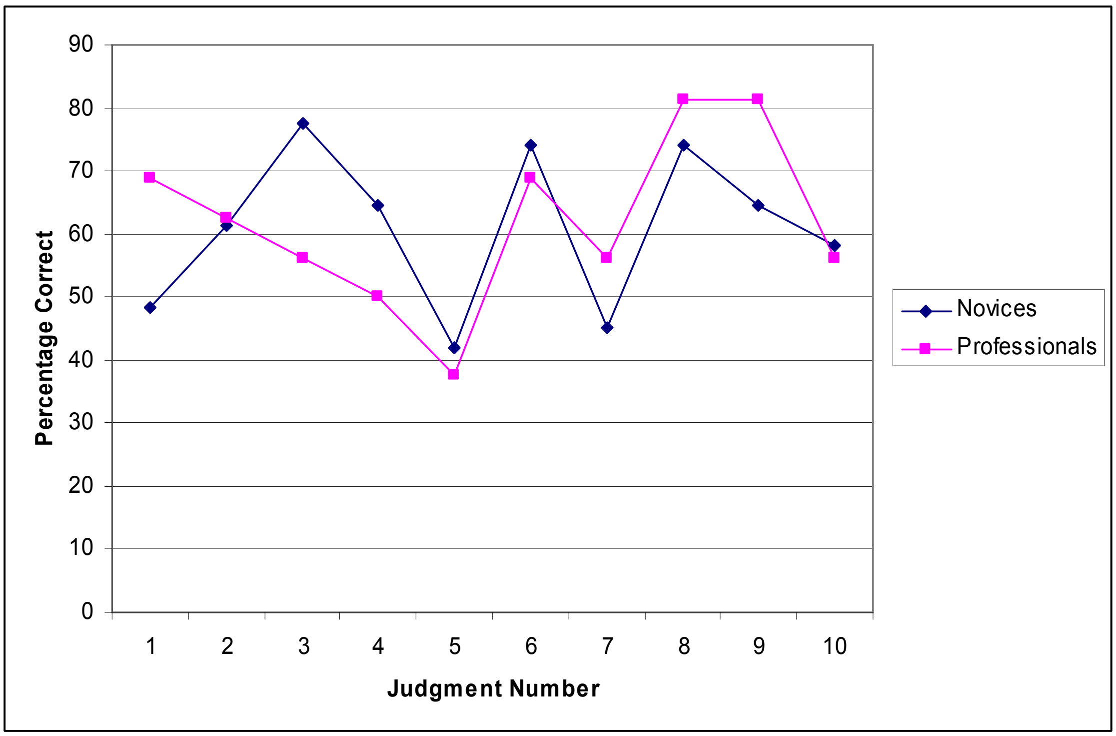


Figure 6-1 Accuracy rates over the detection task

The nature of this effect was examined using a Bonferroni adjusted multiple comparison test ($p < .05$). Results indicated that judgment 5 ($M = 40.43$, $SD = 49.60$) was significantly less than judgment 8 ($M = 76.60$, $SD = 42.80$). A comparison between judgment 1 ($M = 55.32$, $SD = 50.25$) and judgment 10 ($M = 57.45$, $SD = 49.98$) did not yield significant results, suggesting that the amount of improvement that occurred during the detection task was negligible. The lack of improvement was underscored by an examination of the orthogonal polynomials which showed no significant linear trends of accuracy rates for novices or professionals.

6.6.2 Confidence Alignment

To measure the level of alignment between judgment accuracy and judgment confidence among the professionals and novices, the Mean Probability Score (MPS) was again used [146]. Section 5.6.2 contains a detailed description of the MPS and its calculation. It is important to note that the MPS is a direct measure of the level of agreement between one's estimation of the likelihood of an event occurring (i.e,, judgment and confidence level) and the actual occurrence of that event (i.e., deception has occurred). It is scored on a scale of 0 to 1, where 0 is perfect alignment.

The approach adopted in exploring judgment accuracy was utilized in examining MPS. Analysis of MPS took place by means of a two-way 2 (Professional/Novice) × 10 (Judgment Number) mixed-model ANOVA with judgment number as the within-subjects factor.

The resultant mixed-model ANOVA violated assumptions of covariance equality of the dependent measure across groups (Box's M = 129.75, $F(55, 3117.51) = 1.681$, $p = .001$). Therefore Pillai's Trace was used to analyze the multivariate tests [85]. The mixed-model ANOVA also violated assumptions of sphericity (Mauchly's W = .156, $\chi^2(1, N = 44) = 77.68$, $p = .001$). Therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. Levene's test of the within-subjects factor indicated that there were not significant violations of equality of error variance between the judgments of each interview. Table 6-6 shows the final MPSs across the two conditions.

Table 6-6 MPSs of final judgments

	Novices	Professionals
Training/Access to Higher-accuracy BAP	.269	.287

Examination of the between-subject effect indicates that being a professional lie-catcher did not significantly affect alignment between confidence and judgment accuracy, F(1, 45) = .342, p = .562. This finding fails to support H10.

MPSs for each interview in the novice and professional conditions appear in Table 6-7. While the MPSs reported here are organized by the interview number, the order of the interviews seen in the experiment was randomized.

Table 6-7 MPS for each interview

	Interview No.										
Treatment	1	2	3	4	5	6	7	8	9	10	Mean
Novices	.311	.310	.154	.299	.543	.132	.429	.162	.235	.116	.269
Professionals	.190	.248	.155	.198	.607	.266	.532	.271	.237	.164	.287

Examination of the within-subjects factor reveals that judgment number exhibited a significant effect on MPS, $F(7.640, 343.794) = 7.640$, $p = .005$, partial $\eta^2 = .060$. Figure 6-2 displays mean accuracy rates for each judgment number between conditions.

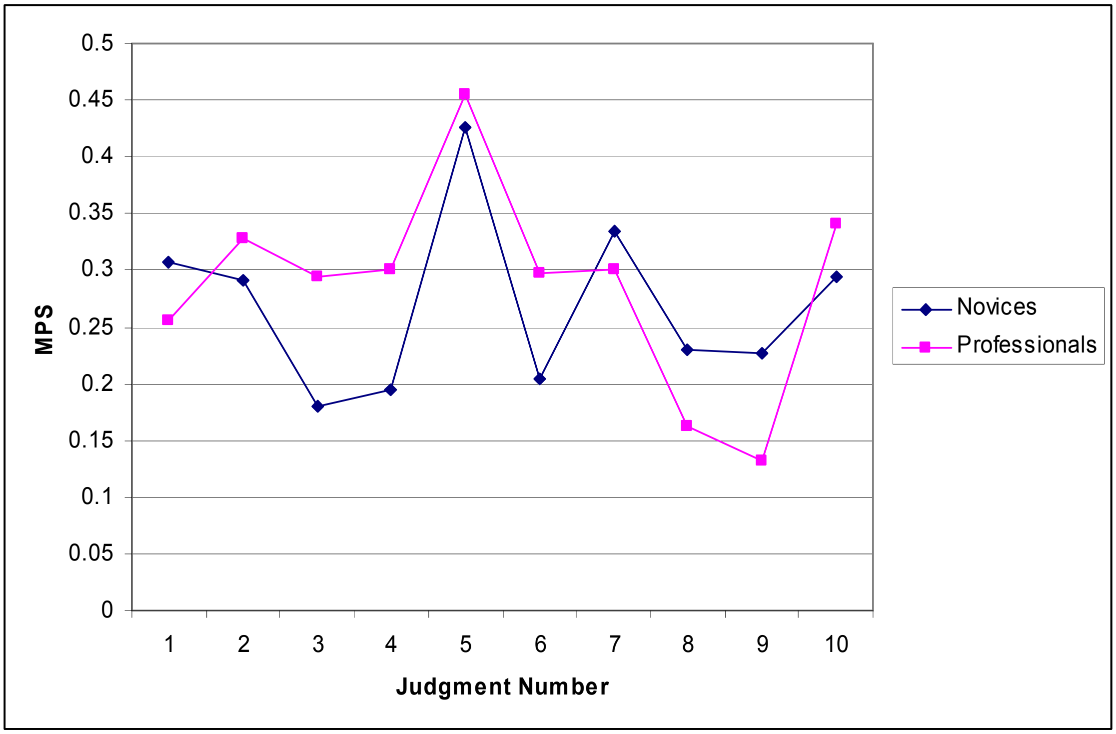


Figure 6-2 Change in MPS over detection task

The nature of this effect was examined using a Bonferroni adjusted multiple comparison test ($p < .05$). Results indicated that judgment 5 (M = .436, SD = .345) was significantly more than judgment 8 (M = .207, SD = .256) and judgment 9 (M = .194, SD = .238). A comparison between judgment 1 (M = .289, SD = .305) and judgment 10 (M = .310, SD = .314) did not yield significant results, suggesting that the amount of improvement that occurred during the detection task was negligible. The lack of improvement was underscored by an examination of the orthogonal polynomials which showed no significant linear trends of MPSs for novices or professionals.

6.6.3 Judgment Improvement

Both the professionals and novices were required to submit an initial judgment before viewing the BAP results. The accuracy rates from the initial judgments of both groups appear in Table 6-8 (see Table 6-9 and Table 6-10 for results with truthful and deceptive interviews). The MPSs of the initial judgments from the professionals and novices are shown in Table 6-11.

Table 6-8 Mean accuracy rates of initial judgments

	Novices	Professionals
Training/Access to Higher-accuracy BAP	46.8	50.0

Table 6-9 Mean initial judgment accuracy rates of truthful interviews

	Novices	Professionals
Training/Access to Higher-accuracy BAP	44.5	45.0

Table 6-10 Mean initial judgment accuracy rates of deceptive interviews

	Novices	Professionals
Training/Access to Higher-accuracy BAP	49.0	55.0

Table 6-11 MPSs of initial judgments

	Novices	Professionals
Training/Access to Higher-accuracy BAP	.368	.354

To determine if there was a difference between the initial accuracy of the novices and professionals, a two sample *t*-test was performed (two-tailed). There was no significant difference between the initial judgment accuracy rates of the professionals and novices, $t(45) = .813$, $p = .421$. Initial MPSs were also examined via a two sample *t*-test (two-tailed). There was no significant difference in MPS between novices' and professionals' initial judgments, $t(45) = .493$, $p = .624$. These findings answer RQ4.

Analysis of judgment improvement for the novices and professionals took place by means of two two-way mixed-model ANOVAs with professional/novice as the between-subjects factor and with initial and final judgments as the within-subjects factor. The first ANOVA involved initial and final correct judgments (judgment accuracy improvement) and the second ANOVA involved initial and final MPSs (judgment alignment improvement).

In the first ANOVA, the analysis conformed to assumptions of covariance equality of the dependent measure across groups (Box's $M = .810$, F(3, 22820.870) = .321, p = .810). However, the mixed-model ANOVA violated assumptions of sphericity; therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. Finally, Levene's test did not indicate a violation of equality of the error variance across groups.

Examination of the within-subjects factor indicated that there was a significant difference in judgment accuracy between users' initial (unaided) judgment and the final (aided) judgment, $F(1, 45) = 48.908$, $p < .001$, partial $\eta^2 = .521$. This test indicated that final, aided judgments ($M = 61.3$, $SD = 13.6$) were significantly more accurate than initial, unaided judgments ($M = 47.9$, $SD = 12.8$). The interaction effect between initial-final judgment and novice/professional was not significant $F(1, 45) = .284$, $p = .537$.

Examination of the between-subjects effects indicated that there were no significant effects for professional experience, $F(1, 45) = .320$, $p = .575$.

Second, the mixed-model ANOVA concerning the improvement in alignment between judgment accuracy and judgment confidence conformed to assumptions of covariance equality of the dependent measure across groups (Box's $M = 7.046$, F(3, 22820.870) = 2.216, p = .084). However, the alignment ANOVA also violated assumptions of sphericity; therefore, degrees of freedom were adjusted for the within-subjects effects using the Huynh-Feldt method. Finally, Levene's test did not indicate a violation of equality of the error variance across groups.

Examination of the within-subjects indicated that there was a significant difference in confidence alignment between users' initial (unaided) judgment and the final (aided) judgment, $F(1, 45) = 95.062$, $p < .001$, partial $\eta^2 = .679$. This test indicated that confidence in final, aided judgments ($M = .275$, $SD = .098$) was significantly more aligned with accuracy than confidence in initial, unaided judgments ($M = .363$, $SD = .092$). The interaction effect between initial-final judgment and novice/professional factor was not significant at the $p < .05$ level; however, the interaction was significant at the $p < .10$ level, $F(1, 45) = 3.507$, $p = .068$, partial $\eta^2 = .072$.

Examination of the between-subjects effects indicated no significant effect for the novice/professional factor, $F(1, 45) = .004$, $p = .949$.

The significant improvement in both judgment accuracy and confidence alignment between initial (unaided) and final (aided) judgments is notable for both novices and professionals. However, the lack of significant difference in judgment accuracy and confidence alignment between novices and professionals does not support H8.

6.6.4 System Anchoring

When users' initial judgments disagree with the recommendations of the system, that conflict had to be resolved before a final judgment could be achieved. To examine the tendencies of novices and professionals in how they resolve this conflict, the number of judgments when initial judgments contradicted system recommendations were tallied. Novices and professionals were classified as *human-anchored* if they tended to retain their initial judgment. Conversely, novices and professionals were classified as *system-anchored* if they tended to adopt the recommendations of the system more frequently.

Figure 6-3 shows the percentage of novices and professionals who were classified as human-anchored, equally anchored, or system-anchored. The percentage of professionals who demonstrated human-anchoring (75%) was significantly larger than the percentage of those who demonstrated system-anchoring (6%), $\chi^2(1, N = 13) = 4$, $p < .046$. This finding supports H12.

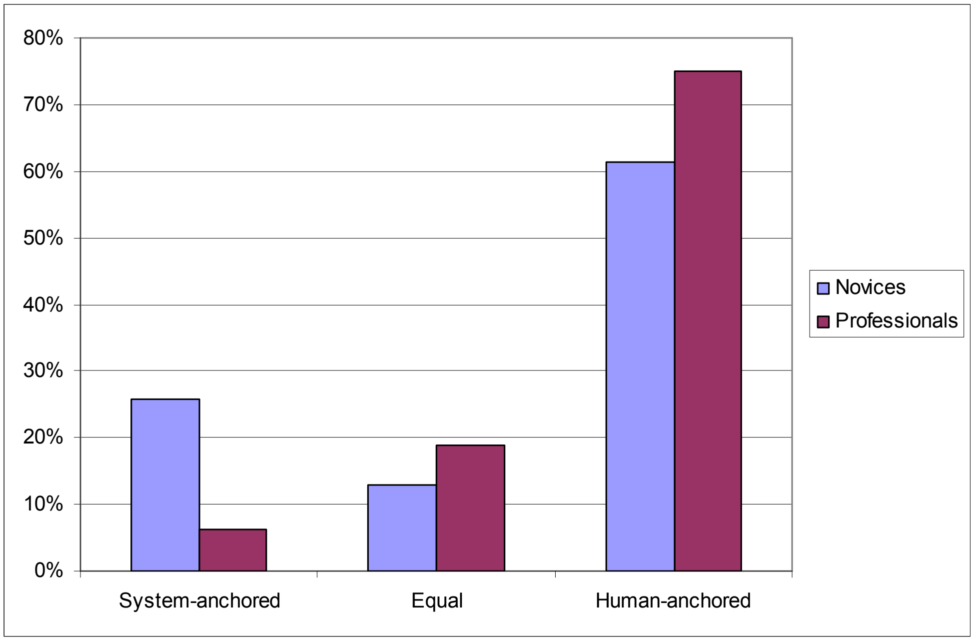


Figure 6-3 Human and system anchoring in novices and professionals

When novices' and professionals' initial judgments disagreed with the system recommendations, they had the opportunity to investigate the reasoning behind the recommendation through viewing explanations. The mean numbers of definitions, cues, and analysis explanations accessed in each condition are shown in Figure 6-4.

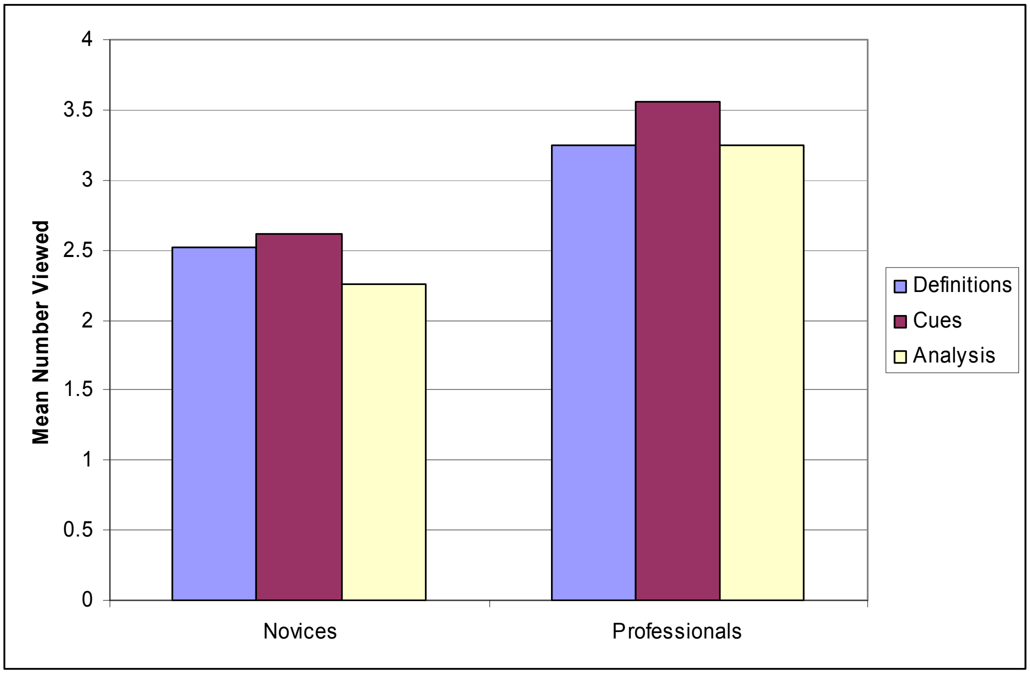


Figure 6-4 Mean BAP explanations accessed

Examination of the number of explanations accessed during the experiment was attempted via a bivariate Pearson's correlation analysis. Results from this analysis indicated that the number of cues explanations were significantly correlated with the number of analysis explanations, $r(47) = .294$, $p = .045$. As a result of possible related dependent measures, a one-way between-subjects design MANOVA was utilized to examine explanation usage.

Review of the data shows that the number of explanations viewed in each category violated assumptions of normality by demonstrating heavy skewness. To remedy this violation, a square-root transformation of the number of definitions, cues, and analysis explanations was adopted. The normality violations were remedied by the transformed variables. Box's *M* test ($p = .789$), indicated variance-covariance equality across factors.

Levene's test indicated equality between the error variances of the number of analysis explanations across groups. Therefore, pairwise comparisons were performed using the Least Significant Difference (LSD) method.

Using Wilks' λ, the dependent variate was not significantly affected by the novice/professional factor, $\lambda = .967$, $F(3, 43) = .365$, $p = .779$. This finding does not support H11.

6.6.5 Trust in the System

Novice and professional opinions about the performance and capability of the BAP were collected via a questionnaire that was administered following the conclusion of the judgment task. Included in the questionnaire were questions concerning four types of perceptions or intentions: emotional trust of the system, intentions to use the system as a delegated agent, intentions to use the system as a decision aid, and competence of the system. Measures indicating the reliability and validity of the items included in the questionnaire are presented in Appendix D. The constructs demonstrated acceptable reliability and convergent and discriminant validity with the exception of system competence. Therefore system competence was excluded from analysis.

Results from a bivariate Pearson correlation indicated that emotional trust significantly correlated with intention to use the BAP as a decision aid, $r(47) = .589$, $p < .001$. As a result of possible related dependent measures, a one-way between-subjects design MANOVA was utilized to examine users' levels of trust in the BAP. The mean factor scores for emotional trust, delegated agent intention, and decision aid intention in each the novice and professional conditions appear in Figure 6-5.

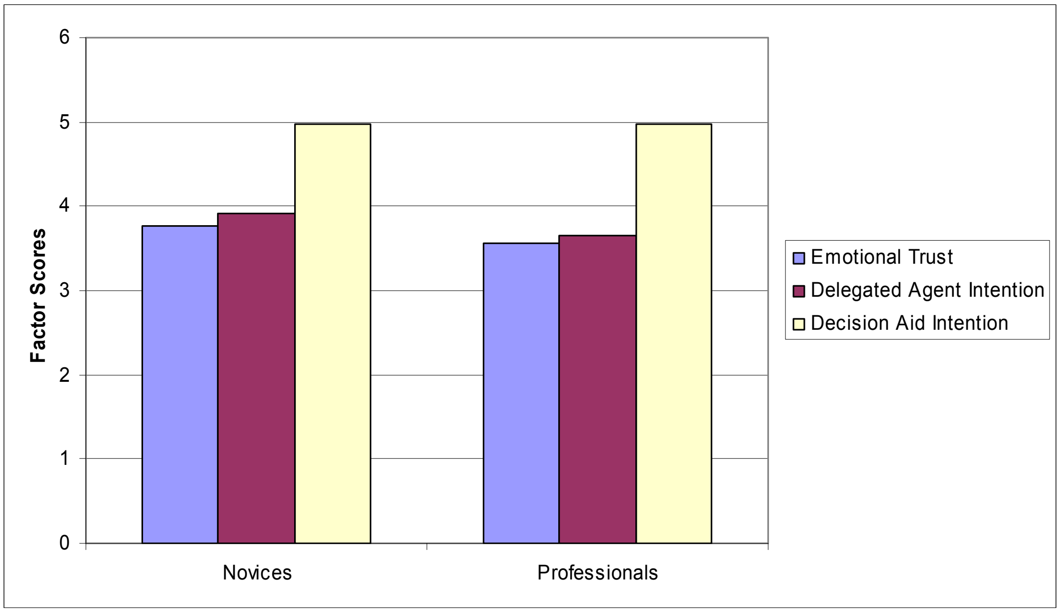


Figure 6-5 Mean factor scores for emotional trust, delegated agent intention, and decision aid intention

No violations of statistical assumptions were noted in the data and Box's *M* test (p = .937) indicated variance-covariance equality across factors. Using Wilks' λ, the dependent variate was not significantly affected by the novice/professional factor, λ = .956, $F(3, 43) = .665$, $p = .578$. This finding does not support H13.

6.7 Discussion

Caution must be used when interpreting and discussing results because the sample size of professionals is small. Thus, significant differences between novices and professionals may exist; however, the experiment may have lacked the power necessary to identify the difference. This may be particularly important in the examination of explanation usage as visual inspection indicates that there may be a systematic difference, yet none was found.

Despite this weakness, the results discussed here comprise the first attempt at understanding the possibility of combining professionals with a BAP-type system to detect deception. Therefore, the conclusions presented are exploratory in nature and could benefit from replication.

Findings concerning novice and professional judgment accuracy, confidence alignment, improvement in judgment, system, anchoring, and trust in the system are summarized in Table 6-12. A graphical summary displaying the supported, hypothesized links between independent and dependent variables appears in Figure 6-6.

Table 6-12 Summary of findings for professional computer-aided deception detection

RQ4: What is the difference in the accuracy and alignment of the initial judgments of novices and professionals	No difference
H8: Aided professionals will see greater improvement in judgment accuracy and alignment between initial and final judgments than aided novices.	Not supported
H9: Aided professionals will demonstrate higher accuracy in final judgments than aided novices.	Not supported
H10: Aided professionals will demonstrate more alignment between judgment accuracy and confidence in final judgments than aided novices.	Not supported
H11: Aided professionals will view more system explanations than aided novices.	Not supported
H12: When professionals' initial judgments conflict with the system recommendations, the professionals will anchor on their initial judgments.	Supported
H13: Aided professionals will demonstrate less emotional trust in the system, be less willing to use the system as a decision aid, and be less likely to use the system as an autonomous agent than aided novices.	Not supported

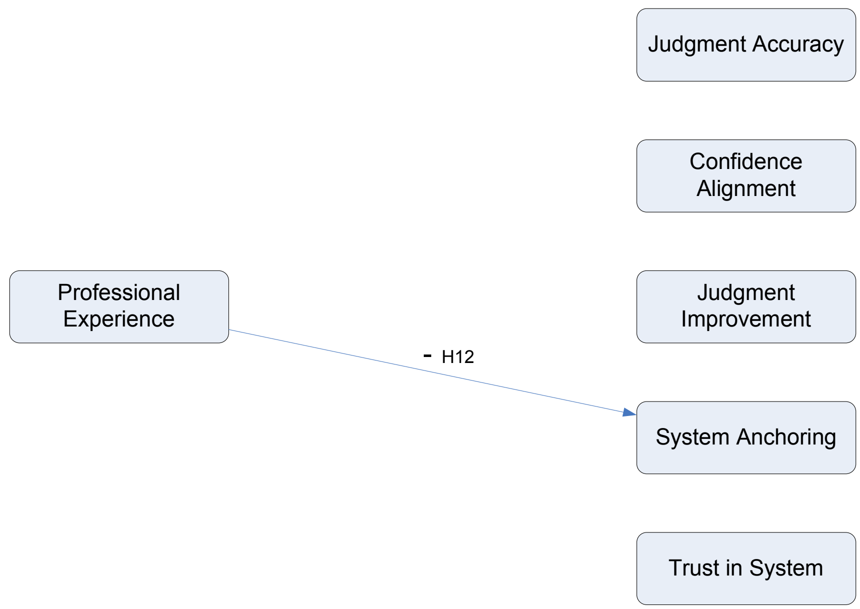


Figure 6-6 Supported hypotheses for professional computer-aided deception detection

Professional lie-catchers without access to the BAP also demonstrated the problems that many who assess credibility face: their accuracy in detecting deception is low and their confidence in their performance is unjustifiably high. The professionals demonstrated initial accuracy of 50%, which closely approximates the 54% accuracy rate suggested in other deception detection studies [9]. This finding echoes other researchers who have also surmised that professionals struggle with deception detection as much as novices do [139]. In line with prior research, this experiment concludes that novices' and professionals' initial judgments are not different from each other.

With access to the BAP, the professionals and novices both perform better in the deception detection task. The improvement in judgment accuracy and alignment between judgment accuracy and level of confidence for both groups was significant. It was believed that the improvement for the professionals would be more pronounced;

however, the results did not support this hypothesis. Rather, there was weak evidence that novices benefited more from access to the BAP than professionals did. However, in other tests, the difference in improvement between professionals and novices turned out to be minor. Similarly, the accuracy and confidence alignment of professionals' final judgments were no different than novices' final judgments. Thus, the BAP improved novice and professional deception detection equally.

As with the novice-only experiment, it is interesting to note that the BAP alone classified 80% of the interviews correctly. Yet novices using the BAP correctly classified 61.0% of the interviews correctly and the professionals correctly classified 61.9% of the interviews. This finding emphasizes the discussion in section 5.7 and demonstrates that professionals, in addition to novices, struggle to detect deception accurately.

In discussions during the debriefing, a common sentiment among the professionals was that the interviewer did not probe or further question the interviewee. As mentioned in section 5.2.1, the same questions were directed at all interviewees. While this consistency benefited comparability across interviews, the professionals considered it a significant departure from the real world where details surrounding an event in question are repeatedly reviewed in search of any inconsistencies or significant omissions[15]. They believed that if probing were introduced, their ability to detect deception might have been enhanced.

---

[15] Some professionals stated that these types of interviews are held for multiple hours. Such interviews are significantly longer than the 2-3 minute interactions used in the experiment.

The professionals also lamented the fact that they could not directly question the interviewee and supposed that they could judge more accurately had they been able to do the questioning. While this may seem to be a limitation on the experiment, recent research has suggested that the ability to ask questions may not be such a benefit [54]. When accuracy rates of judgments from police officers who conducted their own interrogations were compared to judgments of police officers who had seen a video recording of an interrogation, there was no difference between the two. Further, the accuracy rates of both groups were not significantly better than what would be expected by chance alone [54].

The poor accuracy rates of the professionals and novices with access to the BAP further draws into question the notion of a human-computer system for detecting deception (as illustrated by Figure 3-7). Their performance seems to suggest that excluding the human user may improve judgment accuracy and confidence alignment. This finding also seems to question the notion that better deception detection can be achieved through a combination of improved computer-based tools and improved human efforts (see section 1.1). Improved human efforts, evidenced by training or professional experience, do not seem to improve detection ability noticeably. However, as in the discussion of the novice-only experiment, complete exclusion of the human user is not recommended. Rather, a re-examination of the roles of the human user is recommended.

One of the critical concepts forwarded by this dissertation is that humans possess unique abilities, which are not replicable by a computer, which may be useful in identifying deception. In the experiments described in this dissertation, training or experience did not bring these abilities to light. However, deception detection via humans identifying indirect cues to deception has shown promise [140]. Responses to questions such as “Does this person have to think hard when responding?” or “Is this person engaged in the conversation?” may be more diagnostic than a classification question such as “Is this person being deceitful?” [140]. Perceptual measures such as verbal and vocal immediacy have also been shown to be among the most diagnostic of all observable deceptive cues [26]. Such indirect cues to deception could then be considered in light of BAP recommendations to arrive at a final judgment. This approach of combining human- and computer-based detection methods has yet to be tested empirically.

Human users also assume a critical role in the interrogation or interviewing of any individual. This important role was alluded to during the discussion of the novice-only experiment. The BAP simply reports on the behavior that it observes; it does not provide any drill-down capabilities that permit the user to identify the point at which deception occurred. This responsibility falls squarely on the human user. The interviews used in the novice-only and novice/professional experiments contained questions that elicited outright falsification and forced the liars to elaborate on the deception, thus increasing the level of total deception in the interaction. Only through proper interrogation techniques can the BAP fully exposed and subsequently identify deception.

Finally, the human user can effectively take into consideration the context of the interrogation to determine the applicability and relevance of the BAP’s recommendations. Fitts [42] suggests that humans are uniquely proficient at recalling facts at appropriate times. In the context of deception detection this ability may include reconciling evidence or other information with the statements of an interviewee in light of the BAP's recommendations. Possible roles of a potential BAP user appear in Table 6-13.

Table 6-13 Possible responsibilities of a BAP user

Identify indirect cues to deception and incorporate indirect cues into judgment
Conduct interrogation to uncover deception
Reconcile relevant evidence and other information with BAP
Limit use of the BAP to appropriate segments of the interactions

Examination of system-anchoring among professional users produced findings that were in line with what was hypothesized. The results indicated that when professionals' initial judgment conflicted with the BAP's recommendation, the professionals overwhelmingly retained their own judgment. This finding mirrors the finding in the novice-only experiment and suggests that both professionals and novices trusted their own judgment over the judgment of the system. This finding was bolstered by the low levels of emotional trust in the BAP and a general unwillingness to use the BAP as a delegated agent.

From discussions with the professionals, it became clear that some professionals were treating the BAP like a new partner about whom they knew very little. Before they placed any confidence in the BAP, they wanted to see its capabilities and know if their trust would be justified. The novice-only and novice/professional experiments did not involve any immediate feedback as to the performance of the BAP or the human user. Therefore reputation built on known past performance was not possible. Immediate feedback concerning judgment accuracy was not provided in the novice-only and novice/professional experiments because deception is rarely uncovered immediately following an interaction. The lack of feedback was used because it is representative of real interactions. To explore reputation effects, a replication of the novice/professional experiment with immediate judgment feedback could be conducted.

Interestingly, both the novices and professionals responded favorably to questions regarding intentions to use the BAP as a decision aid. As evidence of this finding, some professionals voiced support of using a tool like the BAP because it encouraged them to consider the reasoning behind their own judgments. Although they did not always agree with the BAP, they stopped to think about their judgments and they said they adjusted confidence levels accordingly.

CHAPTER 7

CONCLUSIONS AND FUTURE STEPS

Research has shown repeatedly that humans struggle with evaluating the credibility of the information they receive from others. Numerous studies note that people typically identify deception with accuracy only slightly better than chance (approximately 54%) [9, 69, 139]. Coupled with poor ability in deception detection is poor calibration of confidence in deception judgments. There appears to be a near zero correlation between the accuracy of peoples' judgments and their confidence in their assessments [9, 25]. The combination of poor accuracy in deception detection and poor alignment between deception detection ability and confidence in deception judgments represents a significant problem for business and governmental organizations that rely on credible information to operate effectively.

Computer-based decision aids have been suggested as a method of remedying the problems of low detection accuracy and poorly calibrated confidence in judgments. The purpose of this dissertation is to examine the effectiveness of joining improved human efforts with the capabilities of an ES prototype to improve decision outcomes in deception judgments.

In addition to decision outcomes, decision strategies that ES users utilize in formulating a final judgment of deception were also explored. This work attempts to partially uncover general strategies that trained and untrained novices and professionals use in incorporating ES output in deception decisions. Such decision strategies are bounded on one hand by a complete abdication of decision making in favor of the ES output and on the other hand by complete disuse in favor of human-only decision making.

Finally, this study addresses perceptions of the decision maker utilizing the ES. Central in this investigation are results collected from the users concerning trust in the system, willingness to use the system as a decision aid, and willingness to allow the system to operate without supervision.

7.1 Contributions

Significant contributions concerning decision outcomes, decision strategies, and decision maker perceptions of computer-based deception detection tools are made to the information systems and deception detection fields. They include:

- Summarization and survey of literature concerning computer-aided decision making and deception detection. These bodies of literature are then synthesized and combined to create theoretically-supported, testable hypotheses concerning computer-aided deception detection.
- Summarization of literature concerning unobtrusive computer-aided deception detection methods. Rationale behind kinesic and linguistic analyses is presented and contrasted with more intrusive methods of deception detection (e.g., the polygraph). Kinesic and linguistic analyses are shown to be useful in a natural environment as they do not require the body sensors and can be passively performed.

- A semi-functional expert system prototype that implements kinesic and linguistic analysis. The BAP uses low- and higher-level cues in accordance with kinesic and linguistic analyses and presents recommendations concerning the level of deception it observed in an interaction. The prototype provides explanations concerning the cues it examines and its analysis. Definitions of relevant functions are also provided.

7.1.1 Decision Outcomes

- The BAP significantly improves the accuracy rate of novice and professional lie-catchers
- The BAP significantly improves the level of alignment between judgment accuracy and confidence in the judgment
- Training is not shown to improve judgment accuracy or confidence alignment
- Professionals' accuracy and confidence alignment are not better than novices' accuracy and confidence levels in unaided and aided deception detection
- Combined human-BAP efforts do not exceed the accuracy rate of the BAP alone

7.1.2 Decision Strategies

- Both professionals and novices demonstrate strong human-anchoring when initial judgments and BAP recommendations contradict each other
- Professionals and novices use the same amount of number and types of explanations

- Training decreases the number of explanations novices used

7.1.3 Decision Maker Perceptions

- Novice and professional users are unwilling to use the BAP as a delegated agent, but are willing to use the BAP as a decision aid
- Level of emotional trust in the BAP, intention to use the BAP as a delegated agent, and intention to use the BAP as a decision aid are unchanged by accuracy of the BAP or by training

7.2 Limitations of Studies

The findings of this dissertation must be considered in light of significant limitations. First and foremost, the exploration of the BAP took place in a laboratory within the confines of an experiment. Further, the stimulus materials were also generated via an experimental study. While an experimental approach offered the experimenter a fair amount of control, a good deal of realism was lost. This may have affected the motivations of the novices; but it certainly influenced the performance of the professionals.

Further, the BAP used in the experiments described in Chapter 5 and Chapter 6 was not fully operational and therefore was constrained in its ability to respond to the requests and needs of the user. Although every effort was made to facilitate a realistic interaction between the user and BAP, this lack of full functionality may have influenced user perceptions of the BAP and subsequently affected BAP utilization.

As discussed in section 5.7 the training that novices received was also limited. The training took place over a limited length of time and may not have been fully absorbed. The length of the training may have impacted the training effectiveness in the detection task.

Finally, there were a limited number of individuals in the sample of professional lie-catchers that underpowered statistical tests (power = .478 for a medium effect size) [39]. The small sample size may have hidden important differences in performance and perceptions between professionals and novices.

7.3 Implications of Research

There are many important implications of this research for deception detection and computer-aided decision making. First, this research suggests that it is possible to create technology that can improve on human-only deception detection. Even when the BAP was operating at a lower accuracy, BAP use improved judgment accuracy and confidence alignment. This finding suggests that even incremental improvements in technology to detect deception can have a significant positive impact on deception judgment performance.

This research suggests that the capabilities of the human user in a deception detection task may be limited. Thus, a re-examination of the role of the human user is suggested. This newly-suggested role covers identification of indirect cues to deception, conducting the interview, considering existing evidence, and ensuring the proper operation of the system. These suggested responsibilities reduce the amount of overlap between the system and human user and position the system as the primary tool for detection of signs of deceit.

Results of this work also elevate the importance of continued research in developing technology to monitor behavior for possible signs of deceit. Such technology may be extremely useful in alerting human users to suspicious behavior it observes. The human user may then investigate by examining evidence to resolve the suspicion.

Finally, this study casts doubt on the notion of humans and computers filling complementary roles in a difficult decision-making task. Neither the system nor the novice and professional users demonstrated a high level of skill in detecting deception. When the users were combined with the system, a synergistic compounding of human and computer capabilities did not occur. Rather, the resultant performance was between the human and computer performance. A reasonable extension of this finding may be that when human users are combined with computer-based tools in a difficult decision-making task, the resultant performance of the human-computer system will rarely exceed the higher of the individual performances (i.e., human alone or computer alone).

7.4 Future Steps

In suggesting possible future steps that may be pursued in response to this research, a research framework suggested by Berthon and colleagues [7] was utilized. Berthon et al. suggest that research space can be characterized by three axes: theory, methodology, and context. Research without extensions in any axis is termed replication. Extensions to this research are presented touching on each of these axes.

7.4.1 Replication

The findings of this research are such that replication is critical. Traditionally, researchers seek to discover differences in characteristics and qualities between populations. In contrast, many of the important contributions of this research note the lack of significant difference (e.g., between professional and novice performances). The nature of these findings requires replication. Additionally, replications involving larger sample sizes may alleviate weaknesses that were listed in section 7.2 due to sample size.

7.4.2 Extensions in Methodology

Numerous extensions in methodology are conceivable. First, one may consider providing feedback during the detection task. Both the novices and professionals demonstrated the tendency to anchor on their own beliefs, possibly under-utilizing the BAP. The provision of feedback during the task may inform the users that their reliance on their own beliefs to the exclusion of the BAP may be counterproductive.

Second, improved training may be more influential on judgment accuracy and confidence alignment. Training could be longer, include more interaction (e.g., a question and answer session), and be spread over multiple days. Such improvements would potentially help lie-catchers to retain more of the information that would be useful during the detection task.

Third, different styles of judgment formation could be tested. Such styles would examine the performance of the system and human user along different levels of the Sheridan et al scale of automation (see Figure 3-4). For example, judgment performance may be improved by the human user providing the system with information concerning indirect cues to deception so the system can determine the final judgment.

Finally, a fully functional system could be used in place of the experimental prototype. The fully functional prototype would include more detailed and a greater quantity of explanations. The system would be more stringently trained so the performance of the system would be generalizable to other similar situations. The additional functions of the system may improve user perceptions and increase use of the system. Further, with a fully functioning system, future experiment participants would be able to conduct their own interrogations while using the system in real time. Such an experiment would represent a genuine test of the feasibility of a system similar to the BAP for use in a field setting.

7.4.3 Extensions in Context

With a fully functioning system and after realistic experimentation, the system may be field-tested and tailored to higher-stakes situations. Possible scenarios for such field testing are wide and varied. They range from possible use in criminal interrogations to employment related job interviews. Once the research in computer-aided deception detection has advanced to this stage, careful consideration should be given to privacy and the rights of the suspect or interviewee. This topic has yet to be fully addressed. (However, see [107] for an initial attempt at examining legal ramifications of kinesic analysis).

An additional component of the context that may be changed is the population which is sampled. The sample that was drawn to represent professional lie-catchers can and should be expanded to include other law enforcement personnel (e.g., Federal Bereau of Investigation, Border Patrol Agents, etc.). Additional professional lie-catchers such as certified fraud examiners, lawyers, members of the United States Department of State, and the members of the intelligence community may also be sampled.

Finally, the capabilities of the system may be expanded to cover more deceptive cues. Other methods of deception detection such as vocal analysis and thermal analysis may be incorporated into the system to identify additional deceptive cues. Incorporating vocal analysis and thermal analysis into a BAP-like system would potentially offer greater discrimination between truth and deception and the system would still retain unobtrusive properties. The capabilities of the new system could then be analyzed and compared to more mature technologies such as the polygraph to determine accuracy rates.

7.4.4 Extensions in Theory

Future work may have broader implication for human-computer systems in uncertain circumstances. Computer-based tools are most commonly used in situations where tasks are well defined and can be reliably performed by humans. Future research may shed light on other human-computer systems which must perform under conditions of poor accuracy and reliability. Such situations are likely to arise as more tools and computer agents are constructed to assist in ill-defined tasks such as prediction in social systems, interpretation of human behavior and human-agent teamwork.

7.5 Concluding Remarks

Automatic interpretation of human behavior is a very complex problem. This is particularly true of computer-aided deception detection. This research reiterates previous findings that humans (both novice and professional lie-catchers) struggle with deception detection. Computer-based deception detection tools may improve detection ability. However, more work and research is necessary in both tool development and understanding the interaction between the human user and the computer aid before such applications will be suitable for full implementation and use in consequential deception detection. This dissertation represents initial efforts in this regard.

Appendix A – Sample BAP Results and Explanations

Higher-accuracy BAP Condition

The following screen captures show the BAP recommendations and explanations in the higher-accuracy condition. The sample screen captures were all taken after analyzing a single deceptive interview. Some of the explanation screen captures omit the lower portion of the window and the button that closed the explanations window. However, no text was omitted in the figures.

In order to make the experimental participants believe that the BAP was a fully functioning system, the word "prototype" was excluded from the name of the system. Therefore, during the experiments, the expert deception detection system was called the Behavioral Analysis System (BAS) rather than the Behavioral Analysis Prototype (BAP).

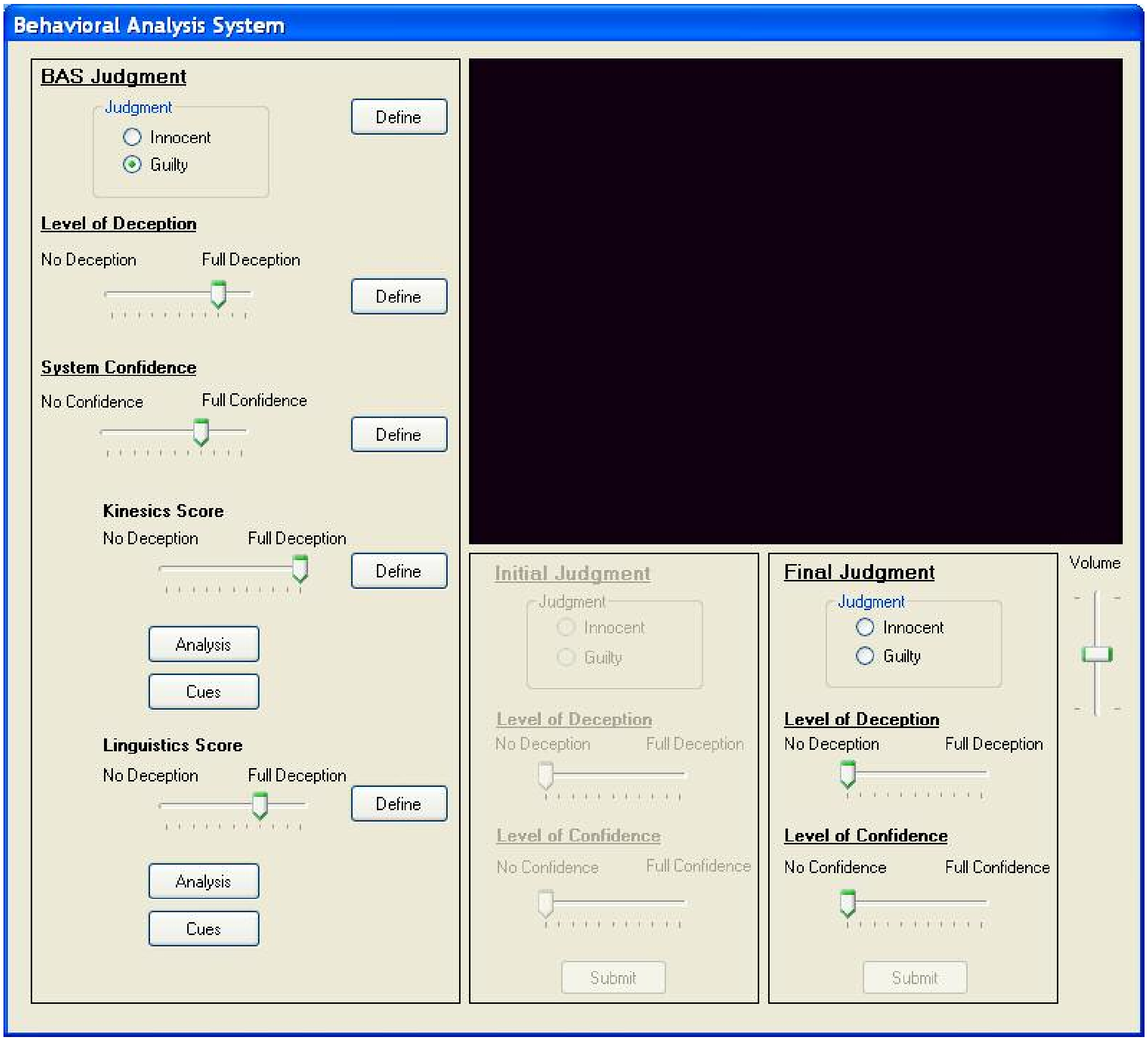


Figure A-1 Sample higher-accuracy BAP output indicating a guilty interviewee

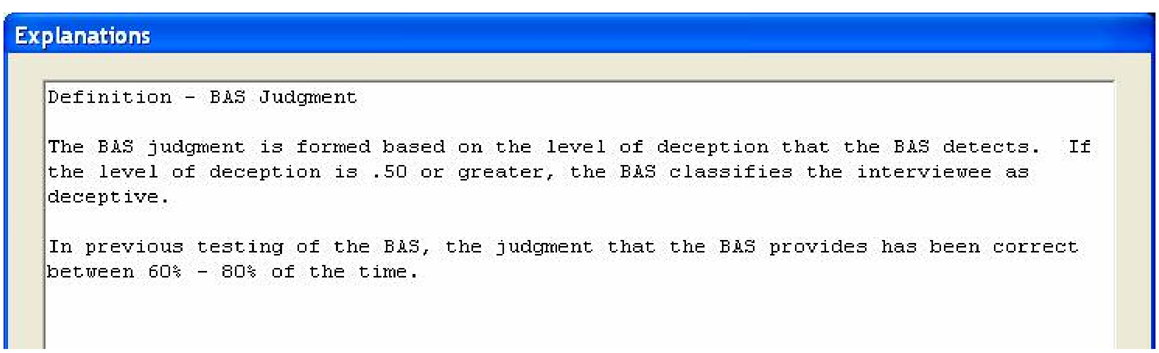


Figure A-2 Explanation of the higher-accuracy BAP's judgment

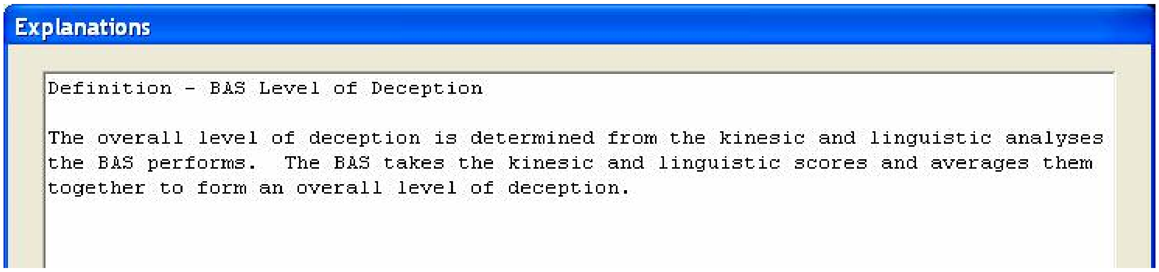


Figure A-3 Explanation of the higher-accuracy BAP's level of deception

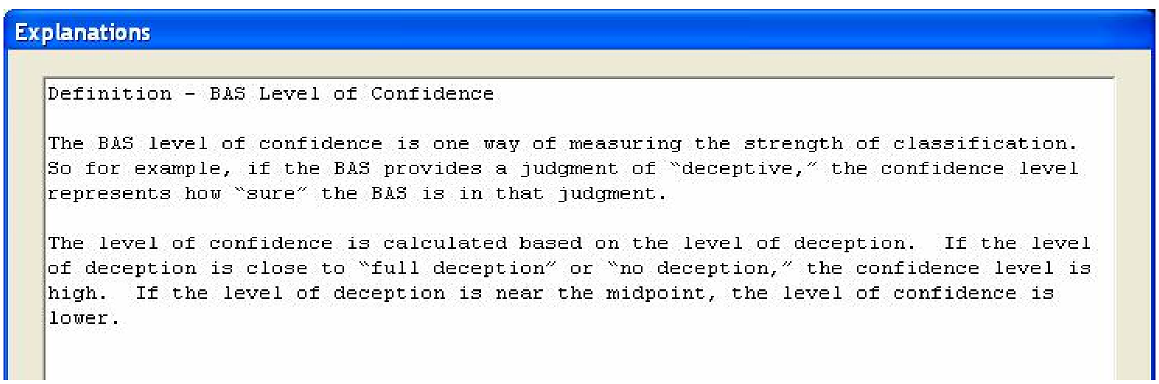


Figure A-4 Explanation of the higher-accuracy BAP's level of confidence

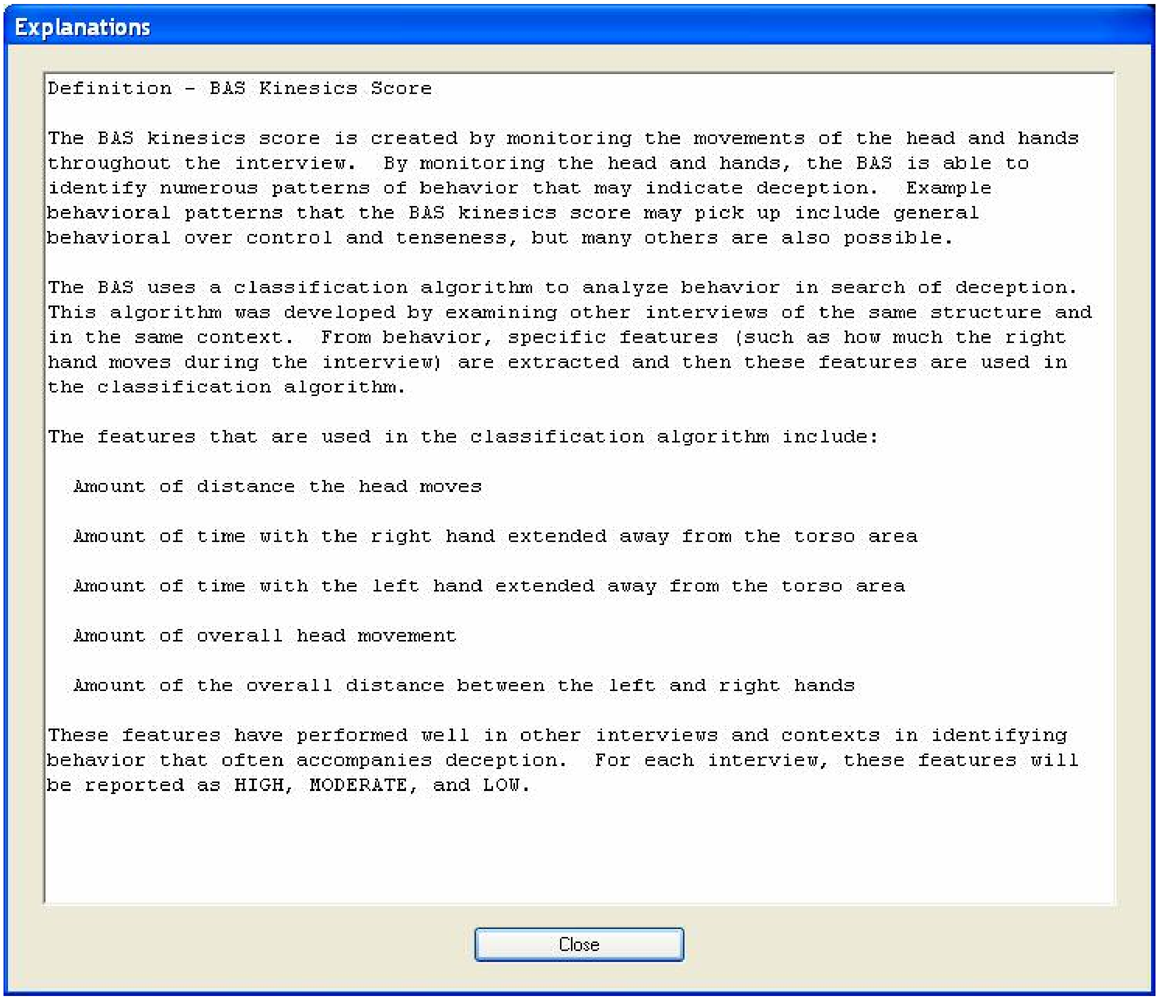


Figure A-5 Explanation of the higher-accuracy BAP's kinesics score

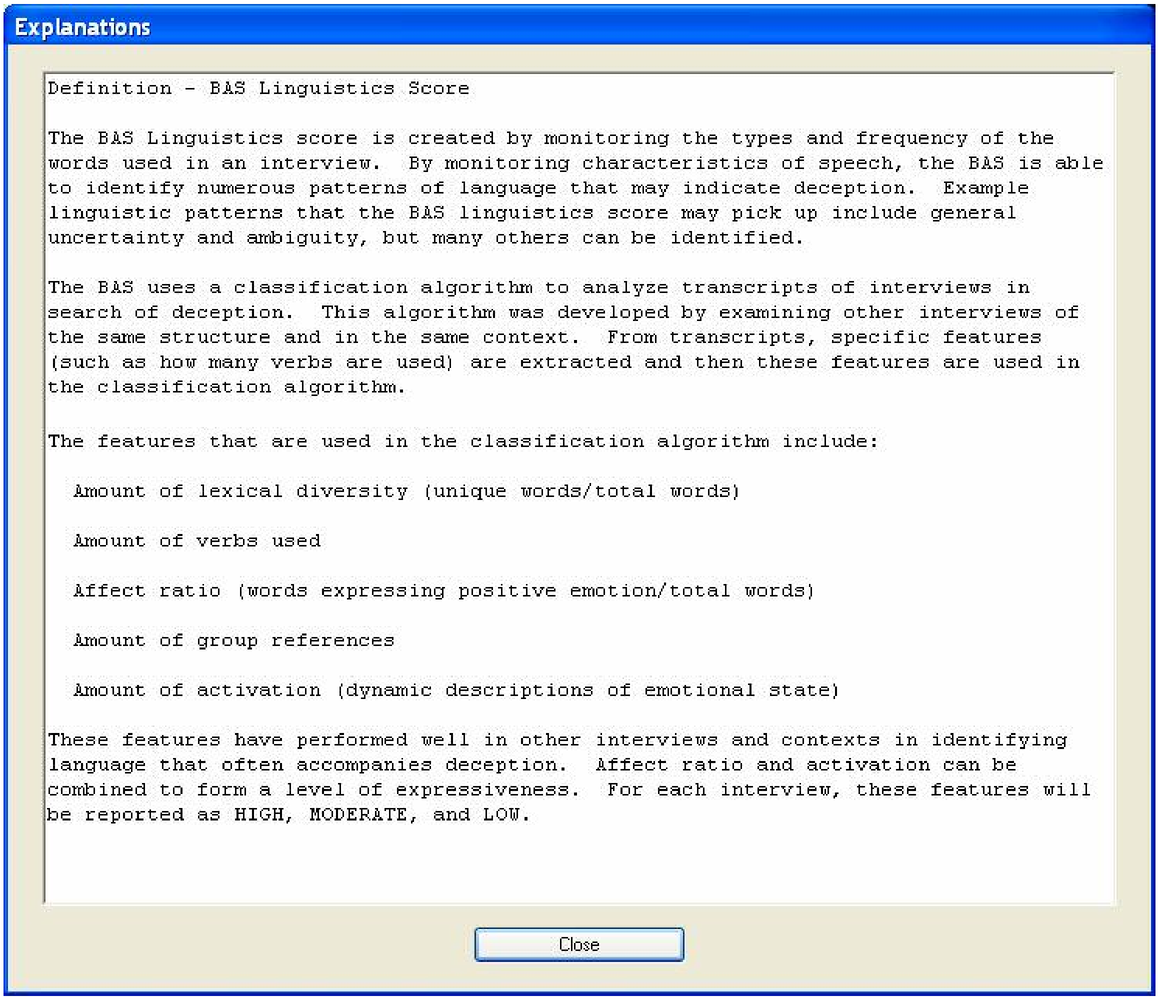


Figure A-6 Explanation of the higher-accuracy BAP's linguistics score

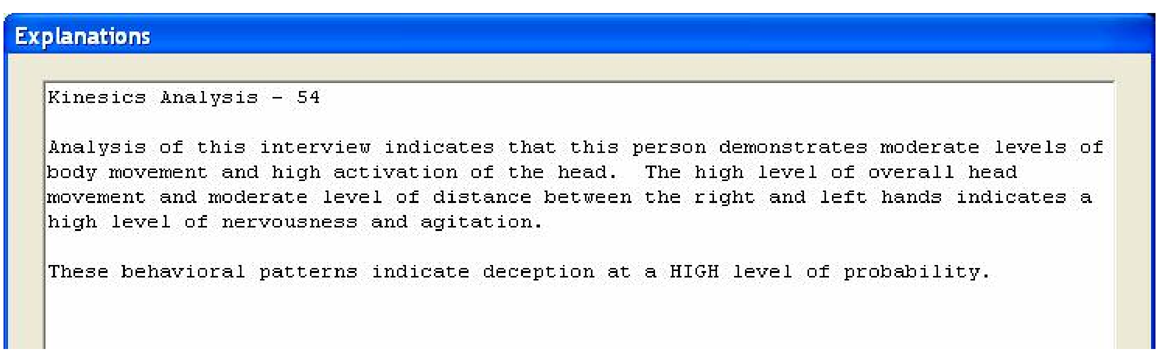


Figure A-7 Explanation of the higher-accuracy BAP's kinesics analysis results

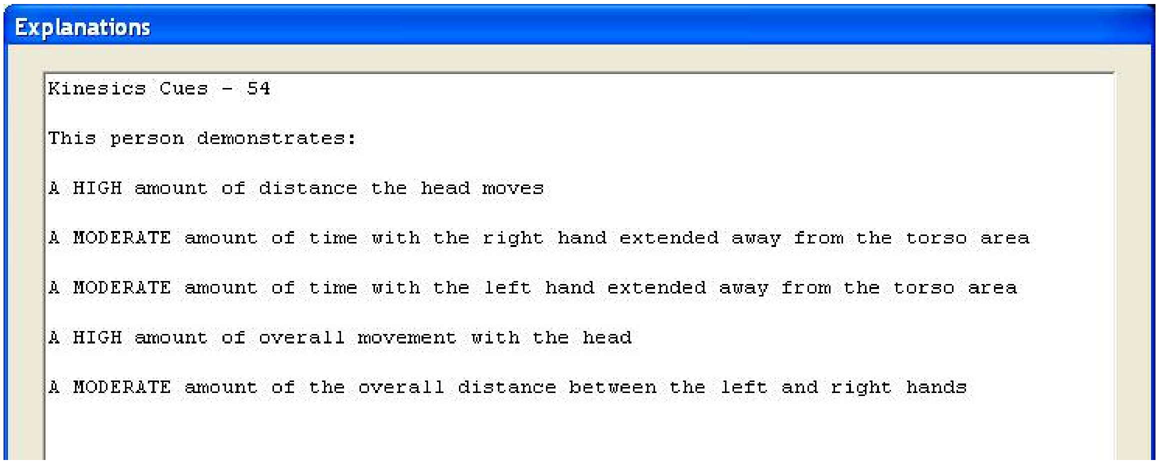


Figure A-8 Explanation of the higher-accuracy BAP's kinesics cues results

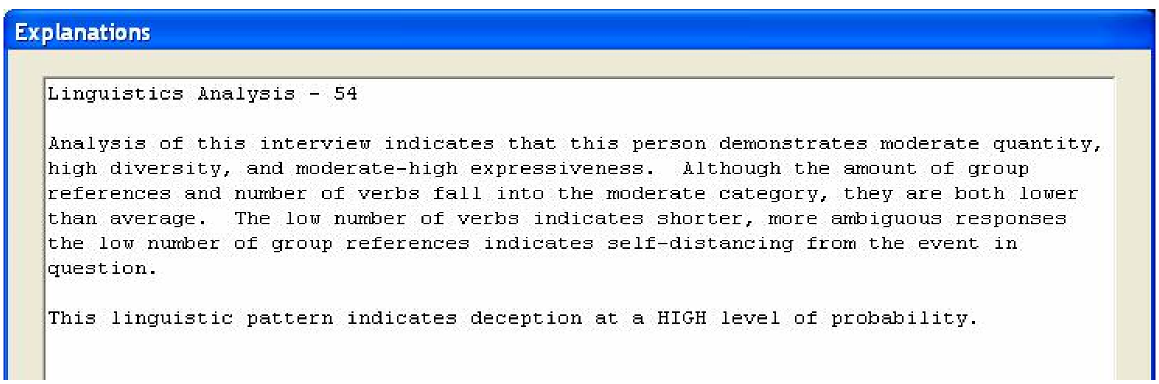


Figure A-9 Explanation of the higher-accuracy BAP's linguistics analysis results

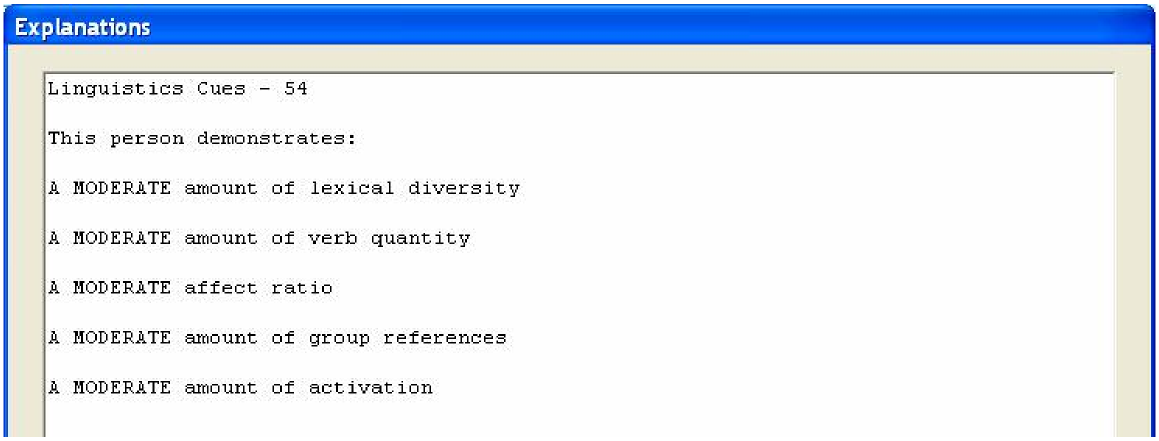


Figure A-10 Explanation of the higher-accuracy BAP's linguistics cues results

Lower-accuracy BAP Condition

The following screen captures show the BAP recommendations and explanations in the lower-accuracy condition. The sample screen captures were all taken after analysis of a single truthful interview. Some of the explanation screen captures omit the lower portion of the window and the button that closed the explanations window. However, no text was omitted in the figures.

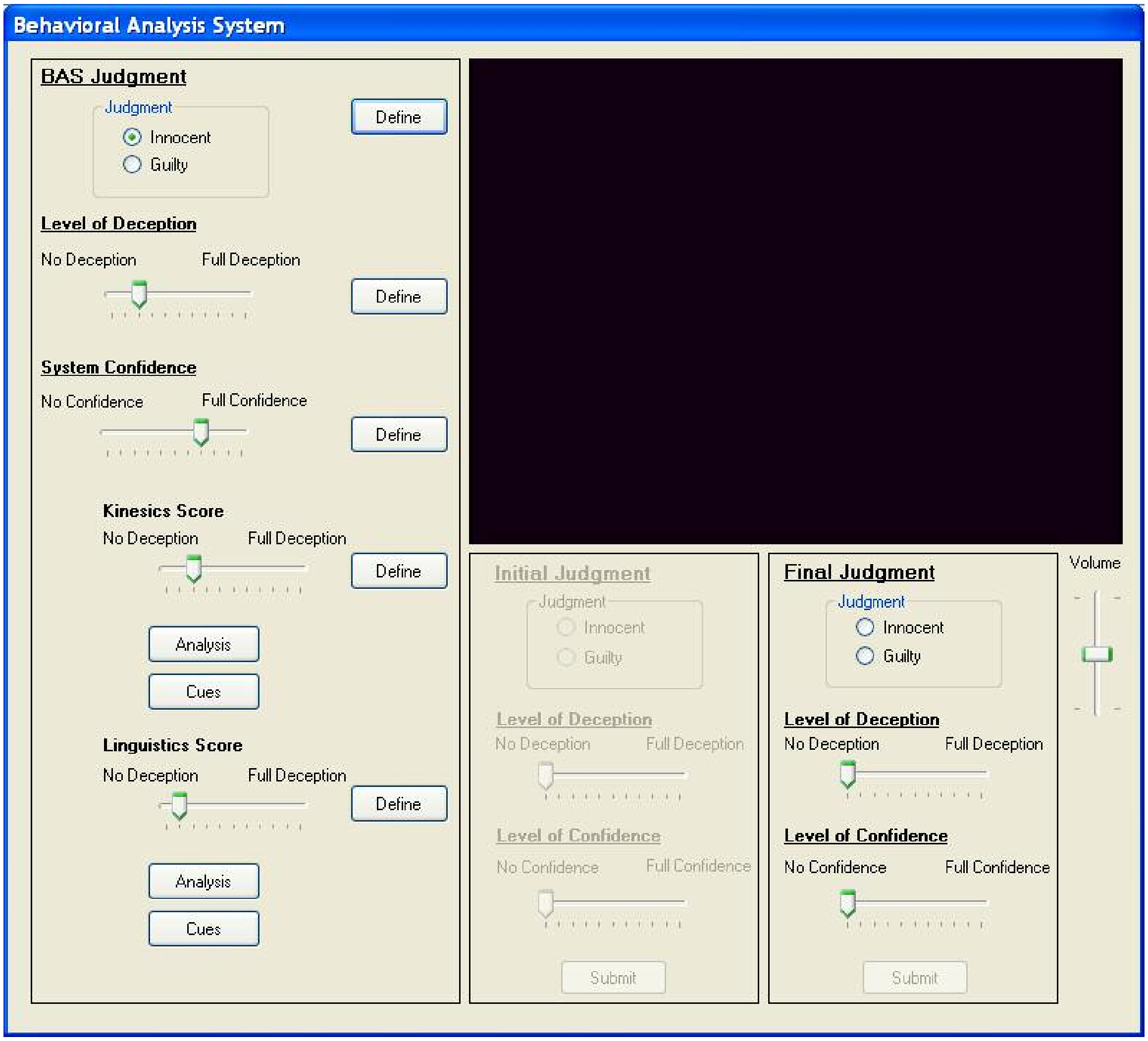


Figure A-11 Sample lower-accuracy BAP output indicating an innocent interviewee

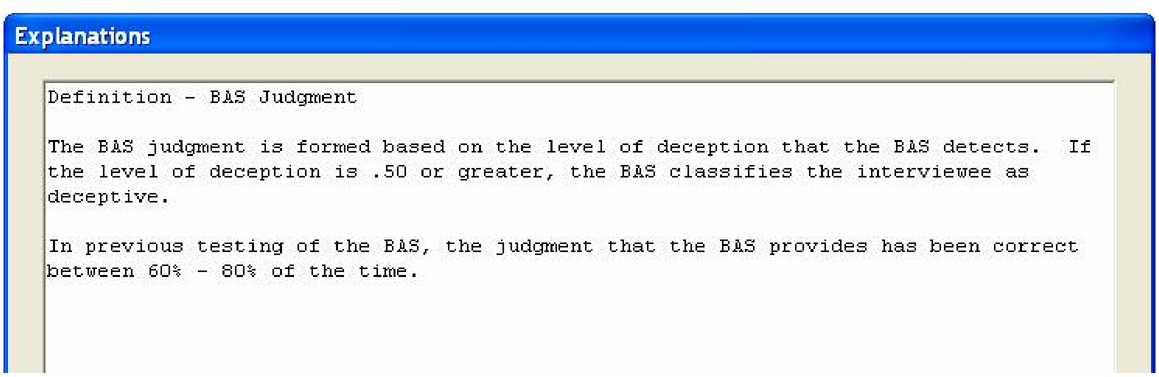


Figure A-12 Explanation of the lower-accuracy BAP's judgment

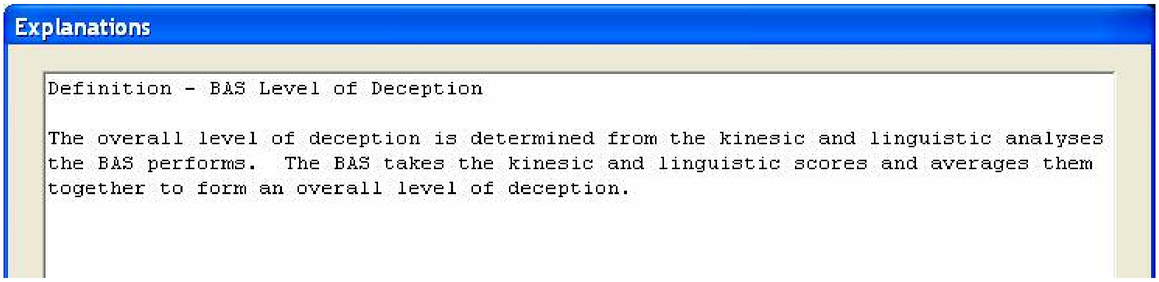


Figure A-13 Explanation of the lower-accuracy BAP's level of deception

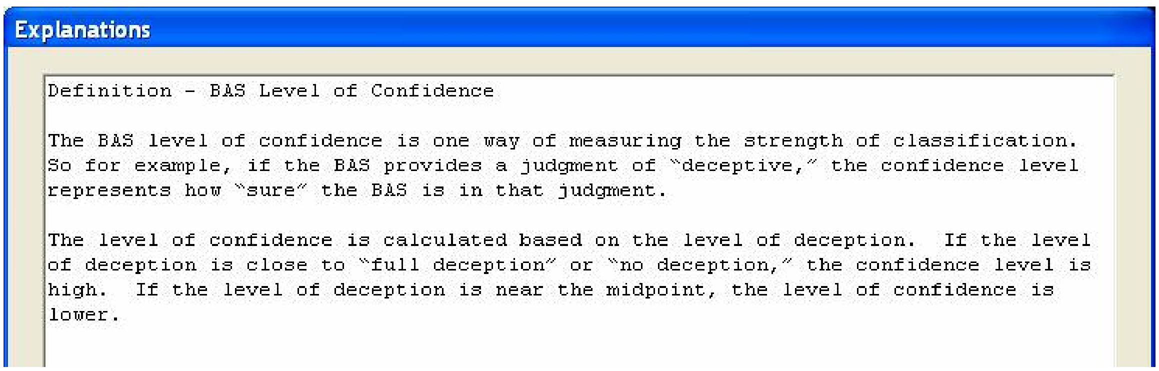
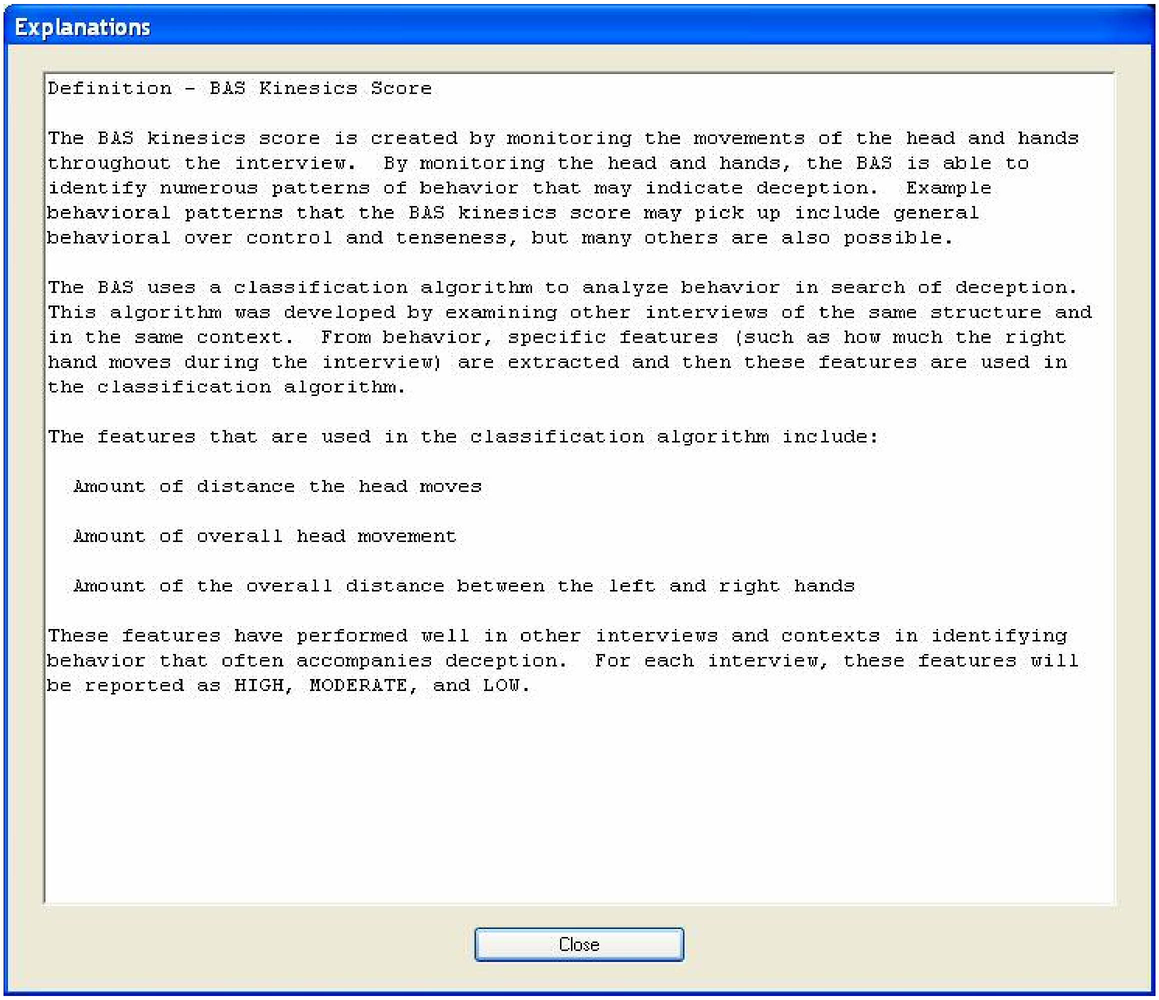


Figure A-14 Explanation of the lower-accuracy BAP's level of confidence


Explanations

Definition - BAS Kinesics Score

The BAS kinesics score is created by monitoring the movements of the head and hands throughout the interview. By monitoring the head and hands, the BAS is able to identify numerous patterns of behavior that may indicate deception. Example behavioral patterns that the BAS kinesics score may pick up include general behavioral over control and tenseness, but many others are also possible.

The BAS uses a classification algorithm to analyze behavior in search of deception. This algorithm was developed by examining other interviews of the same structure and in the same context. From behavior, specific features (such as how much the right hand moves during the interview) are extracted and then these features are used in the classification algorithm.

The features that are used in the classification algorithm include:

Amount of distance the head moves

Amount of overall head movement

Amount of the overall distance between the left and right hands

These features have performed well in other interviews and contexts in identifying behavior that often accompanies deception. For each interview, these features will be reported as HIGH, MODERATE, and LOW.

Close

Figure A-15 Explanation of the lower-accuracy BAP’s kinesics score

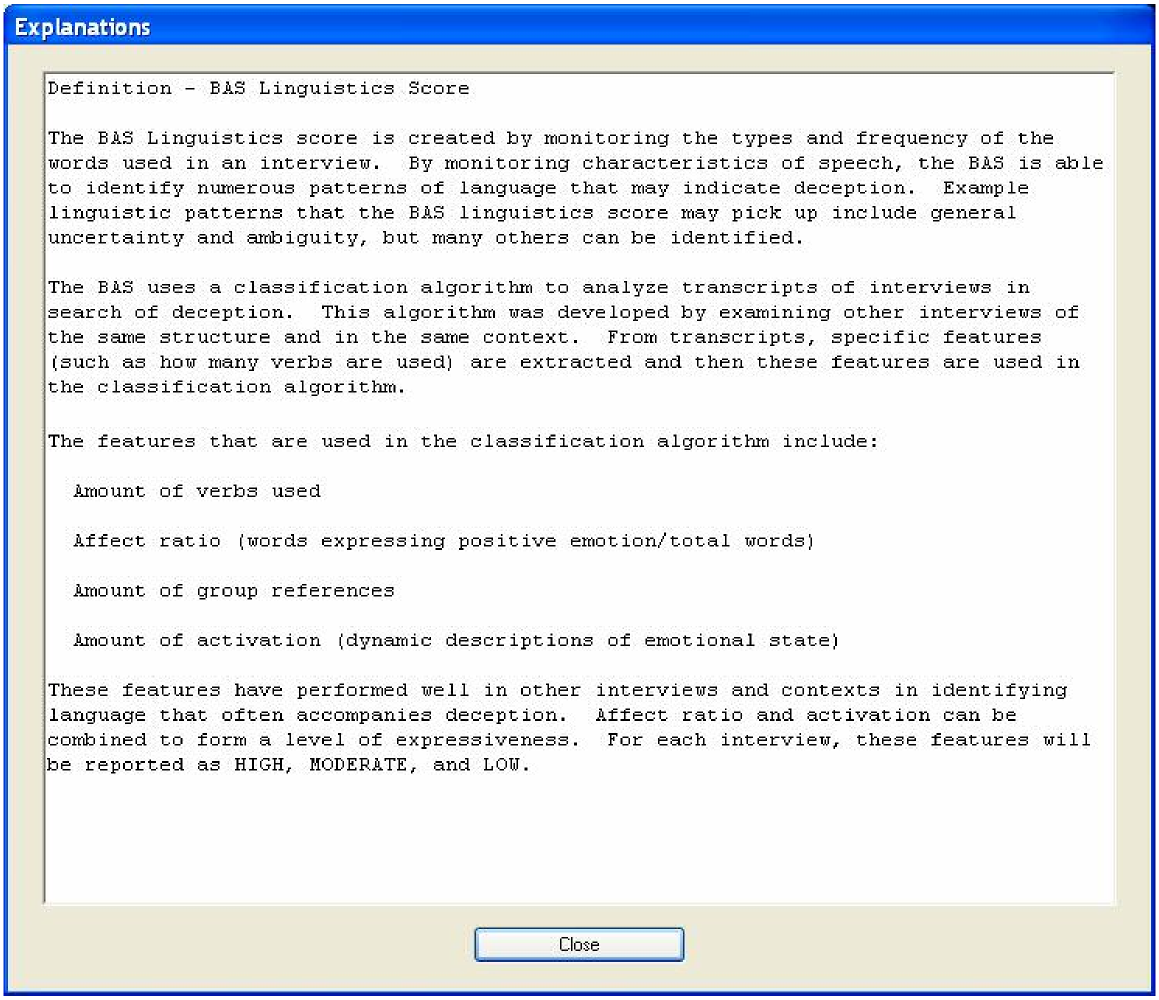


Figure A-16 Explanation of the lower-accuracy BAP's linguistics score

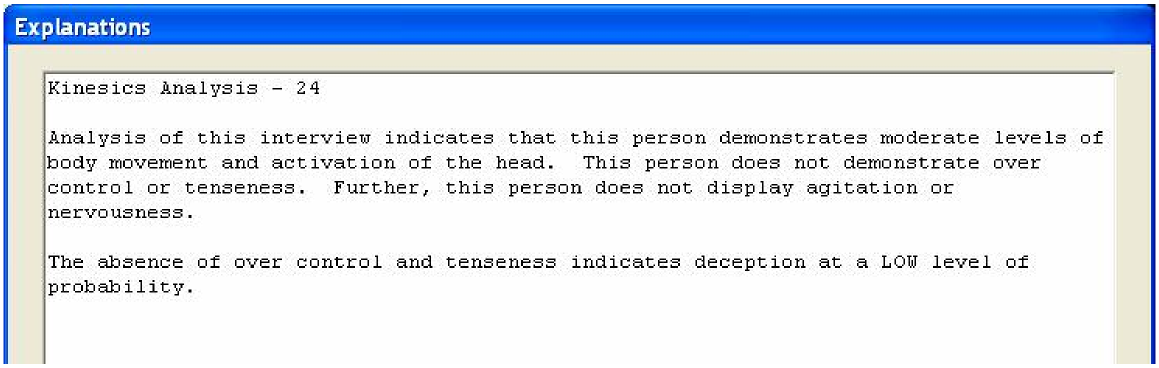


Figure A-17 Explanation of the lower-accuracy BAP’s kinesics analysis results

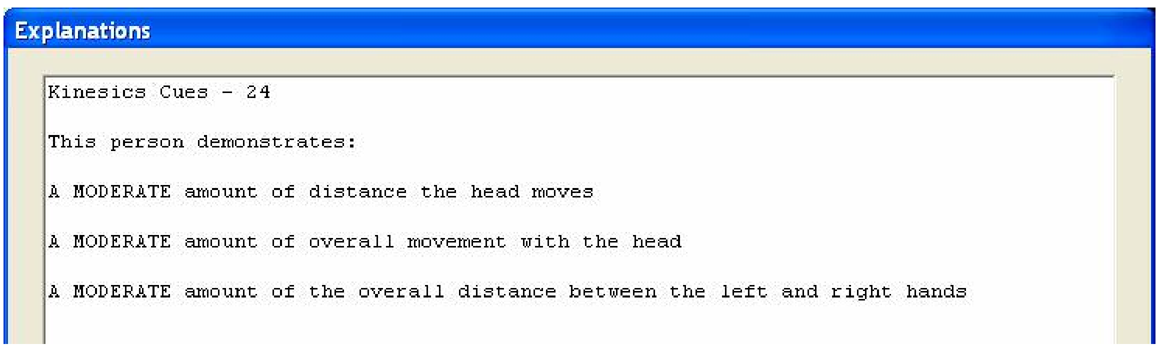


Figure A-18 Explanation of the lower-accuracy BAP’s kinesics cues results

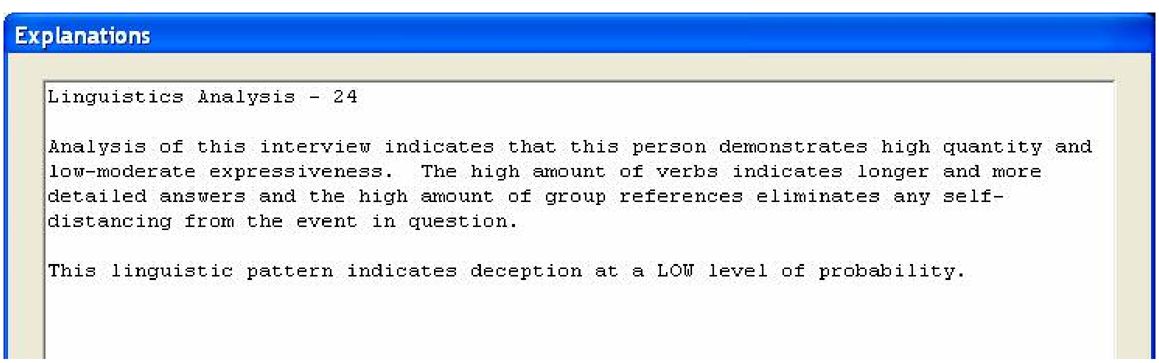


Figure A-19 Explanation of the lower-accuracy BAP’s linguistics analysis results

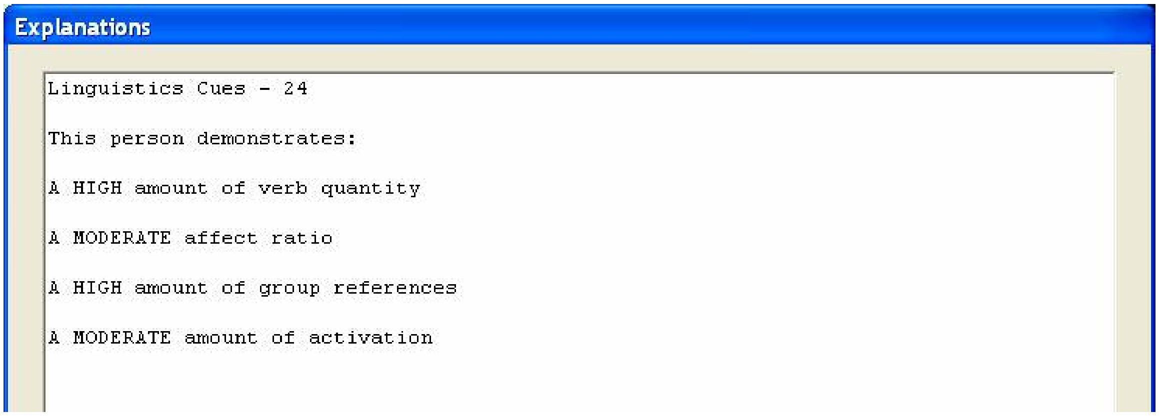


Figure A-20 Explanation of the lower-accuracy BAP’s linguistics cues results

APPENDIX B – TRAINING MATERIALS OUTLINE

The following is an outline of the training materials that were presented to novices and professionals. The training materials were adapted from a training course that was originally produced by Christopher Diller and Karl Wiers of the Center for the Management of information [31].

I. Agenda
    a. What is deception
    b. Human detection ability
    c. Categories of Indicators
        i. Arousal
        ii. Emotion
        iii. Cognitive effort
        iv. Memory
        v. Communicator tactics
    d. Concluding Remarks
II. Deception Defined
    a. Deception is "the intent to deceive a target by controlling information to alter the target's beliefs or understanding in a way that the deceiver knows is false" (Buller and Burgoon 1994)
    b. Deception is a message knowingly transmitted with the intent to foster false beliefs or conclusions
III. Human Detection
    a. Humans are poor lie-detectors
    b. ~54% (including truth and deception)
    c. Approximately the same as flipping a coin
    d. Detection accuracy rates may improve with:
        i. Training
        ii. High stakes scenarios
        iii. Detection in criminal interviews approaches 80% accuracy
IV. Problems for Humans to Overcome
    a. Truth bias
        i. Tendency to assume all tell truth
        ii. Common among lay-people
    b. Othello (lie) bias
        i. Tendency to assume all are lying
        ii. Common among law enforcement personnel
    c. Excessive focus on non-discriminant cues

            i. Gaze aversion
V. Five Categories of Cues
    a. Arousal
    b. Emotion
    c. Cognitive Effort
    d. Memory
    e. Communicator Tactics
    f. In general, deceivers:
        i. Show more arousal
        ii. Show unpleasant emotion
        iii. Show higher cognitive effort
        iv. Demonstrate poor memory
        v. May utilize communicator tactics
            1. Equivocation
            2. Evasion and concealment
            3. Submissiveness and informality
            4. Involvement and expressiveness
VI. Arousal
    a. For some types of deception, deceivers become physiologically aroused
    b. E.g., high stakes lies
    c. Internal arousal is "leaked out" in form of observable deception cues
    d. Leakage is unintentional, involuntary
    e. Not all deceivers become aroused
        i. Everyday lies are not arousing
    f. Not all arousal signals deception
    g. The Othello error: seeing aroused truth tellers as deceivers
    h. Expect these indicators ONLY IF there is reason for the deceiver to be highly aroused
VII. Arousal Indicators – Kinesics
    a. Truth-tellers are at ease
        i. Fairly relaxed faces, bodies and limbs
        ii. Little random movement
        iii. Lopsided posture
        iv. Deceivers show tension
        v. Rigid, still, over-controlled posture
        vi. Few speech-related gestures called illustrators
    b. Deceivers show nervous activation
        i. Increased self- and object-focused gestures called adaptor gestures (e.g., touching face, fidgeting with clothes or hair)
        ii. Restless lower trunk and limb movement (e.g., foot bobbing)
VIII. Arousal Indicators – Vocalics
    a. Truth-tellers have non-tense, relaxed voices
    b. Deceivers show
        i. Tension in the larynx

- ii. Nervous activation
- iii. pitch increase
- iv. tempo change
- v. non-ah nonfluencies (stutters, re-starts, sound repetitions, garbled sounds, etc.)

IX. Emotion
- a. Deception may activate emotions that “leak out” through words and actions
- b. E.g., fear of detection, guilt
- c. Deceivers may try to mask these emotions with more acceptable ones
- d. Some deceivers show “duping delight”
- e. Negative emotion cues should be LACKING with everyday, low-stakes deceptions

X. Emotion – Content
- a. Truth-tellers use more language expressing emotional states and feelings than deceivers

XI. Emotion – Linguistics
- a. Truth-tellers include more emotion-laden language in their responses
- b. Exceptions: deceivers may also use
- c. Negative language reflecting internal negative states OR
- d. Positive affect terms when trying to be persuasive

XII. Emotion – Kinesics
- a. Deceivers
  - i. may lack displays of positive facial emotions like happiness
  - ii. may show micro-momentary negative facial emotions like anger or fear
  - iii. may have fake rather than genuine smiles
    1. Fake smiles don’t use all muscles around mouth or eyes, and may last too long

XIII. Emotion – Vocalics
- a. Deceivers’ voices:
  - i. Show fewer pleasant emotions
  - ii. Are less varied (e.g., flat, monotone pitch)
  - iii. May include forced or nervous rather than relaxed laughter

XIV. Cognitive Effort
- a. Deceiving is harder work mentally than telling the truth
- b. Cognitive effort cues show someone is working hard mentally
- c. Exceptions – If the deceiver has time to:
  - i. Plan it out
  - ii. Mentally rehearse
  - iii. Practice out loud (e.g., coached testimony)
  - iv. Edit written deception before transmitting
- d. Indicators of cognitive effort may not guarantee that the person is lying

XV. Cognitive Effort – Kinesics

a. Deceivers show more
    i. Stopping gestures
    ii. Adaptor gestures
    iii. Looking away while constructing responses

XVI. Cognitive Effort – Vocalics
a. Deceivers display more
    i. Internal and between-turn pauses
    ii. Speech errors

XVII. Cognitive Effort – Between Channels
    i. Multi-channel expectancy violations:
        1. Mental difficulty coordinating all verbal and nonverbal channels
        2. May result in inconsistencies between nonverbal and verbal channels and in unexpected behaviors

XVIII. Memory
a. The accounts of remembered events may differ from fabricated events
b. Truthful accounts should be sufficiently rich in detail, complex, and specific
c. Deceptive accounts lack plausibility, are shorter, and are vague

XIX. Memory – Linguistics
a. Truthful messages are usually longer
b. More sentences
c. More words and syllables
d. Longer talk time
e. Exceptions—when deceivers:
f. Have time to plan, rehearse or edit
g. When the deceptive goal is persuasion
h. Truthful messages should include more specific details
i. Details may be:
    i. Visual
    ii. Auditory
    iii. Spatial
    iv. Temporal
    v. Other senses (touch, smell)
j. Unusual features that a deceiver would be unlikely to invent
k. Truthful messages are more likely to have:
l. Specific personal pronouns (I, me, myself, we, us)
m. Definitive modifiers such as "definitely"
n. Verbatim (direct) quotes of what was said
o. Deceptive messages are more likely to have:
    i. Fewer personal pronouns
    ii. Use of qualifiers, hedges, and modifiers such as "perhaps, possibly, seems"
    iii. Use of modal verbs (could, should, would)

    iv. Use of leveling and generalizing terms
    v. Overgeneralizations
  p. Truth-tellers are more likely to use:
    i. Bigger words
    ii. e.g., “galloped” rather than “ran”, “argued” rather than “said”
    iii. More complex and compound sentences with subordinate clauses and phrases
    iv. Conjunctions such as “and,” “but”, “yet,” “although,” “because” that create compound sentences
  q. Truth tellers are more likely to use more different words
  r. Deceivers are more likely to repeat the same words and phrases
  s. Truthful accounts “make sense”
    i. Need not be orderly but:
    ii. “hang together”
    iii. are plausible
  t. Deceptive accounts seem implausible
  u. Exception: when deceivers can plan, edit or rehearse

XX. Communicator Tactics
  a. Different message forms for different types of deception
  b. Motivation & goal-orientation influence performance
  c. Skillful deceivers and those who can plan, rehearse or edit their messages produce deceptive messages in a more strategic way

XXI. Strategies and Tactics
  a. Equivocation
  b. strategic ambiguity/uncertainty
  c. Evasion and concealment
  d. Submissiveness and informality
  e. Involvement and expressiveness

XXII. Equivocation
  a. Non-immediacy
  b. Disassociates or distances the speaker from his/her utterances, obscures actor’s personal responsibility
  c. Use of passive voice
  d. Lack of first-person pronouns
  e. Increase in group references
  f. Changes in verb tense from past to present or vice versa
  g. Vagueness
  h. Levelers
  i. Modal verbs
  j. Qualifiers

XXIII. Evasion and Concealment
  a. Shorter and simpler messages
  b. Less specificity (details)
  c. More pauses

d. Deceivers use repetition and rephrasing as a delay tactic, e.g.
e. "You were asking, where was I that I could see what happened to the soldiers?"
f. Capitalize on other's comments to "fill in the blanks," create plausible answers

XXIV. Submissiveness and Informality
a. Tag questions—questions at the end of a sentence that make statement tentative, seek listener's agreement
b. E.g., "This is what you want me to say, right?"
c. Informality—slang, typographical errors make speaker appear more approachable
d. Higher pitch--is associated with non-dominance
e. Longer response latencies--shift responsibility to other to "fill in the silences"
f. Softer, quieter voice

XXV. Involvement and Expressiveness
a. At first, deceivers appear less involved and expressive than truth-tellers
b. Monotone voice
c. Limited gesturing
d. Stiff posture
e. Less expressive language
f. BUT, over time, deceivers may increase involvement and expressiveness to appear more normal

XXVI. Conclusions
a. In general, deceivers:
b. Show more arousal
c. Show unpleasant emotion
d. Show higher cognitive effort
e. Demonstrate poor memory
f. May utilize communicator tactics
   i. Equivocation
   ii. Evasion and concealment
   iii. Submissiveness and informality
   iv. Involvement and expressiveness
g. No single indicator is absolute
h. Consider indicators in combination
i. Consider indicators in context
j. Watch for changes over time
k. Consider how to counteract planning and rehearsal effects

Knowledge Test

The following is the knowledge quiz that was provided to novices and professionals as part of the training. Correct answers are indicated by an (*).

Please answer the following questions about deception and deception detection

Deception

a. refers to intentionally or unintentionally sending bad information
b. only includes outright lying
c. includes falsifying or concealing*
d. excludes white lies, evasions, or exaggerations

Typically, people successfully detect deception about _____ of the time

a. 20%
b. 50%*
c. 80%
d. 90%

Which of the following is NOT a reliable kinesic indicator of deception?

a. fake smiles
b. amount of eye contact*
c. amount of gesturing
d. stiff posture

Accuracy in detecting deceit typically is greater when a receiver attends to

a. facial cues rather than body or vocal cues
b. nonverbal cues rather than verbal cues
c. visual cues (eyes, face, body) rather than verbal cues
d. auditory cues (vocal, verbal) rather than visual*

What is a reliable vocal indicator of deception?

a. slowed rate of speech*
b. relaxed laughter
c. few pauses in speech
d. lower voice pitch

Deception is best thought of as

a. a dynamic, changing process*
b. uncontrolled and unconscious actions "leaking out" to betray the truth
c. a set of fixed and stable verbal signs
d. involving senders only, not receivers

What is a reliable kinesic indicator of deception?

a. the speaker is leaning forward
b. the speaker has stiff, wooden posture*
c. a highly expressive face
d. relaxed posture

It is possible that deceivers are having a difficult time lying if you notice that they:

a. respond immediately after being asked a question
b. behave in a normal manner
c. drop the names of others into a conversation
d. stop gesturing*

A _________ may be used by deceivers to mask their negativity toward others

a. non-ah nonfluency
b. self grooming behavior
c. feigned smile*
d. long response latency

According to the leakage hypothesis, the arousal that often accompanies deception will most likely be signaled by:

a. vocalic behavior
b. facial behavior
c. verbal behavior
d. body behavior*

Judgment Test

The following are the explanations that were provided following each of the five judgments that professionals and novices performed as part of the training.

Judgment 1 – Deceptive

Cues exhibited:

Indirect answer "I didn't see anything…"
Limited hand gesturing
Pushed lips together
Shoulder shrugs
Overall tenseness

Judgment 2 – Truthful

Cues exhibited:

Shows positive emotions
Genuine smiles
Expressive head movement

Judgment 3 – Deceptive

Cues exhibited:

Pressed lips together
Smiling at inappropriate times
  -Seems eager to lie
Lack of positive emotions
Limited hand movements

Judgment 4 – Truthful

Cues exhibited:

Involved and detailed explanation
Many gestures
Confidence in voice and demeanor

Judgment 5 – Deceptive

Cues exhibited:

Significant hand fidgeting
Lack of positive emotions
Rigid head movements
Little gesturing

APPENDIX C – EXPERIMENT ORIENTATION

The following is the script that accompanied the orientation video that all participants saw. Participants in the higher-accuracy and lower-accuracy BAP conditions received special instructions that are indicated within the script.

"Deception detection can be a difficult task for humans to perform. In fact, most people can't really tell if someone is deceiving them or not. In an effort to improve detection accuracy, the Center for the Management of Information is developing a computer system called the Behavioral Analysis System, or BAS, that observes human behavior for signs of possible deceit.

"The BAS analyzes the kinesic and linguistic properties of communication from one person in a face-to-face interaction. It then analyzes these characteristics to give insight into whether or not the observed person is being deceitful.

"During this study, you will be shown 10 videotaped interviews and your task will be to decide if the interviewees are being deceptive or truthful. You will also specify how confident you are in that decision. Some of you will have access to the BAS during the detection task. For those who have access to the BAS, you will be free to use it as much or as little as you wish in coming to your decision. The purpose of this study is to explore the effects the BAS has on your decision making and detection accuracy. We will be observing how you use the BAS and how you come to your final judgments.

--BAP Conditions--

“The following is a brief orientation about the BAS. The BAS was developed as part of a research initiative funded by the Department of Homeland Security and the Air Force Office of Scientific Research.

“The BAS analyzes two channels of communication: Kinesics – characteristics of movement; and Linguistics – properties of what is said.

“The kinesic channel is analyzed using technology to track the positions of the head and hands throughout a video segment. From the head and hand tracking, the system creates features which are helpful in identifying when someone is lying.

“Using these features, behavioral signatures can be developed which highlight when someone is trying to mislead or conceal the truth. Such behavioral signatures may indicate tenseness or uncertainty, which can be a sign of deception.

“The linguistic channel is analyzed using technology which reviews a transcript of an interview or other interaction. After reviewing the transcript, features which are helpful in identifying when someone is lying are created. Some of these features include the number and types of words which are used in the interview. From these features, the system can provide insight into whether or not a person is trying to deceive.

“After the linguistic and kinesic analyses are complete, they are merged to form a final judgment. It is important to note that in past testing, the BAS correctly identifies deception 60 to 80 percent of the time. The remaining 40 to 20 percent of the time, the BAS provides an incorrect judgment.

“During this experiment you will have access to the BAS during the detection task. The BAS interface looks like this. During the detection task you will be shown 10 video clips, each one will be deceptive or truthful. The video clips will be shown here and will not be repeated, so please pay careful attention. Should you need to adjust the volume, adjustments are possible here.

"After each clip, you will be asked to provide an initial judgment concerning whether or not you think the person was guilty of cheating, the amount of deception in the interview, and how confident you are about your judgment.

"After your initial judgment, you will have access to the BAS. The BAS will provide scores based on kinesic analysis and on linguistic analysis. Explanations concerning each of these analyses can be accessed by clicking on the Cues or Analyses buttons. The BAS joins the kinesic and linguistic scores and offers an overall score, level of confidence, and judgment. Should you have any questions about any of these items, you may click on Define to learn more.

"After you have reviewed the information the BAS provides, you will then be required to submit a final judgment where you will again indicate whether or not you think the person cheated, the level of deception during the interview, and your level of confidence in your judgment. There will be a brief questionnaire after every final judgment.

--No BAP Conditions--

"During this experiment, you alone will have to determine whether or not you think someone is lying without the help of the BAS. The interface that you will use looks like this.

"During the task you will be shown 10 video clips, each one will be deceptive or truthful. The video clips will be shown here and will not be repeated, so please pay careful attention. Should you need to adjust the volume, adjustments are possible here. After each clip, you will be asked to provide an initial judgment concerning whether or not you think the person was guilty of cheating, the amount of deception in the interview, and how confident you are about your judgment. There will be a brief questionnaire after every judgment.

--All Participants--

"You will also receive training on how to detect deception. Some of you will receive this training before the detection task; others will receive the training after the detection task. The training was compiled from experts in deception detection and will outline numerous cues (or indicators) that deceivers frequently exhibit. The training includes an instructional video, a series of review questions and practice in deception detection.

"The following is brief description of the 10 video clips that you will judge.

"Students from an introductory communication course at a large university were invited to participate in a study. The participants were informed that the study concerned effective teamwork. The participants learned they would be working in pairs to answer difficult trivia questions and they were promised a large cash reward if they performed well on the trivia questions. Unknown to the participants was that the partner they were paired with was a confederate, meaning she was part of the experiment. After the pairing, an experimenter entered the room and asked a number of extremely difficult trivia questions.

“After asking a few questions, the experimenter was called out of the room and left the set and answers in the room with the participant and confederate. The confederate then encouraged the participant to cheat and look at the answers. The participant then either cheated and looked at the answers or refused to cheat.

“After a few minutes, the experimenter returned and finished asking the trivia questions. After all the trivia questions were complete, each participant was brought to an interview where he or she was questioned concerning any possible cheating. All participants were interviewed by a single interviewer and the interviewer posed the same questions to all participants. You will view 10 of these interviews and watch the responses to all 6 questions for each interview. Here is a sample of a truthful interview.

--Novices Only--

“You will all receive extra credit for participating in this experiment. In addition, those who perform in the top 10% of each group will be awarded 10 dollars cash after performance has been assessed and recorded. Groups will be divided according to what you had access to. For instance, those who have training before the detection task will be in one group, those who received training after the detection task will be in another group.

“Please hold on to the slip of paper with your three digit number that you received at the beginning of the experiment. This paper serves as proof of your identity and if you qualified for a reward, you will need to turn it in to receive your compensation.

--All Participants--

“The experiment will now begin. Please pay close attention to the instructions that you will be given. Good luck!”

APPENDIX D – SURVEY INSTRUMENTS

The surveys were administered at 12 junctures during the detection task. Before the detection task, participants filled out a 15-item questionnaire that captured background data and predictions for how well the participants thought they would do (Initial Survey). Then following judgment of each of the 10 interviews, participants answered a brief 6-item questionnaire (Judgment Surveys). Finally, following the conclusion of the detection task, the participants completed a 24-item questionnaire concerning perceptions of the BAP system and estimations about how well participants thought they did (Final Survey).

Not all of the questions in the questionnaire were relevant in all conditions. For example, the novices without access to the BAP were not questioned about their BAP perceptions and novices were not questioned about the number of years they worked as a law enforcement officer. Irrelevant questions were omitted from the questionnaires for such conditions. However, all items are included in the surveys shared below. Unless otherwise indicated, all items were answered on a 7-point scale (1 – Strongly Disagree, 4 – Neutral, 7 – Strongly Agree).

Initial Survey

Please answer the following questions regarding your background and opinions.

1. I consider myself a computer expert
2. I have no experience with computer programs that provide recommendations or suggestions
3. I feel comfortable using a computer to solve problems
4. In general, computers are helpful and provide useful information when I have a decision to make

5. In general, I trust the information that computers provide when I have a decision to make.
6. I am generally a trusting person
7. I don't trust someone or something when I have little knowledge of it
8. Once I make a decision, I rarely change my mind
9. I am generally confident in the choices that I make

The experimental task will require you to make 10 judgments of interviews. You will be asked if the person being interviewed is being deceptive or not. Please answer the following questions about the task that you will complete.

10. I consider myself an expert at detecting when someone is lying
11. I expect this task to require a great deal of effort
12. I believe this task will be a difficult task
13. When I decide someone is lying to me, I am usually sure of my judgment
14. Please state the number (out of ten) interviews that you expect to correctly identify as deceptive or truthful ______ (0-10)
15. I will be satisfied if I get _____ interviews out of 10 correct (0-10)

Judgment Survey

Please answer the following questions relating to your judgment.

1. This judgment was difficult for me to make
2. The BAS helped me make my decision
3. This judgment required a great deal of effort
4. The BAS was useless in forming my decision
5. The BAS compelled me to consider additional information than what I observed
6. Please select the most correct statement
   a. I was unsure about my judgment and I went with what I thought
   b. The BAS and I initially disagreed, but I was sure about my judgment and I ignored the BAS's judgment
   c. The BAS and I came to the same judgment
   d. The BAS and I initially disagreed, but I was sure about the BAS's judgment and I ignored what I thought
   e. I was unsure about my judgment and I went with the BAS's judgment
   f. Other (please explain): (Fill in)

Final Survey

Please answer these questions relating to the experiment and your performance

1. This task required a great deal of effort
2. The BAS and I worked well together
3. This task was difficult
4. I felt confident about relying on the BAS for my decisions
5. I examined behaviors very carefully in order to make sure that all potential deceptive cues were considered
6. I would feel comfortable about relying on the BAS for my decisions in the future
7. I would not be willing to let the BAS make my decisions for me
8. I most frequently tried to observe the cues that the BAS was searching for
9. I would use the BAS as an aid to help with my decision about whether or not someone is being deceptive
10. I most frequently paid attention to the cues that were described in the training I received
11. I would let the BAS suggest to me a number of reasons why a person may be deceptive
12. The BAS is a real expert in assessing deception
13. The BAS has a good knowledge of deception and behaviors associated with deception
14. Please state the number interviews that you think you correctly classified (0-10)
15. I would let the BAS assist me in deciding if someone is being deceptive
16. I will be satisfied if I classified _____ interviews out of 10 correctly. (0-10)
17. In general, I agreed with the BAS's recommendation
18. The cues I was observing did not overlap very much with the cues the BAS was observing
19. I would let the BAS decide for me if a person is being deceptive

Please answer these questions relating to demographics

20. Your sex: (Male/Female)
21. Your age: (Fill in)
22. Number of years of education after high school: (Fill in)
23. Please specify the number of years you have been employed as a law enforcement officer (Fill in)
24. Please state your ethnicity: (Black, Non Hispanic; American Indian or Alaskan Native; Hispanic; White, Non Hispanic; Other)

Multi-item Validation

In the final survey, a total of 9 items were adapted from Komiak [68] that elicited information about emotional trust in the BAP, intention to use the BAP as a delegated agent, intention to use the BAP as a decision aid, and competence of the decision aid. Komiak demonstrated proper item reliability and construct validity of the items and constructs that were used in this dissertation. However, items and constructs were reviewed again to ensure reliability and validity.

First, items were reviewed by experts in survey research for face validity and item clarity. Then items were utilized in a pretest (pilot experiment) where respondents had the opportunity to interact with the experimenter concerning wording and presentation of the questionnaires.

Following the conduct of the experiments, validity and reliability measures were examined for the constructs outlined by Komiak. Abbreviations of the items and the constructs they are meant to capture are shown in Table D-1. Reliability across items within a construct are presented via Cronbach's α (N = 139). With the exception of perceptions of BAP competence, the items demonstrate reliability ($\alpha \geq .70$) [53].

Table D-1 Multi-item survey constructs

Constructs	Cronbach's α	Items
Emotional Trust	.79	ConfidentRelyingOnBAS
		FeelComfortableRelyingOnBAS
Delegated Agent Intention	.70	NotLetBASJudgeForMe (reverse coded)
		LetBASDecideForMe
Decision Aid Intention	.82	LetBASHelp
		LetBASSuggestReasons
		LetBASAssistMe
BAP Competence	.49	BASIsExpert
		BASHasGoodKnowledgeOfDeception

To demonstrate convergent validity among same-construct items and discriminant validity between different-construct items, a factor analysis was performed with a maximum likelihood extraction method and a Varimax rotation with Kaiser Normalization. As four constructs were taken from Komiak's survey instruments, the factor analysis was constrained to four factors. Table D-2 shows the results of the rotated factor matrix.

Table D-2 Rotated factor matrix with survey items

Constructs	Items	Factor 1	Factor 2	Factor 3	Factor 4
Emotional Trust	ConfidentRelyingOnBAS	**.733**	.309	.229	.094
	FeelComfortableRelyingOnBAS	**.555**	.544	.292	.077
Delegated Agent	NotLetBASJudgeForMe (reverse coded)	.129	**.873**	.007	.034
Intention	LetBASDecideForMe	.368	**.566**	.033	-.045
Decision Aid	LetBASHelp	.268	.055	**.679**	.269
Intention	LetBASSuggestReasons	.092	.015	**.449**	.802
	LetBASAssistMe	.120	.065	**.986**	.095
BAP Competence	BASIsExpert	.447	.472	.332	**.202**
	BASHasGoodKnowledgeOfDeception	.134	.095	.524	**.228**

Review of the rotated factor matrix uncovers serious problems with convergent and discriminant validity in the BAP competence construct. This finding taken together with the low reliability between items undermined the value of the measures. Therefore the BAP competency construct was excluded from analysis.

There were additional minor problems with the remaining constructs. Notably, LetBASSuggestReasons and FeelComfortableRelyingOnBAS demonstrated significant cross loading. However, the remaining items appeared to load on the proper factors. Although there were violations of discriminant validity, these violations were considered minor and the items were retained and analyzed. Additionally, Komiak noted in her analysis of these measures that there was significant correlation between emotional trust, delegated agent intention, and decision aid intention [68].

APPENDIX E – NOVICE ANCHORING AND ADJUSTMENT PILOT STUDY

According to studies examining the anchoring and adjustment bias, individuals who are given an initial value in response to a question anchor on the given value (see [104] for a succinct review). Once anchored, the individuals rarely deviate very far – even when anchoring values are randomly assigned [132]. In the context of deception detection with the BAP, the anchoring and adjustment bias may have become evident resulting from the request for an initial judgment. Under this possibility, the act of submitting an initial judgment would serve as an anchor and deviations from that initial judgment would be unlikely. If the anchoring and adjustment bias were affecting the human users' judgments, that effect would dissipate if the request for an initial judgment were removed.

To examine the possibility that the human-anchoring seen in both the novice-only and novice/professional experiments were results of the anchoring and adjustment bias, a pilot study was conducted. The pilot study examined novice BAP usage with and without an initial judgment. If the human-anchoring was resultant from the anchoring and adjustment bias, the amount of human-anchoring should have decreased when the initial judgment request was removed.

Two groups of novice participants were used for this pilot study. The first group of participants was originally used as part of the novice-only experiment described in Chapter 5. The 31 participants that were included in this pilot study comprised the training/access to lower-access BAP group. In the first group, the mean age of the participants was 21.9, mean years of secondary education was 3.9, and of all the participants, 16.0 percent were female.

The second group of participants was recruited just for this pilot study. They were recruited from an MIS graduate seminar class and numbered 8 in total. In the second group, the mean age of the novice participants was 23.0, mean years of secondary education was 5.0, and of all the participants, 13.0 percent were female.

The MSU cheating interviews produced by Timothy Levine and colleagues [72] were used as the stimulus materials in the pilot study (see section 5.2 for a full description).

The single independent variable in the pilot study was the requirement of an initial judgment. The experimental design matrix and sample size in each cell are illustrated in Table E-1.

Table E-1 Experimental design matrix for anchoring and adjustment pilot study

	Initial Judgment	No Initial Judgment
Training/Access to Lower-accuracy BAP	$N = 31$	$N = 8$

To determine if there was a difference in the utilization and adoption of the BAP recommendation between the initial judgment condition and no initial judgment condition, a two sample *t*-test was performed (two-tailed). The *t*-test dealt with the number of final judgments that matched the BAP recommendation. There was no significant difference in the number of times the novices accepted the BAP recommendation across the initial judgment and no initial judgment conditions, $t(37) = 1.57$, $p = .123$.

The small sample size of the no initial judgment condition is an acknowledged weakness of the pilot study, so care should be taken in interpreting the results. Nevertheless, these preliminary findings seem to suggest that there is no difference in BAP utilization when an initial judgment is requested and when an initial judgment is not requested. This finding suggests that the anchoring and adjustment bias is not the source of the human anchoring seen in both the novice-only and novice/professional experiments.

References

[1] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton, "The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions," *MIS Quarterly,* vol. 30, 2006.

[2] G. Ben-Shakhar and E. Elaad, "The Validity of Psychophysiological Detection of Information with the Guilty Knowledge Test: A Meta-Analytic Review," *Journal of Applied Psychology,* vol. 88, pp. 131-151, 2003.

[3] M. Benaroch and M. Tanniru, "Conceptualizing structurable tasks in the development of knowledge-based systems," *Decision Sciences,* vol. 27, pp. 415-427, 1996.

[4] I. Benbasat, G. DeSanctis, and B. R. Nault, "Empirical research in managerial support systems: A review and assessment," in *Recent developments in decision support systems*, C. W. Holsapple and A. B. Whinston, Eds. New York, NY: Springer-Verlag, 1993.

[5] I. Benbasat and R. W. Zmud, "The identity crisis within the IS discipline: Defining and communicating the disciplines core properties," *MIS Quarterly,* vol. 27, pp. 183-194, 2003.

[6] J. L. Bennett, *Building decision support systems*. Reading, MA: Addison-Wesley, 1983.

[7] P. Berthon, P. Leyland, M. Ewing, and C. L. Carr, "Potential research space in MIS: A framework for envisioning and evaluating research replication, extension, and generation," *Information Systems Research,* vol. 13, pp. 416-427, 2002.

[8] R. H. Bonczek, C. W. Holsapple, and A. B. Whinston, *Foundations of decision support systems* vol. 14. New York, NY: Academic Press, 1981.

[9] C. F. Bond and B. M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Review,* vol. 10, pp. 214-234, 2006.

[10] K. Bontcheva, H. Cunningham, V. Tablan, D. Maynard, and O. Hamza, "Using GATE as an Environment for Teaching NLP," in *ACL Workshop on Effective Tools and Methodologies in Teaching NLP*, 2002.

[11] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review,* vol. 78, pp. 1-3, 1950.

[12] R. Bull, "Can training enhance the detection of deception," in *Credibility Assessment*, J. C. Yuille, Ed. Norwell, MA: Kluwer Academic Publishers, 1989.

[13] D. Buller, J. Burgoon, C. White, and A. Ebesu, "Interpersonal deception: VII. Behavioral profiles of falsification, equivocation and concealment," *Journal of Language and Social Psychology,* vol. 13, pp. 366-395, 1994.

[14] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Communication Theory,* vol. 6, pp. 203–242, 1996.

[15] D. B. Buller, J. K. Burgoon, A. Buslig, and J. Roiger, "Interpersonal deception: VIII. Nonverbal and verbal correlates of equivocation from the Bavelas et al. (1990) research," *Journal of Language and Social Psychology,* vol. 13, pp. 396-417, 1994.

[16] D. B. Buller, J. K. Burgoon, A. Buslig, and J. Roiger, "Testing interpersonal deception theory: The language of interpersonal deception," *Communication Theory,* vol. 6, pp. 268-289, 1996.

[17] J. K. Burgoon, M. Adkins, J. Kruse, M. L. Jensen, T. Meservy, D. P. Twitchell, A. Deokar, J. F. Nunamaker, S. Lu, G. Tsechpenakis, D. N. Metaxas, and R. E. Younger, "An approach for intent identification by building on deception detection," in *38th Annual Hawaii International Conference on System Science (CD/ROM)*, Hawaii, 2005.

[18] J. K. Burgoon, D. B. Buller, L. K. Guerrero, W. A. Afifi, and C. M. Feldman, "Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful mesages," *Communication Monographs,* vol. 63, pp. 50-69, 1996.

[19] J. K. Burgoon, D. B. Buller, and W. G. Woodall, *Nonverbal communication: The unspoken dialogue*. New York, NY: The McGraw-Hill Companies, Inc., 1996.

[20] J. K. Burgoon, M. L. Jensen, J. Kruse, T. O. Meservy, and J. Jay F. Nunamaker, "Deception and intention detection," in *Handbooks in Information Systems*. vol. 2, H. Chen, T. S. Raghu, R. Ramesh, A. Vinze, and D. Zeng, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2007.

[21] K. Burns, "Mental models and normal errors," in *How professionals make decisions*, H. Montgomery, R. Lipshitz, and B. Brehmer, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

[22] I. N. Chengalur-Smith, D. P. Ballou, and H. L. Pazer, "The impact of data quality information on decision making: An exploratory analysis," *IEEE Transactions on Knowledge and Data Engineering,* vol. 11, pp. 853-864, 1999.

[23] M. S. Cohen, "Three paradigms for viewing decision biases," in *Decision making in action: Models and methods*, G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok, Eds. Norwood, NJ: Ablex Publishing Corp., 1993.

[24] H. Cunningham, "GATE, a General Architecture for Text Engineering," *Computers and the Humanities,* vol. 36, pp. 223-254, 2002.

[25] B. DePaulo, K. Charlton, H. Cooper, J. Lindsay, and L. Muhlenbruck, "The accuracy-confidence correlation in the detection of deception," *Personality and Social Psychology Review,* vol. 1, pp. 346-357, 1997.

[26] B. DePaulo, J. Lindsay, B. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychological Bulletin,* vol. 129, pp. 74-118, 2003.

[27] B. M. DePaulo, "Spotting lies: Can humans learn to do better?," *Current Directions in Pyschological Science,* vol. 3, pp. 83-86, 1994.

[28] B. M. DePaulo and D. A. Kashy, "Everyday lies in close and casual relationships," *Journal of Personality and Social Psychology,* vol. 74, pp. 63-79, 1998.

[29] B. M. DePaulo and R. L. Pfeifer, "On-the-job experience and skill at detecting deception," *Journal of Applied Social Psychology,* vol. 16, pp. 249–267, 1986.

[30] J. S. Dhaliwal and I. Benbasat, "The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation," *Information Systems Research,* vol. 7, pp. 342-362, 1996.

[31] C. B. R. Diller and K. Wiers, "Deceiving Others," in *Nonverbal Course*, J. K. Burgoon, Ed., 2004.

[32] H. J. Einhorn and R. M. Hogarth, "Unit weighting schemes for decision making," *Organizational Behavior and Human Performance,* vol. 13, pp. 171-192, 1975.

[33] P. Ekman, *Telling lies*. New York: W. W. Norton & Company, 1985.

[34] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry,* vol. 32, pp. 88-106, 1969.

[35] P. Ekman and M. O'Sullivan, "Who can catch a liar?," *American Psychologist,* vol. 46, pp. 913-920, 1991.

[36] R. V. Exline, J. Thibaut, C. B. Hickey, and P. Gumpert, "Visual interaction in relation to Machiavellianism and an unethical act," in *Studies in Machiavellianism*, R. Christie and F. L. Geis, Eds. New York, NY: Academic Press, 1970.

[37] D. L. Faigman, S. E. Fienberg, and P. C. Stern, "Limits of the Polygraph," *Issues in Science and Technology,* pp. 40-46, Fall 2003 2003.

[38] L. A. Farwell and E. Donchin, "The truth will out: interrogative polygraphy ("lie detection") with event-related brain potentials," *Psychophysiology,* vol. 28, pp. 531-47, 1991.

[39] F. Faul and E. Erdfelder, "GPOWER: A priori, post hoc, compromise power analysis for MS-DOS," Bonn, Germany: Bonn University, 1992.

[40] C. W. Fisher, I. Chengalur-Smith, and D. P. Ballou, "The impact of experience and time on the use of data quality information in decision making," *Information Systems Research,* vol. 14, pp. 170-188, 2003.

[41] S. Fiske and S. Taylor, *Social cognition*. Reading, MA: Addison-Wesley, 1991.

[42] P. M. Fitts, "Human engineering for an effective air navigation and traffic control system," Ohio State University Research Foundation Report, Columbus, OH 1951.

[43] M. G. Frank and T. H. Feeley, "To Catch a Liar: Challenges for Research in Lie Detection Training," *Journal of Applied Communication Research,* vol. 31, pp. 58-75, 2003.

[44] S. Freud, *Collected papers*. New York, NY: Basic Books, 1959.

[45] G. Ganis, S. M. Kosslyn, S. Stose, W. L. Thompson, and D. A. Yurgelun-Todd, "Neural Correlates of Different Types of Deception: An fMRI Investigation," *Cerebral Cortex,* vol. 13, pp. 830-836, 2003.

[46] H. Gelernter, "Realization of a geometry-theorem proving machine," in *International Conference on Information Processing*, Paris, France, 1959, pp. 273-282.

[47] J. F. George, D. P. Biros, M. Adkins, J. K. Burgoon, and J. F. Nunamaker, *Testing Various Modes of Computer-Based Training for Deception Detection* vol. 3073. Heidelberg: Springer-Verlag, 2004.

[48] J. F. George, K. Marett, J. Crews, J. Cao, M. Lin, D. P. Biros, and J. K. Burgoon, "Training to detect deception: an experimental investigation," in *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, Big Island, HI, 2004.

[49] T. G. Gill, "Early expert systems: Where are they now?," *MIS Quarterly,* vol. 19, pp. 51-81, 1998.

[50] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," *MIS Quarterly,* vol. 19, pp. 213-236, 1995.

[51] S. Gregor and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," *MIS Quarterly,* vol. 23, pp. 497-530, 1999.

[52] V. Grover, C. C. Lee, and D. Durand, "Analyzing methodological rigor of MIS survey research from 1980-1989," *Information & Management,* vol. 24, pp. 305-317, 1993.

[53] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Multivariate data analysis*. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

[54] M. Hartwig, P. A. Granhag, L. A. Stromwall, and A. Vrij, "Police officers' lie detection accuracy: Interrogating freely verses observing video," *Police Quarterly,* vol. 7, pp. 429-456, 2004.

[55] F. Hayes-Roth, "Rule-based systems," *Communications of the ACM,* vol. 28, pp. 921-932, 1985.

[56] F. Hayes-Roth and N. Jacobstein, "The state of knowledge-based systems," *Communications of the ACM,* vol. 37, pp. 26-39, 1994.

[57] A. R. Hevner, S. March, J. Park, and S. Ram, "Design science research in information systems," *MIS Quarterly,* vol. 28, pp. 75-105, 2004.

[58] G. P. Huber, "Cognitive style as a basis for MIS and DSS designs: Much ado about nothing?," *Management Science,* vol. 29, pp. 567-579, 1983.

[59] M. L. Jensen, "An Examination and Cross-validation of Automatically-generated, Behavioral Features in Deception Detection," Tucson, AZ, 2005.

[60] M. L. Jensen, T. O. Meservy, J. K. Burgoon, and J. F. Nunamaker, "Automatic, multimodal evaluation of human interaction," Tucson, AZ, 2007, pp. 1-19.

[61] M. L. Jensen, T. O. Meservy, J. Kruse, J. K. Burgoon, and J. Jay F. Nunamaker, "Identification of Deceptive Behavioral Cues Extracted from Video," in *International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria, 2005.

[62] S. M. Kassin and C. T. Fong, ""I'm innocent!": Effects of training on judgments of truth and deception," *Law and Human Behavior,* vol. 23, pp. 499-516, 1999.

[63] P. Keen and M. S. Scott Morton, *Decision support systems: An organizational perspective*. Reading, MA: Addison-Wesley, 1978.

[64] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for Automatic Segmentation of Audio Data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[65] G. Keppel and T. D. Wickens, *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice Hall, 2004.

[66] B. D. Klein, D. L. Goodhue, and G. B. Davis, "Can humans detect errors in data? Impact of base rates, incentives, and goals," *MIS Quarterly,* vol. 21, pp. 169-194, 1997.

[67] M. R. Klein and L. B. Methlie, *Knowledge-based decision support systems with applications in business*. New York, NY: John Wiley & Sons, 1995.

[68] S. Y. X. Komiak, "The impact of internalization and familiarity on trust and adoption," in *Sauder School of Business*. vol. PhD Vancouver, BC: University of British Columbia, 2003, p. 182.

[69] R. Kraut, "Humans as lie detectors," *Journal of Communication,* vol. 30, pp. 209–216, 1980.

[70] J. Kruger and D. Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of Personality and Social Psychology,* vol. 77, pp. 1121-1134, 1999.

[71] S. Le Cessie and J. C. Van Houwelingen, " Ridge estimators in logistic regression," *Applied Statistics,* vol. 41, pp. 191-201, 1997.

[72] T. R. Levine, "MSU Trivia Videos," Michigan State University, 2004.

[73] T. R. Levine, T. H. Feeley, S. A. McCornack, H. M., and C. M. Harms, "Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group," *Western Journal of Communication,* vol. 69, pp. 203–217, 2005.

[74] T. R. Levine, R. K. Kim, H. S. Park, and M. Hughes, "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,* vol. 73, pp. 243 - 260, 2006.

[75] T. R. Levine, H. S. Park, and S. A. McCornack, "Accuracy in detecting truths and lies: Documenting the "veracity effect"," *Communication Monographs,* vol. 66, pp. 125-144, 1999.

[76] S. Lu, G. Tsechpenakis, D. N. Metaxas, M. L. Jensen, and J. Kruse, "Blob analysis of the head and hands: A method for deception detection," in *38th Hawaii International Conference on System Science (CD/ROM)*, Hawaii, 2005.

[77] S. Mann, A. Vrij, and R. Bull, "Detecting true lies: Police officers' ability to detect suspects' lies," *Journal of Applied Psychology,* vol. 89, pp. 137-149, 2004.

[78] J. Mao and I. Benbasat, "The use of explanations in knowledge-based systems: Cognitive perspectives and process-tracing analysis," *Journal of Management Information Systems,* vol. 17, pp. 153-179, 2000.

[79] R. C. Mayer, J. H. Davis, and E. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review,* vol. 20, pp. 709-734, 1995.

[80] S. A. McCornack and M. R. Parks, "Deception detection and relationship development: The other side of trust," in *Communication Yearbook*, M. L. McLaughlin, Ed. Beverly Hills, CA: Sage, 1986.

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

[82] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, and J. F. Nunamaker, "Automatic extraction of deceptive behavioral cues from video," in *Intelligence and Security Informatics: Proceedings of the IEEE International Conference on Intelligence and Security Informatics ISI 2005, Atlanta, GA, USA*. vol. 3495, P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen, and R. Merkle, Eds. Berlin: Springer - Verlag, 2005, pp. 198-208.

[83] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, and J. F. Nunamaker, "Detecting deception in a security screening scenario," in *38th Hawaii International Conference on System Science (CD/ROM)*, Hawaii, 2006.

[84] T. O. Meservy, M. L. Jensen, J. Kruse, D. P. Twitchell, G. Tsechpenakis, J. K. Burgoon, D. N. Metaxas, and J. F. Nunamaker, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems,* vol. 20, Sept/Oct 2005 2005.

[85] L. S. Meyers, G. Gamst, and A. J. Guarino, *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage Publications, Inc, 2006.

[86] G. R. Miller and J. B. Stiff, *Deceptive communication*. Newbury Park, CA: Sage, 1993.

[87] K. L. Mosier and L. J. Skitka, "Human decision makers and automated decision aids: Made for each other?," in *Automation and Human Performance*. vol. 1, R. Parasuraman and M. Mouloua, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 1996, pp. 201-220.

[88] R. R. Nelson, P. A. Todd, and B. H. Wixom, "Antecedents of information and system quality: An empirical examination within the context of data warehousing," *Journal of Management Information Systems,* vol. 21, pp. 199-235, 2005.

[89] A. Newell and H. A. Simon, "The logic theory machine: A complex information processing system," *IEEE Transactions on Information Theory,* vol. 2, 1956.

[90] A. Newell and H. A. Simon, "GPS, A program that simulates human thought," in *Lernende Automaten*, H. Billing, Ed. Munich, Germany: Oldenbourg, R, 1961, pp. 109-124.

[91] A. Newell and H. A. Simon, *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

[92] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin,* vol. 29, pp. 665 - 675, 2003.

[93] J. F. Nunamaker, M. Chen, and T. D. M. Purdin, "Systems development in information systems research," *Journal of Management Information Systems,* vol. 7, pp. 89-106, 1990.

[94] M. O'Sullivan, P. Ekman, and W. V. Friesen, "The effect of comparisons on detecting deceit," *Journal of Nonverbal Behavior,* vol. 12, pp. 203-215, 1988.

[95] J. Orasanu and T. Connolly, "The reinvention of decision making," in *Decision making in action: Models and methods*, G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok, Eds. Norwood, NJ: Ablex Publishing, Corp., 1993.

[96] W. J. Orlikowski and C. S. Iacono, "Desperately seeking the "IT" in IT research-A call to theorizing the IT artifact," *Information Systems Research,* vol. 12, pp. 121-134, 2001.

[97] R. Parasuraman and V. Riley, "Humans and automation: use, misuse, disuse, abuse," *Human Factors,* vol. 39, pp. 230-254, 1997.

[98] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 30, pp. 286-297, 2000.

[99] H. S. Park, T. R. Levine, S. A. McCornack, K. Morrison, and M. Ferrara, "How people really detect lies," *Communication Monographs,* vol. 69, pp. 144-157, 2002.

[100] A. S. Patrick, "Building trustworthy software agents," *IEEE Internet Computing,* vol. 6, pp. 46-53, 2002.

[101] I. Pavlidis, N. L. Eberhardt, and J. Levine, "Seeing through the face of deception: Thermal imaging offers a promising hands-off approach to mass security screening," *Nature,* vol. 415, p. 35, January 2002 2002.

[102] I. Pavlidis and J. Levine, "Thermal Image Analysis for Polygraph Testing," *IEEE Engineering in Medicine and Biology,* pp. 56-64, November/December 2002 2002.

[103] J. W. Payne, J. R. Bettman, and E. J. Johnson, *The adaptive decision maker*. Cambridge: Cambridge University Press, 1993.

[104] S. Plous, *The psychology of judgment and decision making*. New York, NY: McGraw Hill, Inc., 1993.

[105] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *Journal of the American Statistical Association,* vol. 73, pp. 699-705, 1978.

[106] F. Quek, X.-F. Ma, and R. Bryll, "A Parallel Algorithm for Dynamic Gesture Tracking," in *International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Corfu, Greece, 1999.

[107] J. P. Raman, "Deception detection through blob analysis aids homeland security and may reduce racial profiling," in *The Internet Law Journal*. vol. 2007, T. Kurtzman, Ed. Cleveland, OH: Case Western Reserve University, 2006.

[108] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall, 2003.

[109] J. E. Russo and P. J. H. Schoemaker, "Managing overconfidence," *Sloan Management Review,* vol. 33, pp. 7-28, 1992.

[110] J. E. Russo and P. J. H. Schoemaker, *Winning decisions: Getting it right the first time*. New York, NY: Doubleday, 2002.

[111] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research,* vol. 3, pp. 210-229, 1959.

[112] T. B. Sheridan and W. L. Verplank, "Human and computer control of undersea operators," MIT Man-Machine Systems Laboratory, Cambridge, MA 1978.

[113] M. S. Silver, "Decision support systems: Directed and non-directed change," *Information Systems Research,* vol. 1, pp. 47-70, 1990.

[114] H. A. Simon, "A behavioral model of rational choice," *Quarterly Journal of Economics,* vol. 69, pp. 99-118, 1955.

[115] H. A. Simon, *Administrative behavior*. New York, NY: Macmillian, 1957.

[116] H. A. Simon, *Models of my life*. New York, NY: Basic Books, 1991.

[117] L. J. Skitka, K. Mosier, and M. D. Burdick, "Accountability and automation bias," *International Journal of Human-Computer Studies,* vol. 52, pp. 701-717, 2000.

[118] L. J. Skitka, K. L. Mosier, and M. Burdick, "Does automation bias decision-making?," *International Journal of Human-Computer Studies,* vol. 51, pp. 991-1006, 1999.

[119] S. L. Sporer, "The less traveled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events," *Applied Cognitive Psychology,* vol. 11, pp. 373-397, 1997.

[120] R. H. Sprague, "A framework for the development of decision support systems," *MIS Quarterly,* vol. 4, 1980.

[121] R. H. Sprague and E. D. Carlson, *Building effective decision support systems*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1982.

[122] D. W. Straub, "Validating instruments in MIS research," *MIS Quarterly,* vol. 13, pp. 147-169, 1989.

[123] J. A. Swets, "Indices of discrimination or diagnostic accuracy: Their ROCs and implied models," *Psychological Bulletin,* vol. 99, pp. 301-340, 1986.

[124] J. A. Swets, "Enhancing diagnostic decisions," in *Judgment and Decision Making: An Interdisciplinary Reader*, T. Connolly, H. R. Arkes, and K. R. Hammond, Eds. Cambridge, England: Cambridge University Press, 2000.

[125] The Global Deception Research Team, "A world of lies," *Journal of Cross-Cultural Psychology,* vol. 37, pp. 60-74, 2006.

[126] R. G. Tippett, "A comparison between decision accuracy rates obtained using the polygraph instrument and the computer voice stress analyzer in the absence of jeopardy," 1993.

[127] P. Todd and I. Benbasat, "An experimental investigation of the impact of computer based decision aids on decision making strategies," *Information Systems Research,* vol. 2, pp. 87-115, 1991.

[128] P. Todd and I. Benbasat, "An experimental investigation of the computer based DSS on effort," *MIS Quarterly,* vol. 16, pp. 373-393, 1992.

[129] P. Todd and I. Benbasat, "The impact of information technology on decision making: A cognitive perspective," in *Framing the Domains of IT Management: Projecting the Future through the Past*, R. W. Zmud, Ed. Cincinnati, OH: Pinnaflex Educational Resources, Inc., 2000.

[130] E. Turban and P. R. Watkins, "Integrating expert systems and decision support systems," *MIS Quarterly,* vol. 10, pp. 121-136, 1986.

[131] A. Tversky, "Elimination by aspects, A theory of choice," *Psychological Review,* vol. 79, pp. 281-299, 1972.

[132] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science,* vol. 185, pp. 1124-1131, 1974.

[133] D. P. Twitchell, D. P. Biros, M. Adkins, N. Forsgren, J. K. Burgoon, and J. F. Nunamaker, "Automated determination of the veracity of interview statements from people of interest to an operational security force," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Big Island, HI, 2006.

[134] D. P. Twitchell, J. F. Nunamaker Jr., and J. K. Burgoon, "Using speech act profiling for deception detection," in *Lecture Notes in Computer Science 3073: Intelligence and Security Informatics: Proceedings of the Second Symposium on Intelligence and Security Informatics*, Tucson, AZ, 2004.

[135] D. P. Twitchell, J. F. Nunamaker Jr., and J. K. Burgoon, "Using speech act profiling for deception detection," in *Lecture Notes in Computer Science 3073: Intelligence and Security Informatics: Proceedings of the Second Symposium on Imtelligence and Security Informatics*, Tucson, AZ, 2004.

[136] U. Undeutsch, "The development of statement reality analysis," in *Credibility Assessment*, U. Undeutsch, Ed. Dordrecht, The Netherlands: Kluwer, 1989, pp. 101-121.

[137] A. Villringer, "Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults," *Neuroscience Letters,* vol. 154, p. 101, 1993.

[138] A. Vrij, "Credibility judgments of detectives: The impact of nonverbal behavior, social skills, and physical characteristics on impression formation," *Journal of Social Psychology,* vol. 133, pp. 601-611, 1993.

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

[140] A. Vrij, K. Edward, and R. Bull, "Police officer's ability to detect deceit: The benefits of indirect deception detection," *Legal and Criminological Psychology,* vol. 6, pp. 185-196, 2001.

[141] A. Vrij and S. Mann, "Telling and detecting lies in a high-stake situation: The case of a convicted murderer," *Applied Cognitive Psychology,* vol. 15, pp. 187-203, 2003.

[142] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: IV. Reliability, validity, and applications," *Perceptual and Motor Skills,* vol. 62, pp. 875-888, 1986.

[143] C. D. Wickens and S. R. Dixon, "Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature," Micro Analysis and Design, Boulder, CO 2005.

[144] I. E. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann, 2005.

[145] C. R. Wren, A. J. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 780-785, July 1997 1997.

[146] F. J. Yates, *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall, 1990.

[147] L. R. Ye and P. E. Johnson, "The impact of explanation facilities on user acceptance of expert systems advice," *MIS Quarterly,* vol. 19, pp. 157-172, 1995.

[148] L. Zhou, J. K. Burgoon, J. F. J. Nunamaker, and D. P. Twitchell, "Automated linguistics based cues for detecting deception in text-based asynchronous computer-mediated communication: An empirical investigation," *Group Decision and Negotiation,* vol. 13, pp. 81-106, January 2004 2004.

[149] L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, and J. F. Nunamaker Jr., "An Exploratory Study into Deception Detection in Text-Based Computer-Mediated Communication," in *Thirty-Sixth Annual Hawaii International Conference on System Sciences (CD/ROM)*, Big Island, Hawaii, 2003.

[150] L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, and J. F. Nunamaker Jr., "Toward the Automatic Prediction of Deception - An empirical comparison of classification methods," *Journal of Management Information Systems,* vol. 20, pp. 139-166, Spring 2004 2004.

[151] I. Zigurs and B. K. Buckland, "A theory of task/technology fit and group support systems effectiveness," *MIS Quarterly,* vol. 22, pp. 313-334, 1998.

[152] M. Zuckerman, B. M. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," in *Advances in experimental social psychology*. vol. 14, L. Berkowitz, Ed. New York, NY: Academic Press, 1981, pp. 1-59.

[153] M. Zuckerman and R. E. Driver, "Telling Lies: Verbal and Nonverbal Correlates of Deception," in *Multichannel Integrations of Nonverbal Behavior*, A. W. Siegman and S. Feldstein, Eds. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1985.