4-2016

Robustness of Multiple Indicators in Automated Screening Systems for Deception Detection

Nathan Twyman
Missouri University of Science and Technology

Jeffrey Gainer Proudfoot
Bentley University

Ryan M. Schuetzler
University of Nebraska at Omaha, ryan.schuetzler@byu.edu

Aaron Elkins
San Diego State University

Douglas C. Derrick
University of Nebraska at Omaha, dcderrick@unomaha.edu

Robustness of Multiple Indicators in Automated Screening Systems for Deception Detection

By:
Nathan W. Twyman, Jeffrey Gainer Proudfoot, Ryan M. Schuetzler,
Aaron C. Elkins & Douglas C. Derrick

Abstract

This study investigates the effectiveness of an automatic system for detection of deception by individuals with the use of multiple indicators of such potential deception. Deception detection research in the information systems discipline has postulated increased accuracy through a new class of screening systems that automatically conduct interviews and track multiple indicators of deception simultaneously. Understanding the robustness of this new class of systems and the limitations of its theoretical improved performance is important for refinement of the conceptual design. The design science proof-of-concept study presented here implemented and evaluated the robustness of these systems for automated screening for deception detection. A large experiment was used to evaluate the effectiveness of a constructed multiple-indicator system, both under normal conditions and with the presence of common types of countermeasures (mental and physical). The results shed light on the relative strength and robustness of various types of deception indicators within this new context. The findings further suggest the possibility of increased accuracy through the measurement of multiple indicators if classification algorithms can compensate for human attempts to counter effectiveness.

Key words and phrases

automated screening systems, deception countermeasures, deception detection, design science, human–computer interaction, human risk assessment, human screening

Traditional deception detection systems used for assessing human veracity have been criticized for lacking sufficient control, requiring extensive human skill, and being too invasive and costly for widespread application [51, 65, 89]. Recent work in the information systems (IS) domain has sought to address these issues directly through design and evaluation of automated deception detection systems. Automated approaches have shown promise for decreasing human skill requirements and minimizing interviewer effects, and have discovered behavioral and psychophysiological indicators of deception that can be measured using noninvasive methods. One of the assumptions driving this research has been that measuring many indicators of deception at the same time could improve the reliability of an assessment, as well as make the overall system more robust to *countermeasures*, or methods of deceiving the system [26, 89].

The potential for increasing accuracy and discovering robustness provides clear motivation for a study investigating multiple indicators of deception in an automated screening system. Analyses that use multiple indicators may more accurately capture an underlying psychophysiological or behavioral correlate of deception, or they may capture more than one correlate. Either option should theoretically

increase detection accuracy. However, some indicators of deception are more easily controlled than others, and before multiple-indicator automated screening systems can render a bigger impact, it will be important to understand and improve their robustness in the face of directed attempts to defeat them. This need is similar to that which prompted examinations of the robustness of more invasive systems such as the polygraph (e.g., [47,48]) and brain imaging techniques (e.g., [38, 81]), which identified susceptibility to human countermeasures that needed be addressed.

Of particular import to both theory and practice is an increased understanding of how easily certain indicators of deception can be controlled or masked by the deceiver. Prior knowledge about the task and the indicators used to evaluate veracity allows an interviewee to change his or her behavior in an effort to appear innocent. Some behaviors appear to be more difficult to control than others [19], and an understanding of how different behavioral manipulations affect the robustness of an indicator can reveal further insight into the mechanisms that generate the indicator. Thus, different types of countermeasures and different types of indicators must be measured to effectively evaluate an automated screening system.

This paper details a study designed to evaluate the performance of a human screening deception detection system using multiple sensors in normal conditions and when various types of countermeasures are employed.

Background on Automated Deception Detection Systems

Technology has been used to aid deception detection since at least 1895, when Cesare Lombroso, an Italian criminologist, used a medical device for measuring blood pressure changes during police interrogations [87]. In the 1920s and 1930s, John Larson and Leonarde Keeler developed the now widely known polygraph machine that measures blood pressure, respiration, and skin conductance (a measure of arousal) [1]. Polygraph systems have since become the standard tool for aiding in detecting deception, despite a general scientific consensus that the most common polygraph interviewing technique lacks validity [51]. Accordingly, much academic research on deception detection focuses on alternative techniques that feature more scientific control [60, 61, 79]. The current study likewise frames contributions within the context of a highly controlled interviewing technique, as detailed in previous IS research [89].

Technology and systems clearly have potential to improve deception detection accuracy. Unaided human deception detection accuracy rates hover near chance levels [9]. Furthermore, the interviewing process for detecting deception is lengthy and requires specialized sensors configured by highly skilled interviewers. Each of these factors prevents broad application. Opportunities for improvement in this area include identification of new deception indicators and measurement tools, decision support systems, or fully automated deception detection systems. Recently, additional technologies for human screening have been investigated, including noncontact technologies for measuring heart rate and blood pressure [68], vocalic features [37, 44], linguistic variables [35, 92], oculometric factors [82, 84], thermal features [73, 74], and kinesic factors [17, 88]. Beyond individual technologies, at least two automated system designs have been proposed to cover the interview protocol, technologies, and interaction facilitation [68, 89].

To date, automated screening systems have generally focused on identifying indicators of deception. The combination of several indicators and an analysis of the robustness of indicators in this new context

have been left for future research. Multiple disparate indicators may provide a more holistic view of a deceptive state. A combination of indicators may also provide robustness if a deceiver focuses on suppressing or manipulating only a single type of indicator [19], or if the countermeasures employed are more effective against only one type of indicator. For instance, in some contexts, the mental countermeasure tactic of story recall is effective at reducing electrodermal responses but is not effective against the respiration response associated with deception [7]. Evidence such as this suggests that automated screening systems may benefit from multiple technologies measuring disparate indicators.

Deception Detection

Deception is a deliberate attempt to foster false beliefs or perceptions in the recipient [56]. Theories describing deception and its effects explain and predict human behavioral and physiological differences when a truth is conveyed as compared to a deception. These behavioral and physiological differences are sometimes referred to as *indicators* of deception. To the extent that these indicators are reliable and discoverable, detecting deception should be possible.

Several social psychology and communication theories of deception have built on Ekman and Friesen's [27] leakage hypothesis. The leakage hypothesis predicted that liars would experience abnormal arousal and affect, which would unintentionally "leak" deception indicators that would manifest particularly in the hands, legs, and feet. For instance, liars may relieve their tension and discomfort through nervous movements and adaptors (e.g., foot tapping, face touching). In addition to arousal-induced behaviors, the leakage hypothesis predicted that liars would also inhibit certain behaviors and natural gesturing. Zuckerman et al.'s [95] four-factor theory added increased cognitive effort and overt behavioral control as sources of leakage. Liars are predicted not only to exhibit leakage indicators due to arousal and negative affect but also to experience more cognitive load while managing their lie and appearance. This increased vigilance in appearance causes overcontrol of normally automatic and natural gesturing, resulting in rigid and inhibited behavior.

The type and strength of indicators of deception may change throughout the course of an interaction. Interpersonal deception theory [15] posits that deceivers will employ dynamic strategies in their effort to appear credible. The type and amount of variation depend partially on the skill of the sender and his or her relationship with the receiver (e.g., boss, parent, loved one). A liar may start out feeling confident and reveal few behavioral indicators of deception, but after sensing suspicion he or she may begin compensating by exhibiting different behaviors, and therefore different indicators.

While leaked and strategic behavioral indicators of deception are particularly interesting in common interpersonal interactions, such as interviews or conversations, formal deception assessments using the polygraph and similar tools traditionally rely on cognitive and psychophysiological indicators [36, 63]. These practices draw on theories describing the psychophysiological orienting of attention and the defensive response (traditionally termed a "fight or flight" response [10]) that are expected to occur when a deceiver is presented with a stimulus (e.g., a question or an image) that is perceived as personally significant and threatening. When attention orients toward a personally significant stimulus [18], measurable physiological changes occur, such as changes in pupil dilation [11, 59] and skin conductance [6, 8]. If that stimulus is perceived as threatening, a defensive response also occurs, which also exhibits psychophysiological variations and can also include defensive behaviors such as those identified in communication and social psychology literature [18, 90].

Countermeasures

Because deception detection is based on the measurement and interpretation of behavioral or psychophysiological responses, the employment of countermeasures to appear innocent is a threat to the validity of any deception detection system. Countermeasures have been shown to have a significant effect on polygraph tests [7, 46, 47, 49] and brain imaging tests [38, 81, 85]. Traditional polygraph countermeasures are characterized primarily by overt attempts to manipulate behavioral or physiological signals in a manner that is expected to minimize the difference between baseline (i.e., truthful) responses and responses to questions or other stimuli that may result in deception. Countermeasures fall primarily into two categories: physical and mental.

Physical countermeasures are deployed using a variety of behaviors, including finger movements [38], pressing toes against the floor [46, 47, 49], and biting the tongue [47, 49]. These physical countermeasures are employed during the portion of the interview that is designed to capture baseline physiology. By mimicking a physiological response (e.g., generating pain induces arousal) during baseline stimuli, examinees may effectively obfuscate their deception if evaluations reveal no significant difference between baseline and deceptive responses.

Mental countermeasures can also be used to mimic physiological responses, but more often they are used to suppress such responses through distraction. Mental countermeasures include silent counting [31, 46], recalling past emotional events [7], or distractions, such as mentally reciting one's own first and last name [85]. Mental countermeasures are employed either during the baseline portion if the goal is to mimic, or during the entire interview if the goal is to suppress. In a polygraph study, Elaad and Ben-Shakhar [31] found that counting sheep throughout the length of the deception test decreased detection rates. In a P300-based deception detection experiment, Sokolovsky et al. [85] showed that participants who mentally recited their first and last name during two of the four baseline stimuli were able to modify their responses enough to evade detection.

Where countermeasures are shown to be effective at manipulating deception detection results, cognitive psychology research has turned to detecting countermeasures. Physical countermeasures are especially vulnerable to detection. Honts et al. [49] showed that 90 percent of countermeasure users could be identified using electromyography to measure muscle activity in the legs and head. Similarly, Ganis et al. [38] showed that while slight finger movements reduced detection accuracy in fMRI-based tests, the movements also increased activation of the motor cortex, the part of the brain responsible for movement. Increases in reaction time allowed countermeasure detection in P300 mental countermeasures [80].

Deception Indicators in a Controlled Human Screening System

The study of deception and indicators of deception has extended for decades [27]. Throughout the years, a variety of indicators have been examined and reexamined to assess their capabilities for detecting deception. Early systems, such as the traditional polygraph, use physiological measures such as pulse, breathing, and electrodermal activity to identify deceptive responses when examinees are presented with *target* stimuli (i.e., deception-relevant questions) [1]. Responses to target stimuli are then compared to responses to baseline stimuli. These systems are restricted in their potential application, partly because the necessary sensors require extensive skill to interpret and require considerable time to connect and calibrate to the examinee. Newer studies of the neural correlates of

deception have employed fMRI [38] or EEG [85] systems to identify cognitive correlates of deception, but these tools require similar or sometimes greater amounts of time, training, and calibration. Automated screening systems have focused on indicators that can be captured using sensors that require minimal or no contact and calibration, such as ordinary and specialized cameras, platforms, and microphones [64, 68, 88, 89]. Such systems can operate with greater autonomy, thereby creating potential application to use cases where other deception detection systems are not feasible—cases such as health screening, job interviews, visa applications, financial audits, and more.

Whereas traditional deception detection tools measure one or two distinct indicators, next-generation screening systems may show increased resilience to countermeasures if they capture multiple types of signals from multiple underlying behavioral and psychophysiological processes. To the extent that individuals are limited in the number of activities to which they can be simultaneously attentive, they should be less able to counter a deception detection system that tracks and measures multiple heterogeneous indicators of deception. Many potential indicators have been identified in deception literature [25], but only three indicators that can be captured using noninvasive sensors—pupil dilation, kinesic rigidity, and gaze aversion—have been examined in the context of a highly controlled screening system. The current study examines these three confirmatory indicators in the same experiment and also includes three additional indicators that are exploratory, in that they have not been previously examined in this specific context. The exploratory indicators—vocal pitch, proximity, and frowning—were selected due to their heterogeneity and their feasibility for use in automated screening. Each of these six indicators of deception theoretically captures some distinct correlate of deception, and therefore may be differentially affected by physical and mental countermeasures. For example, pupil dilation occurs as a function of the orienting reflex in this context [89], while variation in vocal pitch stems from the emotional state of the deceiver [83]. By assessing deception from a variety of angles, a system may become more robust to intentional manipulation. Table 1 provides a brief description of each indicator and includes an overview of the effects of both deception and countermeasures on each indicator. Each indicator is discussed in more detail below.

Table 1. Summary of Confirmatory and Exploratory Indicators

Indicators		Predicted deception effect	Expected effects of countermeasures
Confirmatory indicators	Pupil dilation: Increased ocular pupil diameter	Pupils will dilate compared to baseline as a result of the orienting reflex	Mental: No significant effect Physical: Inhibited detection as a result of lesser disparity between baseline and target orienting reflex
	Kinesic rigidity: Constriction of body movement	Body movement will lessen compared to baseline as a result of the freeze response	No significant effects
	Gaze aversion: Oculomotor avoidance of a stimulus	Gaze will fixate more in the center of the screen (away from stimuli) compared to baseline as a defensive response	No significant effects
Exploratory indicators	Vocal pitch: Variations in vocal fold vibration rates	Vocal pitch will increase compared to baseline as a result of increased emotional arousal	Uncertain, but mental countermeasures may decrease arousal disparity, inhibiting detection
	Proximity: Physical distance between individual and target	Proximity will decrease compared to baseline as a result of increased cognitive interest	Uncertain; Mental countermeasures may diminish cognitive interest, inhibiting detection
	Frowning: Facial expression reflecting negative affect	Frowning will increase compared to baseline as a result of increased negative affect	Uncertain; Mental countermeasures may diminish negative affect, while physical countermeasures may increase negative affect

Pupil Dilation

Changes in pupil dilation can be linked with a number of cognitive functions. Early research provided evidence that changes in pupil dilation are linked with short-term and long-term memory retrieval [4, 39, 52] and can also indicate activation and arousal in autonomic activity [40, 70]. Controlled deception screenings with simple responses that require minimal recall are optimally suited for capturing dilation resulting only from autonomic physiological changes. In these settings, pupil dilation occurs as part of the autonomic *orienting reflex*, a psychophysiological response to novel or personally significant stimuli, as compared to a given baseline stimuli set [11, 53, 57, 59].

The orienting response has traditionally been a key human factor of interest in highly controlled deception detection interviews such as the concealed information test [2, 62]. Tracking differential

electrodermal activity (i.e., skin sweatiness) compared to an established within-subject baseline is the standard method for measuring the orienting response. However, the physiological activation triggered by the orienting response also triggers several other physiological changes, including pupil dilation. Within the context of a scientifically controlled interview, a common mechanism is thought to trigger both pupil dilation and increased electrodermal activation in response to presentation of a stimulus (either verbally or visually on a screen [57]) that represents purposely concealed knowledge [11]. In this context, at least a portion of pupil dilation is attributable to perception of the stimulus representing concealed knowledge rather than the verbal act of lying [53]. The system design used in the current study extends the common method of visually presenting several baseline stimuli on a screen (representations that are not relevant to the illicit activity in question) together with relevant stimuli. This study anticipated results similar to previous work that noted increased pupil dilation during presentation of relevant stimuli compared to baseline stimuli [59]. Hypothesis 1:

Deceptive individuals will have a larger pupillary response to target stimuli than to baseline stimuli.

The employment of mental countermeasures to artificially increase pupillary response could help artificially raise the baseline of comparison, resulting in a less accurate determination of deception. Because pupil dilation is also related to cognitive processing [54], taxing mental tasks may increase pupil dilation on demand. However, because examinees must take plausible note of stimuli and respond as they are employing the countermeasure, dilation should still increase beyond the baseline. Thus, it is unlikely that mental countermeasures will distract enough to prevent generating a detectable autonomic orienting reflex. Hypothesis 2:

Deceptive individuals using mental countermeasures will exhibit increased pupil dilation during responses to target stimuli.

In addition to cognitive effort, pupil dilation can also be triggered by pain. Using electrical stimulation, Chapman et al. [20] showed that pupil dilation increases nearly immediately at the onset of pain, and that this dilation increases with increasing pain intensity. Further research has shown that pupil dilation not only is immediate but also lasts for the duration of the pain [34]. These results indicate that physical countermeasures such as biting the tongue, as employed in polygraph studies, may also work to artificially manipulate the pupil dilation baseline in an automated screening paradigm. Inducing pain during baseline stimuli should increase pupil dilation, reducing the difference between target and baseline pupil dilation. Hypothesis 3:

Physical countermeasures will reduce the pupil dilation differential between target and baseline stimuli.

Kinesic Rigidity

Kinesic rigidity is the constriction of body movement. In communication research, rigidity has been found to be an indicator of low veracity during open-ended or semistructured interviewing techniques [19, 93]. When lying, participants tend to exhibit less overall movement, especially expressive or illustrative gestures, and the movement that does occur tends to be spatially constricted and appears forced rather than natural [14, 93, 94]. Recent IS research has discovered that this phenomenon is also present in the more controlled interview setting and has developed a method for automatic detection of rigidity via comparison of body movement during baseline stimuli to body movement during target stimuli [88].

Likely because of the high cost of traditional measurement of rigidity, this deception indicator is not commonly used in practice. It has also received almost no attention in countermeasures research. One psychology study determined that controlling rigidity is very difficult in semistructured interviews, at least when trying to control it directly [94]. However, there are still many unknowns, including how effective traditional countermeasures are and how well rigidity can be overtly controlled in a setting in which many behaviors must be controlled simultaneously.

In semistructured interviews, rigidity has been hypothesized to stem from high cognitive load. The explanation is that greater-than-normal cognitive effort is placed on mentally constructing and relaying a plausible story, leaving fewer cognitive resources for nonverbal presentation. This lack of resources is thought to lead to less overall movement and more constricted movement [24, 28]. A second theory suggests that rigidity itself may be an unconscious or overt countermeasure, in that because people generally falsely believe that liars exhibit increased movement, liars minimize their own movement to appear truthful [23]. A third possible explanation is the biologically driven freeze response that all humans experience when confronted with a stimulus that is perceived as threatening [41]. Previous IS research discovered rigidity in a controlled interview that required no communicative or illustrative movement, so cognitive overload will not be a likely driver in the highly controlled, automated format employed by the system design used in the current study.

The effectiveness of traditional countermeasures on kinesic rigidity is therefore expected to be low. Physical countermeasures do not eliminate the perception of a threat or the desire to appear truthful. Mental countermeasures should be effective to the extent that they not only cognitively distract the examinee but also help him or her to ignore the potential threat or the desire to appear truthful. However, because individuals have to at least verbally respond and be visually attentive, mental countermeasures are limited in how much they can distract, leaving plenty of opportunity for recognition of the potential threat and triggering a natural freeze response. More research is needed to fully understand these dynamics, but at this stage it appears unlikely that traditional countermeasures will be particularly effective against the rigidity effect. Hypothesis 4:

Deceptive individuals will exhibit less overall movement when viewing and responding to a target stimulus.

Hypothesis 5:

Deceptive individuals employing mental or physical countermeasures will exhibit less overall movement when viewing and responding to a target stimulus.

Gaze Aversion

Earlier research supports the hypothesis that when presented with several equidistant stimuli on a single screen during a visual multiple-answer question set, individuals concealing guilt have a tendency to spend more time gazing away from all stimuli by spending more time looking at the center of the screen. This effect happens if one of the stimuli is highly associated with the guilt being concealed [89]. This study replicates the same hypothesis. Hypothesis 6:

Deceptive individuals will show increased gaze time at the screen center when presented with a target stimulus.

This gazing tendency may stem from an autonomic avoidance response, or it may be an overt defensive behavior designed to help avoid suspicion. If an overt defensive behavior, traditional countermeasures designed to mentally distract or corrupt physiological responses should not naturally translate into controlling visual gaze. Hypothesis 7:

Deceptive individuals employing mental or physical countermeasures will exhibit greater gaze duration at the screen center when viewing and responding to a target stimulus.

Vocal Pitch

Whereas the above-named correlates of deception stem from autonomic psychophysiological processes and overt defensive behavior, vocal pitch is thought to correlate more with emotional arousal. To speak, the diaphragm pushes air through vocal folds in the larynx [86]. The frequency of the air pressure changes affected by vibration of the vocal folds is perceived as vocal pitch. The vocal fold vibrations are facilitated by muscles about the larynx in the vocal tract. Just like other muscles in the body, the larynx muscles exhibit tension when an individual experiences stress or arousal. Tension around the larynx causes vocal folds to increase the frequency of vibration, thereby increasing vocal pitch.

Increases in mean and range in vocal pitch have been predictive of deceptive speech [25, 33, 77] and heightened emotion and arousal [3, 83]. Because vocal pitch provides primarily emotional arousal-based information, it has not been analyzed in the type of highly controlled interview implemented in this study, as studies using these controlled interviews traditionally rely on psychophysiological measurements. However, emotional arousal may be present in a controlled interview, even though it has not traditionally been measured. Hypothesis 8:

Deceptive individuals will exhibit greater vocal pitch when responding to a target stimulus.

Vocal pitch and the next two indicators, proximity and frowning, are in an exploratory stage for this context, so discussion of countermeasure effectiveness is more speculative. Nevertheless, mental countermeasures may be effective against vocal pitch to the extent that they are able to distract enough to diminish emotional reaction. Physical countermeasures such as stepping on a tack or even biting one's tongue may or may not ultimately lead to tension in the larynx muscles, so their effectiveness is uncertain.

Proximity

Interpersonal proximity—the physical distance between two social actors—has been hypothesized to increase with deception in interpersonal communication [16]. This view assumes that proximity is a type of nonverbal *immediacy*, which is the degree to which a communication is direct, relevant, clear, and personal. Nevertheless, insufficient evidence supports the relationship between deception and proximity [25]. It is possible that proximity is highly influenced by many other factors beyond immediacy, limiting its potential as an indicator in unstructured communications. Furthermore, in an automated screening interaction in which allowed responses are highly restricted, immediacy is not likely to vary, which may suggest little potential for exploring this indicator in this context.

However, in one relevant study that used a structured interaction and a virtual screening agent, proximity significantly *decreased* among deceivers [72]. That study more closely matches the current context because it used an automated interview. It is possible that proximity decreased as a result of

increased cognitive interest: attention may focus in more on a stimulus that is of higher interest compared to baseline stimuli. This interest factor may naturally draw examinees in closer to the target perceived to be more interesting. If interest is the key construct, mental countermeasures would be only as effective as their ability to distract from interesting stimuli. Hypothesis 9:

Deceptive individuals will move closer to a target stimulus.

Frowning

Regarding facial expressions, "leakage" of emotional indicators is the dominant explanation for deception indicators. When the act of lying and/or perceptions of guilt generate negative affect, these emotions have a natural tendency to show up in the face. Some evidence supports the notion that controlling facial expressions can be a difficult venture [13, 50, 76], though hiding less-intense emotions may be easier than hiding very intense emotions [75]. Since expressions are emotionally linked, mental countermeasures may have potential to distract inasmuch as they diminish emotional response. Physical countermeasures may amplify negative affect in baseline responses.

Little or no research has investigated leaked emotion in facial expressions in a controlled, automated screening interview, in which no natural, dynamic conversation occurs. In such a setting, deceivers' emotions should remain constant except when presented with target stimuli or when increased negative affect such as guilt or fear could lead to a more negative expression. Hypothesis 10:

Deceptive individuals will increase frowning when presented with a target stimulus.

Determining whether individuals can be successful when they attempt to counter many factors at once is especially important to this area of research. Thus, in addition to testing these ten hypotheses, this study also investigates whether countermeasures are less effective when multiple countermeasures are employed simultaneously.

Research Approach

This study is one large step within a design science-driven program of research centered on human risk assessment systems. Using this research framework, knowledge is discovered through the design, implementation, and evaluation of system prototypes [42, 45, 66, 67, 69]. Projects iterate through prototyping, experimentation, field studies, and theoretical development in a nonlinear path. Because of the multimethodological nature of the research approach, contributions may be as diverse as revised understanding of a problem space, new theoretical insights, or evaluation of a proposed system design. The goals of this study were not only to evaluate overall system performance but also to test hypothesized outcomes. Thus, the research approach is both design science and behavioral in nature.

Because of the relative novelty of the problem space, a laboratory experiment was an appropriate method of investigation at this stage. An experiment was designed to evaluate the ability of deceivers to successfully bypass an automated screening system through the use of countermeasures. The experimental task was patterned after a number of experiments designed to test the ability of noninvasive sensors to identify deception and concealed information. The experiment was composed of five conditions, including (1) guilty with no countermeasures; (2) guilty with mental countermeasures; (3) guilty with physical countermeasures; (4) guilty with mental, physical, and additional countermeasures; and (5) an innocent (control) group. Measures were repeated within-subjects and

within-question sets for a total of 20 measurements (captured during responses to 20 questions) per individual, per indicator. Detailed information about the data collection process and the experimental task is provided in the following subsections.

Participants

Participants were recruited from undergraduate and graduate business courses at a large American university. Human subjects review approval was obtained, and all human subjects procedures were followed. Although the ideal population would be individuals who regularly participate in illicit activities, such a population was not feasibly obtainable. Students were therefore selected as a target population due to (1) the cultural and ethnic diversity of student populations and (2) the empirically supported similarities between physiological response patterns exhibited by interviewees in both field and laboratory controlled interviews [30, 32, 62, 71, 91]. Participants' (*N* = 175) mean age was 21.86, with a median age of 21, a minimum age of 18, and a maximum age of 36. The sample was composed of individuals with a diverse range of ethnic backgrounds, including: 33 percent Asian, 3 percent black, 1 percent Hawaiian/Islander, 46 percent white, 13 percent Hispanic, 3 percent Middle Eastern, and 1 percent unidentified. Females accounted for 41 percent of the total participants. Over half (126) of the participants spoke English as a second language. Of the initial 175 participants, 18 were disqualified because they either failed two manipulation check questions,[1] or they did not follow instructions (i.e., failed to complete the task as instructed). Participants received course credit for participation, as well as $20 if they successfully made it through the checkpoint and completed their task without arousing suspicion. The requirements to avoid detection and to complete a task after the screening were a mild deception to encourage realism. No participants were required to complete their task once they passed beyond the screening.

Experiment Task

Each participant was required to provide written consent prior to beginning the experiment. The experimental task mirrored that of Twyman et al. [89], except that it included additional conditions (summarized in Table 2). After providing consent, each individual was randomly assigned to one of the five conditions. The number of participants assigned to each group was roughly equal.

Table 2. Experiment Conditions

Condition		Task
Innocent	Innocent	Pack a bag with innocuous items and bring it through screening into a secure portion of the building.
Guilty	Guilty, no countermeasures	Pack a bag with innocuous items and construct and pack an IED. Bring the bag through screening into a secure portion of the building.
	Guilty, mental countermeasure	Pack a bag with innocuous items and construct and pack an IED. Bring the bag through screening into a secure portion of the building. Use a mental countermeasure to thwart the screening system.
	Guilty, physical countermeasure	Pack a bag with innocuous items and construct and pack an IED. Bring the bag through screening into a secure portion of the building. Use a physical countermeasure to thwart the screening system.
	Guilty, many countermeasures	Pack a bag with innocuous items and construct and pack an IED. Bring the bag through screening into a secure portion of the building. Use mental, physical, and additional countermeasures simultaneously to thwart the screening system.

Participants in the guilty conditions were tasked with constructing an improvised explosive device (IED), concealing it in a bag with standard travel items, and delivering the bag to a person located in a secure portion of a building (see Figure 1). Participants in the three countermeasures groups also completed training on how to use their assigned countermeasure(s) to defeat the system. Members of the control group packed only innocuous household items in a bag prior to proceeding to the simulated screening; they did not receive any information about the system or countermeasures training. After packing the bag, they proceeded to a simulated screening environment, purportedly for security screening prior to entry.

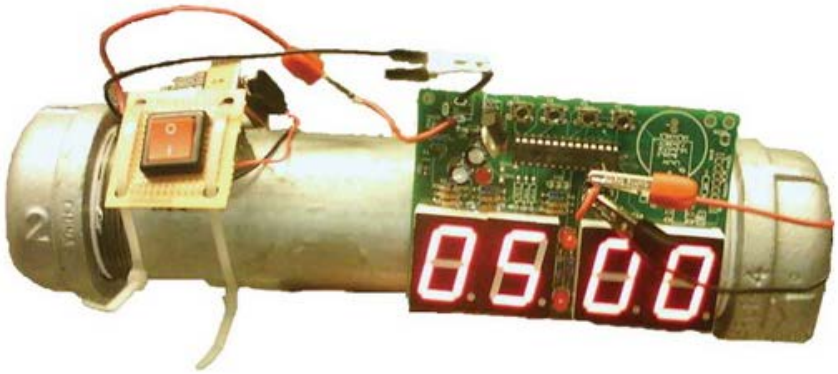

Figure 1. Improvised Explosive Device Built by Participants in Guilty Conditions

Countermeasures training entailed watching a video describing the sensors that would be used during the screening interview to identify deception, as well as methods that could be used to thwart the effectiveness of the sensors. The countermeasures selected for use in this experiment were adapted from the most commonly employed countermeasures used to thwart the polygraph exam. Participants in the mental countermeasures condition were instructed to count backward by seven from a three-digit number. This tactic is designed to occupy the interviewee's mental processes, thereby cognitively blocking out most of the screening interview that is taking place and minimizing any behavioral or physiological differences in behavior that could occur. This countermeasure has previously demonstrated its efficacy for polygraph examinations [48]. Participants in the physical countermeasures condition were instructed to moderately but carefully bite their tongue when baseline stimuli were presented during the interview, in an attempt to create physiological responses on baseline stimuli that could mimic the expected response of a guilty person to a stimulus of interest. This action could theoretically result in a false negative classification, as the guilty participant should appear as much or more responsive when confronted with control stimuli than when confronted with target stimuli of interest, and would thus appear innocent [47, 48]. The final countermeasures group was trained to use the mental and physical countermeasures described previously, as well as to directly control kinesic rigidity (not "freeze up"), control their vocal pitch such that it would always be the same, keep a neutral facial expression, and look at all images naturally—not focus more on one stimulus on the screen than another.

All participants entered the screening environment, placed the packed bag on a counter, and stood in front of the automated screening system shown in Figure 2. The system was designed to identify the presence of an interviewee and begin the interview autonomously. The first step of the interview entailed automatic simple calibration of the interviewee to the eye-tracking device. During this brief process, the person must fixate his or her eye gaze on nine different locations on the screen, resulting in a more accurate and consistent oculometric data collection than if no calibration were obtained. After calibration, the screening interview commenced, which entailed displaying to the examinee five slides containing images of banned items. Each slide contained four small images of banned items (example shown in Figure 3), and an embodied conversational agent asked the interviewee, "Are you carrying any of these items?" immediately prior to the display of each slide. The interviewee was given five seconds to respond before the question repeated and the next slide appeared. Participants were asked to vocally respond either "yes" or "no." One of the five slides randomly contained the target stimulus, which was an image that resembled an explosive device. Participants in the guilty conditions had to verbally lie in response to this question to appear truthful. The five-slide set was repeated four times, with the slide order and the spatial location of the images changed each time. In total, each participant responded to 20 questions, the response to 4 of which required an outright lie.

Figure 2. Automated Screening Kiosk Used to Question Participants

Figure 3. Example Slide Used in the Automated Screening

After the screening, each participant completed a brief survey designed to check experimental manipulations and to collect information on his or her experience during the interview.

Measures

The sensors used in this experiment were selected based on their potential for noncontact, automated interviewing. Kinesic data were captured using a Microsoft Kinect sensor, which captured three-dimensional Cartesian coordinates of 20 distinct body points in real time at approximately 30 Hz. These raw data were used to calculate overall movement via the same procedure detailed in previous work on automated rigidity detection [88]. These raw data were averaged by participant (N = 157) and response (N = 20) for each body point. Proximity was measured using the same data and process, except the raw data used for input was distance from the sensor for each major body point.

Pupil diameter was measured via an EyeTech^{TM} VT2 eye-tracking sensor. The VT2 captured pupil diameter at approximately 50 Hz. The mean pupil diameter for both eyes was calculated for each measurement; these data were likewise summarized by participant and response. Center-gaze-duration ratio was calculated using Cartesian coordinate data collected by the VT2. Percentage of time viewing the center of the screen was likewise calculated for each slide-participant combination.

Raw vocal data were captured using an array microphone capturing at 48 kHz. For each response, the maximum, mean, and standard deviation of vocal pitch from the beginning to the end of each vocal utterance (always the word "no") were extracted from the raw vocal data. Open-ended responses with a diversity of possible sounds can greatly influence pitch measurements between utterances. Having all vocal responses restricted to one-word, uniform responses in this study increased comparability between participants.

Frowning was measured by analyzing video captured at 15 fps by a standard high-definition webcam. Frown data were generated from the videos using CERT [58], which generates the level of smile (or frown) for each video frame using an algorithm trained on a database of diverse images of faces.

To control for effects stemming from highly variable interpersonal differences, such as wide variance in nervousness, stillness, eye size, and vocal range, the data points from each of these indicators were standardized using within-subject *z*-scores [5], meaning each subject's observations were representative of a personal baseline as opposed to a population baseline. All observations were also standardized within each group of slides to take advantage of the question-specific baseline [89]. In the case of body movement, movement was also standardized for each body point separately to account for natural differences in movement patterns between body points.

For each sensor and indicator, some cases of data loss occurred due to misconfiguration, difficulty with calibration, or low-fidelity data capture. For instance, technical malfunctions with the raw vocal data sensor occurred on 5.6 percent of the measurements. After removing cases with missing data, the total number of usable data points was 2,172 for each indicator except body movement, which had 3,139 measurements for each of the 20 body points (total N = 62,780).

Analysis and Results

Because any of several measures of vocal pitch variation may have been most useful in a controlled screening context, vocal pitch variation underwent a preliminary analysis to explore three possible measures. Regression analyses were performed to examine the relative effectiveness and robustness of the confirmatory and exploratory indicators.

The vocal measures investigated included mean pitch, pitch standard deviation, and max pitch. As with body movement and pupil dilation indicators, each of these variables was normalized within-subjects and within each group of slides before being submitted to repeated measures ANOVA (Condition x Target Stimulus). The interaction of Condition and Target Stimulus was not significant for mean pitch (p = .28) or pitch standard deviation (p = .065). Max pitch, a measurement of high-end pitch, was significant for the Condition and Target interaction, $F(4, 2989) = 2.32$, p = .05.

Pupil dilation, center gaze duration, vocal pitch, proximity, and frowning were specified as dependent variables in a multivariate regression model. The dependent variables are standardized scores representing standard deviations from an individual baseline. The independent variables include Target Stimulus (a binary variable indicating whether the stimuli slide included an IED image), Time (a value between 1 and 4 representing the temporal order of the four groups of stimuli slides), and Condition (a dummy coded variable representing the four guilty conditions using the Innocent condition as the baseline for comparison). Interaction effects between Condition and Target Stimulus were the key independent variables of interest, necessary to test hypotheses.

Whereas other indicators were measured with a single signal, overall body movement involved measurement of 20 distinct body points. No clear best method for aggregating or filtering these data has yet been identified for this context. Therefore, a separate regression model was specified with movement as the sole dependent variable. All other model specifications were the same as the multivariate model. English as a second language was initially included as a covariate in both models, but was not statistically significant and was subsequently removed. A Bonferroni correction was applied to account for the use of multiple models. The body movement model was statistically significant, $F(10, 62769) = 21$, $p < .001$. The multivariate model showed overall significance of Target Stimulus, $F(4,20) = 26.792$, $p < .001$, and the Condition-Target Stimulus interaction, $F(4, 20) = 4.684$, $p < .001$. No significant main effects of Condition or Time were found. A post hoc power analysis revealed that with this sample, small effect sizes (.004) would be detectable. Observed effect sizes (.058 and .042, respectively) exceeded this threshold.

Indicator-specific results of the models are shown in Table 3. For convenience, overall movement results are displayed in the same table as the multivariate model results. Beta weights can be interpreted in terms of standard deviations—the amount of difference from normal for an individual. For instance, when individuals in the guilty condition responded to a question involving a target stimulus, their frown was on average 0.442 standard deviations deeper than when they were responding to questions not involving a target stimulus.

Table 3. Regression Results

	Pupil diameter	Overall movement	Center gaze duration	Max vocal pitch	Proximity	Frown
Fixed effects	β (S. E.)	β (S. E.)	β (S. E.)	β (S. E.)	β (S. E.)	β (S. E.)
(Intercept)	0.001	0.005	0.056	0.075	0.007	0.013
	(0.060)	(0.013)	(0.061)	(0.062)	(0.070)	(0.062)
Target Stimulus	−0.072	−0.026	**−0.251***	**−0.278****	0.217	−0.078
	(0.010)	(0.019)	(0.101)	(0.103)	(0.116)	(0.103)
Time	0.002	0.000	0.001	−0.005	−0.005	0.003
	(0.016)	(0.003)	(0.017)	(0.017)	(0.019)	(0.015)
Guilt	−0.105	0.015	−0.065	−0.065	0.031	−0.091
	(0.061)	(0.012)	(0.062)	(0.063)	(0.071)	(0.063)
Mental Countermeasure (MC)	**−0.179****	0.019	−0.046	−0.043	0.045	0.064
	(0.068)	(0.013)	(0.069)	(0.071)	(0.079)	(0.071)
Physical Countermeasure (PC)	**−0.146***	**0.041****	−0.074	−0.089	0.007	−0.008
	(0.064)	(0.013)	(0.065)	(0.066)	(0.074)	(0.066)
All Countermeasures (AC)	**−0.199****	0.013	−0.053	−0.044	−0.002	−0.050
	(0.064)	(0.012)	(0.065)	(0.067)	(0.075)	(0.067)
Guilt x Target Stimulus	**0.609*****	**−0.076****	0.222	**0.281***	**−0.377***	**0.442****
	(0.138)	(0.027)	(0.140)	(0.143)	(0.160)	(0.143)
MC x Target Stimulus	**0.892*****	**−0.097*****	0.166	0.147	**−0.398***	−0.246
	(0.157)	(0.028)	(0.159)	(0.162)	(0.182)	(0.162)
PC x Target Stimulus	**0.702*****	**−0.203*****	**0.343***	**0.375****	−0.301	−0.042
	(0.146)	(0.028)	(0.148)	(0.151)	(0.169)	(0.151)
AC x Target Stimulus	**0.922*****	**−0.063***	0.257	0.154	−0.130	0.109
	(0.145)	(0.028)	(0.147)	(0.150)	(0.168)	(0.150)

* $p < .05$; ** $p < .01$; *** $p < .001$; models fit using maximum likelihood. Statistically significant results are in bold.

Confirmatory Indicators

When the target stimulus (IED image) was displayed on the screen while an examinee responded to a question, pupil dilation was significantly larger for participants in the guilty (β = .609, $p < .001$), mental countermeasures (β = .892, $p < .001$), physical countermeasures (β = .702, $p < .001$), and all countermeasures (β = .922, $p < .001$) groups. The all countermeasures group showed the largest effect.

Kinesic rigidity was detected among all groups who were smuggling the IED. When the target stimulus was present on the screen, guilty (β = –.076, $p = .006$), physical countermeasures (β = –.203, $p < .001$), mental countermeasures (β = –.097, $p = .001$), and all countermeasures (β = –.063, $p = .024$) groups exhibited rigidity. Those performing only physical countermeasures showed the greatest amount of rigidity.

Though increased center-of-the-screen gazing was consistent for all guilty conditions, this was not a significant indicator of deception in this study except in the physical countermeasures group (β =.343, $p < .05$).

Exploratory Indicators

Max vocal pitch was used as the dependent variable for vocal pitch variation in the multilevel regression model detailed in Table 3. Individuals in both the guilty without countermeasures (β = .281, $p < .05$) and the physical countermeasures (β = .375,$p < .05$) conditions demonstrated significant increases in max pitch when responding to target stimuli. Though max pitch among participants in the other two guilty conditions was calculated to be higher than baseline, neither the mental countermeasures condition (β = .147, $p = .37$), nor the condition using several countermeasures (β = .154, $p = .31$) was significantly different from the innocent condition. Raw calculations of proximity were smaller among those in the guilty conditions, though significantly so only in the guilty with no countermeasures condition (β = –.377, $p =.019$) and the mental countermeasures condition (β = –.398, $p =.029$). Physical countermeasures were suggestive (β = – .301, $p = .08$), but all countermeasures together were not (β = –.130, $p = .44$). Frowning significantly increased among participants in the guilty condition when a target stimulus was presented (β = .442, $p < .01$), but no such difference was found for those in countermeasures conditions.

Prediction

Results showing how deception influences certain indicators from an explanatory standpoint provide unique insights into the strength and robustness of indicators under various conditions. Using these indicators to effectively predict deception is an altogether different challenge, and can provide insight into the potential robustness of a multisensor system. The overall system used in this study was evaluated for its predictive capability compared to innocent responses when different countermeasure types were used.

In many cases, including this one, generating a post hoc machine learning algorithm that achieves very high accuracy on a given data set is almost trivial. Overfitting a prediction model to a specific data set can result in seemingly high accuracy, but the generalizability of such algorithms is questionable [55]. Several steps were taken to produce conservative estimates of performance so that reported results have a high likelihood of performing similarly to how they would in real-world applications.

First, rather than selecting the best-performing algorithm, the output of several prediction algorithms was combined. Though many types of machine-learning algorithms have been used for deception detection, there is no consensus on which is most appropriate, either in general or in this particular context. We therefore selected a subset of several common types of algorithms for each indicator and indicator group,[2] including naive Bayes [43], logistic regression [22], random forest [12], and SVM (support vector networks) [21]. A naive ensemble algorithm [78] equally weighted the output of each of these. By using a naive ensemble classifier, the influence of a single type of algorithm is muted. Therefore, overall performance remains reliable even if a particular algorithm by chance outperformed or underperformed others for this data set.[3]

Second, each prediction algorithm used a two-thirds/one-third training/testing split, meaning that each classification algorithm was generated by analysis of only two-thirds of the data. Performance was then determined by using that prediction algorithm on the one-third that was held back. This approach penalizes overfitting algorithms to the current data set. Third, each training phase used tenfold cross-validation, an iterative training method designed to further mitigate overfitting. This combination of actions is likely to result in conservative accuracy estimates.

Separate models were created to compare the efficacy of the confirmatory and exploratory indicators examined in this study. Pupil dilation is further reported in a single-indicator model, since it has been repeatedly shown to be a strong indicator of deception and appears similarly robust in this context, given the regression analysis results. In addition to these smaller models, a complete model combining all confirmatory and exploratory indicators in a single model was created. Ensemble results for the combined model (labeled “All indicators”) and each of the other models are reported in Table 4.

Table 4. Prediction Capability of Indicators in a Controlled Screening System

Indicators	Accuracy	Sensitivity	Specificity
Baseline (no countermeasures)			
All indicators	0.86	0.90	0.82
Confirmatory indicators	0.86	1.00	0.64
Exploratory indicators	0.67	0.40	0.90
Pupil dilation only	0.76	0.80	0.72
Mental countermeasures			
All indicators	0.86	0.90	0.82
Confirmatory indicators	0.81	0.90	0.73
Exploratory indicators	0.52	0.10	0.91
Pupil dilation only	0.86	0.90	0.82
Physical countermeasures			
All indicators	0.62	0.40	0.82
Confirmatory indicators	0.62	0.40	0.82
Exploratory indicators	0.57	0.50	0.64
Pupil dilation only	0.62	0.60	0.64
All countermeasures at once			
All indicators	0.71	0.80	0.64
Confirmatory indicators	0.80	0.70	0.91
Exploratory indicators	0.33	0.20	0.45
Pupil dilation only	0.76	0.90	0.64

Notes: Sensitivity = detecting guilt; specificity = detecting innocence.

Discussion

IS research is helping to pioneer automated deception detection systems that can be more broadly applied, more economical and easier/simpler to deploy, and more scientifically sound. The current study sought to examine the robustness of one such system to individual attempts to counter its detection methods.

Summary of Results

The regression analyses provided direct support for some but not all hypotheses. Analysis of pupil dilation and general body movement replicated prior work showing dilated pupils (H1) and kinesic rigidity (H4) during presentation of target stimuli in a controlled interview. Traditional countermeasures were not effective at countering these responses, supporting rigidity expectations (H5) and pupil dilation expectations with regard to mental countermeasures (H2), but contrary to expectations for pupil dilation (H3). Pupil dilation was the strongest effect among those investigated and appeared to be the most resilient to countermeasures. The pupil dilation resulting from the orienting response was pronounced, and no decrease in this effect occurred when mental distraction or pain was used, though self-induced pain did appear to affect the prediction performance of this factor. Center-of-screen gaze appeared to increase among guilty participants during responses to relevant stimuli; contrary to prior research, the increase was not statistically significant (H6, H7). The most likely explanation for this difference is that the effect is not as pronounced as other indicators. Participants in the guilty with no countermeasures condition exhibited greater maximum vocal pitch (H8), closer proximity (H9), and increased frowning (H10), as hypothesized. A review of the results also suggests that mental countermeasures may be at least somewhat effective against vocal pitch, though physical countermeasures were not. Proximity significantly decreased; and unlike vocal pitch, proximity appears to be more robust to mental countermeasures than to physical countermeasures. Though increased frowning was significant in the baseline guilty condition, it was nonexistent in all of the countermeasures conditions, suggesting that it may be easily overtly controlled. Table 5 summarizes results from explicit hypotheses.

Table 5. Summary of Hypothesis Testing Results

	Hypothesis	Supported?
1	Deceptive individuals will have a larger pupillary response to target stimuli than to baseline stimuli.	Yes
2	Deceptive individuals using mental countermeasures will exhibit increased pupil dilation during responses to target stimuli.	Yes
3	Physical countermeasures will reduce the pupil dilation differential between target and baseline stimuli.	No
4	Deceptive individuals will exhibit less overall movement when viewing and responding to a target stimulus.	Yes
5	Deceptive individuals employing mental or physical countermeasures will exhibit less overall movement when viewing and responding to a target stimulus.	Yes
6	Deceptive individuals will show increased gaze time at the screen center when presented with a target stimulus.	No
7	Deceptive individuals employing mental or physical countermeasures will exhibit greater gaze duration at the screen center when viewing and responding to a target stimulus.	No
8	Deceptive individuals will exhibit greater vocal pitch when responding to a target stimulus.	Yes
9	Deceptive individuals will move closer to a target stimulus.	Yes
10	Deceptive individuals will increase frowning when presented with a target stimulus.	Yes

At a more general level, the prediction capability of a controlled screening system appeared to be more robust when multiple indicators for this context were used for prediction, at least when compared to the strongest predictor by itself (i.e., pupil dilation). However, the multiple-signal buoyancy effect was not necessarily robust to countermeasures. Mental countermeasures appeared to alter behavior and physiology in various ways, but these changes did not effectively undermine the overall effectiveness of the system. Nevertheless, physical countermeasures manipulated behavior and physiology in such a manner that despite leaking clear signals, overall system performance decreased. Performance likewise decreased when multiple types of countermeasures were attempted simultaneously, although not to the same degree. Possibly, physical countermeasures were the key driver of the performance drop in this group as well. Interestingly, although physical countermeasures failed to mask many indicators of deception, they succeeded in foiling the system. These results suggest a need for further research to examine the drivers of this phenomenon, and how to increase system resilience. This research will likely involve improved classification algorithms, improved detection of deception indicators, or incorporation of countermeasure detection. Each of these efforts will likely require a more nuanced understanding of human behavior and psychophysiology within this context.

Contributions

This study provides new insight into the robustness of IS deception detection screening systems. Whereas prior research provides very limited evidence that a multisensor, noninvasive, controlled screening approach can work, this study directly examines a multi-indicator system, and by doing so illuminates potential strengths and limitations of its potential. Relative strengths of various indicators in a controlled automated screening context are noted, and areas for greatest improvement identified. In

particular, pupil dilation as a function of the orienting response appears to be the strongest and most robust indicator. Kinesic rigidity and gazing at the screen center are relatively weaker indicators, though physical countermeasures seem to strengthen these two indicators from an explanatory perspective. These may ultimately prove valuable for detecting or counteracting physical countermeasures, which this study shows is an apparent weakness of this type of system.

While rigidity and pupil dilation indicators were more robust to countermeasures, the exploratory indicators vocal pitch, proximity, and frowning were clearly affected, suggesting that they may be more overtly controllable. Depth of frowning in particular appears to be easily controlled, even when attempting to concurrently control additional behaviors simultaneously.

Technologies that can identify deception rapidly and without contact have the potential to be used in a variety of interviewing and screening contexts, changing how integrity and security are managed. As with the polygraph, new systems will encounter some individuals who will attempt to mitigate their effectiveness through the use of countermeasures. Ultimately, the results of this study should inform future controlled screening system design—driving revised detection algorithms, refined interactions, and key procedural modifications. Thus, this study represents one step in a much larger effort to create system-driven solutions to detect deception. In this way, this paper also contributes an example of how an impact-driven design science study can make theoretical and practical contributions at a proof-of-concept stage in the Nunamaker framework [69].

Implications for Theory

Prominent deception theories outline how deception influences cognitive and behavioral processes, leading to deception indicators. As evidence for many types of indicators continue to be uncovered in the deception literature, it is moderating factors—limitations and boundaries to each type of indicator—that become increasingly important to generating a more nuanced understanding of the underlying processes and to understanding their comparative strength and robustness.

This study provides evidence of the relative strength of various indicator types, and that overt countermeasures influence nontraditional deception indicators differentially. Leakage theory claims that despite efforts to maintain truthful appearances, no one is capable of controlling every verbal and nonverbal process simultaneously [29, 96]. Interpersonal deception theory proposes that while a deceiver may engage in strategic behavior to manipulate some indicators, other indicators may increase in intensity [15]. This study adds specificity to those claims, in that physical countermeasures, designed to introduce noise into cognitive arousal indicators, seemed to also increase the intensity of indicators not directly stemming from cognitive arousal. Mental countermeasures, in a similar vein, seem to diminish indicator intensity on certain emotion and tension indicators but not indicators resulting from autonomic processes.

From an indicator standpoint, the results of this study suggest that indicators stemming from cognitive arousal and low-level perception (e.g., pupil dilation, rigidity) may generally be less susceptible to noise and overt attempts at manipulation because of their more autonomic nature, as compared to indicators that stem more from emotional variations (e.g., vocal pitch, frowning). Cognitive distraction may to some extent mitigate the intensity of experienced emotions, thereby decreasing the prevalence of emotional indicators. Facial emotion in particular appears to be easily controllable compared to other indicators.

There is still uncertainty as to the role of proximity in this context. The traditional explanation for this behavioral variation in alternative contexts is immediacy, but the results of this study suggest a different driving factor. While we considered cognitive interest to be a possible explanation, mental distraction did not appear to affect this indicator, suggesting either that people find it difficult to suppress their interest or that proximity decreases may stem from an altogether different source.

Limitations and Future Directions

Just as no single veracity indicator has proved to be a "Pinocchio's nose," or guaranteed sign of deception, no system seeking to measure a physiological or behavioral process is expected to be foolproof for all cases and circumstances. It is important that research in this area be contextualized, clearly explaining boundaries and limitations for results and implications. The current study focuses on fully automated screening system interviews, specifically automated systems employing a highly controlled interview.

One limitation of this work is the set of countermeasures used by participants in the countermeasures treatments. The countermeasures taught to participants had been identified in previous deception literature; however, they were traditionally employed to thwart polygraph examinations. While the concept of using countermeasures to create bogus physiological and behavioral responses certainly applies to the sensors used in this study, it is possible that alternative actions may be conceived that are better able to thwart this new class of systems. As such, an area of future research is the identification of new types of countermeasures that may be employed specifically for use against the sensors evaluated in this research. Future work should also assess the effectiveness of these novel countermeasures. The results of this experiment indicate that physical countermeasures are most effective at reducing classification accuracy and are therefore the most pressing area of study.

Training is also a limitation of this study, since participants had only a few minutes to learn countermeasures and practice them on their own. Reasonably, countermeasure effectiveness may improve with greater training and practice, and future studies should examine the effects of various levels of training. However, individuals rarely train for extended periods before taking a polygraph exam or smuggling illicit items into a sports arena or secure building, so these results may generalize well to many scenarios. Nevertheless, highly motivated individuals may undergo extensive training, and the effectiveness of such training in this context is an open question for research.

Finally, because of the novelty of the problem space, the prediction algorithms used in this study are not representative of a field standard. The purpose of this study was not to suggest that automated screening systems have currently achieved a specific performance rate. Rather, the goal was to determine relative performance under disparate conditions, challenging their robustness. Determining optimal approaches and reliable estimates for prediction will require a stream of follow-up research evaluating various options for improving classification. For instance, classification algorithms that consider a range of readings and common patterns in indicator readings may be better able to identify and adapt to physical countermeasures. Physical countermeasures both enhanced and failed to mask some signals, but those signals did not easily translate into straightforward prediction affordance.

Conclusion

Automated screening systems for deception detection have the potential to be a major disruptor in practice. The lower cost and noninvasive nature of systems that integrate a variety of deception indicator types provide for wider application, faster processes, and less human bias than traditional systems. This potential for high impact motivates this study, which exemplifies a proof-of-concept contribution within an impact-driven design science program of research. Nevertheless, results from this study indicate that additional research is needed to improve the robustness of the design concept. Results of this study further provide evidence that pupil dilation is among the strongest and most robust of indicators in this context. Other indicators also seem to have potential, though it appears that classification algorithms will need to be hardened to various potential countermeasures that may be employed to undermine such systems. Where a single indicator might be weak to a particular type of countermeasure, the use of many indicators stemming from diverse psychophysiological and behavioral mechanisms may improve predictive accuracy, if classification algorithms can sufficiently adapt.

Acknowledgments

The Department of Homeland Security's (DHS) National Center for Border Security and Immigration (BORDERS) and the Center for Identification Technology Research (CITeR), a National Science Foundation (NSF) Industry/University Cooperative Research Center (I/UCRC), provided funding for this research. Statements provided herein do not necessarily represent the opinions of the funding organizations. Valuable input for this paper was provided by Jay F. Nunamaker, Judee K. Burgoon, anonymous reviewers, and several graduate students at the Center for the Management of Information (CMI).

Notes

1. The first manipulation check question was "Were you carrying any illicit objects through the checkpoint today?" The second question, "Which of the following people were you asked to deliver the bag to?" was followed by four distinct images of faces, one of which was the same image provided in the instructions as the target recipient of the bag (see "Experiment Task" section). Participants who answered incorrectly to both questions were disqualified.

[2.] Missing values were imputed using a random forest approach.

[3.] As an interesting side note, although the best reported ensemble classifiers produced 86 percent overall accuracy, the best prediction performance resulted from a trained logistic regression model, when no countermeasures were used and all indicators were included (90 percent overall, 90 percent sensitivity, 90 percent specificity). As noted, however, whether logistic regression would outperform other classification algorithms in a replication is uncertain.

References

1. Alder, K. To tell the truth: The polygraph exam and the marketing of American expertise. *Historical Reflections*, 24, 3 (1998), 487–525.

2. Ambach, W.; Bursch, S.; Stark, R.; and Vaitl, D. A Concealed Information Test with multimodal measurement. *International Journal of Psychophysiology*, 75, 3 (2010), 258–267.

3. Bachorowski, J.A., and Owren, M.J. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 6, 4 (1995), 219–224.

4. Beatty, J., and Kahneman, D. Pupillary changes in two memory tasks. *Psychonomic Science*, 5, 10 (1966), 371–372.

5. Ben-Shakhar, G. Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology*, 22, 3 (1985), 292–299.

6. Ben-Shakhar, G. The roles of stimulus novelty and significance in determining the electrodermal orienting response: Interactive versus additive approaches. *Psychophysiology*, 31, 4 (1994), 402–411.

7. Ben-Shakhar, G., and Dolev, K. Psychophysiological detection through the guilty knowledge technique: Effect of mental countermeasures. *Journal of Applied Psychology*, 81, 3 (1996), 273–281.

8. Ben-Shakhar, G., and Meijer, E. Skin conductance, respiration, heart rate, and P300 in the concealed information test: A meta analysis. *International Journal of Psychophysiology*, 85, 3 (2012), 324–325.

9. Bond, C.F., and DePaulo, B.M. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10, 3 (2006), 214–234.

10. Bracha, H.S.; Ralston, T.C.; Matsukawa, J.M.; Williams, A.E.; and Bracha, A.S. Does “fight or flight” need updating?*Psychosomatics*, 45, 5 (2004), 448–449.

11. Bradley, M.M.; Miccoli, L.; Escrig, M.A.; and Lang, P.J. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45, 4 (2008), 602–607.

12. Breiman, L. Random forests. *Machine Learning*, 45, 1 (2001), 5–32. **13.** Brinke, L. ten; Porter, S.; and Baker, A. Darwin the detective: Observable facial muscle contractions reveal emotional high-stakes lies. *Evolution and Human Behavior*, 3, 4 (2011), 411–416.

14. Buller, D.B., and Aune, R.K. Nonverbal cues to deception among intimates, friends, and strangers. *Journal of Nonverbal Behavior*, 11, 4 (1987), 269–290.

15. Buller, D.B., and Burgoon, J.K. Interpersonal deception theory. *Communication Theory*, 6, 3 (1996), 203–242.

16. Burgoon, J.K.; Buller, D.B.; Afifi, W.; White, C.; and Buslig, A. The role of immediacy in deceptive interpersonal interaction. In*Conference of the International Communication Association*. Chicago, 1996.

17. Burgoon, J.K.; Proudfoot, J.G.; Wilson, D.; and Schuetzler, R. Hidden patterns of nonverbal behavior associated with truth and deception. *Journal of Nonverbal Behavior*, 38, 3 (2014), 325–254.

18. Campbell, B.A.; Wood, G.; and McBride, T. Origins of orienting and defensive responses: An evolutionary perspective. In P.J. Lang, R.F. Simons, and M. Balaban (eds.), *Attention and Orienting: Sensory and Motivational Processes*. Mahwah, NJ: Erlbaum,1997, pp. 41–67.

19. Caso, L.; Maricchiolo, F.; Bonaiuto, M.; Vrij, A.; and Mann, S. The impact of deception and suspicion on different hand movements. *Journal of Nonverbal Behavior*, 30, 1 (2006), 1–19.

20. Chapman, C.R.; Oka, S.; Bradshaw, D.H.; Jacobson, R.C.; and Donaldson, G.W. Phasic pupil dilation response to noxious stimulation in normal volunteers: Relationship to brain evoked potentials and pain report. *Psychophysiology*, 36, 1 (1999), 44–52.

21. Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning*, 20, 3 (1995), 273–297.

22. Cox, D.R. The regression analysis of binary sequences. *Journal of Royal Statistical Society*, 20 (1958), 215–242.

23. DePaulo, B.M., and Kirkendol, S.E. The motivational impairment effect in the communication of deception. In J. Yuille (ed.),*Credibility Assessment*. Deurne, Belgium: Kluwer, 1989, pp. 51–70.

24. DePaulo, B.M.; Kirkendol, S.E.; Tang, J.; and O'Brien, T.P. The motivational impairment effect in the communication of deception: Replications and extensions. *Journal of Nonverbal Behavior*, 12, 3 (1988), 177–201.

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

26. Derrick, D.C.; Elkins, A.C.; Burgoon, J.K.; Nunamaker, J.F., Jr.; and Zeng, D. Border security credibility assessments via heterogeneous sensor fusion. *IEEE Intelligent Systems*, 25, 3 (2010), 41–49.

27. Ekman, P., and Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry*, 32, 1 (1969), 88–106.

28. Ekman, P., and Friesen, W.V. Hand movements. *Journal of Communication*, 22, 4 (1972), 353–374.

29. Ekman, P., and Friesen, W.V. Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29, 3 (1974), 288–298.

30. Elaad, E. Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology*, 75, 5 (1990), 521–529.

31. Elaad, E., and Ben-Shakhar, G. Effects of mental countermeasures on psychophysiological detection in the Guilty Knowledge Test. *International Journal of Psychophysiology*, 11, 2 (1991), 99–108.

32. Elaad, E.; Ginton, A.; and Jungman, N. Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology*, 77, 5 (1992), 757–767.

33. Elkins, A.C.; Derrick, D.C.; and Gariup, M. The voice and eye gaze behavior of an imposter: Automated interviewing and detection for rapid screening at the border. In *Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, 2012.

34. Ellermeier, W., and Westphal, W. Gender differences in pain ratings and pupil reactions to painful pressure stimuli. *Pain*, 61, 3 (1995), 435–439.

35. Fuller, C.M.; Biros, D.P.; and Wilson, R.L. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46, 3 (2009), 695–703.

36. Furedy, J.J. The North American polygraph and psychophysiology: Disinterested, uninterested, and interested perspectives.*International Journal of Psychophysiology*, 21, 2 (1996), 97–105.

37. Gamer, M.; Rill, H.G.; Vossel, G.; and Gödert, H.W. Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology*, 60, 1 (2006), 76–87.

38. Ganis, G.; Rosenfeld, J.; Meixner, J.; Kievit, R.; and Schendan, H. Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage*, 55, 1 (2011), 312–319.

39. Gardner, R.M.; Beltramo, J.S.; and Krinsky, R. Pupillary changes during encoding, storing, and retrieval of information.*Perceptual and Motor Skills*, 41, 3 (1975), 951–955.

40. Goldwater, B.C. Psychological significance of pupillary movements. *Psychological Bulletin*, 77, 5 (1972), 340–355.

41. Gray, J.A. *The Psychology of Fear and Stress*. 2nd ed. Cambridge: Cambridge University Press, 1988.

42. Gregor, S., and Hevner, A.R. Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37, 2 (2013), 337–355.

43. Hand, D.J., and Yu, K. Idiot's Bayes—not so stupid after all? *International Statistical Review*, 69, 3 (2001), 385–389.

44. Harnsberger, J.D.; Hollien, H.; Martin, C.A.; and Hollien, K.A. Stress and deception in speech: Evaluating layered voice analysis. *Journal of Forensic Sciences*, 54, 3 (2009), 642–650.

45. Hevner, A.R.; March, S.T.; Park, J.; and Ram, S. Design science in information systems research. *MIS Quarterly*, 28, 1 (2004), 75–105.

46. Honts, C.R.; Devitt, M.K.; Winbush, M.; and Kircher, J.C. Mental and physical countermeasures reduce the accuracy of the Concealed Knowledge Test. *Psychophysiology*, 33, 1 (1996), 84–92.

47. Honts, C.R.; Hodes, R.L.; and Raskin, D.C. Effects of physical countermeasures on the physiological detection of deception.*Journal of Applied Psychology*, 70, 1 (1985), 177–187.

48. Honts, C.R., and Kircher, J.C. Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, 79, 2 (1994), 252–259.

49. Honts, C.R.; Raskin, D.C.; and Kircher, J.C. Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology*, 1, 3 (1987), 241–247.

50. Hurley, C.M., and Frank, M.G. Executing facial control during deception situations. *Journal of Nonverbal Behavior*, 35, 2 (2011), 119–131.

51. Iacono, W.G., and Lykken, D.T. The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology*, 82, 3 (1997), 426–433.

52. Janisse, M.P. *Pupillometry: The Psychology of the Pupillary Response*. Washington, DC: Hemisphere Publishing, 1977.

53. Janisse, M.P., and Bradley, M.T. Deception, information and the pupillary response. *Perceptual and Motor Skills*, 50, 3 (1980), 748–750.

54. Kahneman, D. *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall, 1973.

55. Kearns, M.; Mansour, Y.; Ng, A.Y.; and Ron, D. An experimental and theoretical comparison of model selection methods.*Machine Learning,* 27, 1, (1997), 7–50.

56. Knapp, M.L., and Comadena, M.E. Telling it like it isn't: A review of theory and research on deceptive communications.*Human Communication Research*, 5, 3 (1979), 270–285.

57. Krapohl, D.J.; McCloughlan, J.B.; and Senter, S.M. How to use the Concealed Information Test. *Polygraph*, 35, 3 (2009), 34–49.

58. Littlewort, G.; Whitehill, J.; Wu, T.; et al. The Computer Expression Recognition Toolbox (CERT). In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*. Santa Barbara, CA, 2011, pp. 298–305.

59. Lubow, R.E., and Fein, O. Pupillary size in response to a visual Guilty Knowledge Test: New technique for the detection of deception. *Journal of Experimental Psychology-Applied*, 2, 2 (1996), 164–177.

60. Lykken, D.T. *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. New York: Plenum Trade, 1998.

61. MacLaren, V.V. A quantitative review of the guilty knowledge test. *Journal of Applied Psychology*, 86, 4 (2001), 674–683.

62. Matsuda, I.; Nittono, H.; and Allen, J.J.B. The current and future status of the Concealed Information Test for field use.*Frontiers in Psychology*, 3 (2012), 1–11.

63. Matte, J.A. *Forensic Psychophysiology Using the Polygraph*. Williamsville, NY: J.A.M. Publications, 1996.

64. Meservy, T.O.; Jensen, M.L.; Burgoon, J.K.; et al. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20, 5 (2005), 36–43.

65. National Research Council. *The Polygraph and Lie Detection*. Washington, DC: National Academies Press, 2003.

66. Nunamaker, J.F., Jr., and Briggs, R.O. Toward a broader vision for information systems. *ACM Transactions on Management Information Systems*, 2, 4 (2011), 1–12.

67. Nunamaker, J.F., Jr.; Chen, M.; and Purdin, T.D.M. Systems development in information systems research. *Journal of Management Information Systems*, 7, 3 (1991), 89–106.

68. Nunamaker, J.F., Jr.; Derrick, D.C.; Elkins, A.C.; Burgoon, J.K.; and Patton, M.W. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, 28, 1 (2011), 17–48.

69. Nunamaker, J.F., Jr.; Twyman, N.W.; and Giboney, J.S. Breaking out of the design science box: High-value impact through multidisciplinary design science programs of research. In *Americas Conference on Information Systems*. Chicago, IL, 2013.

70. Nunnally, J.C.; Knott, P.D.; and Duchowski, A. Pupillary response as a general measure of activation. *Attention, Perception, and Psychophysics*, 2, 4 (1967), 149–155.

71. Osugi, A. Gap and connection between laboratory research and field applications of the CIT in Japan. *International Journal of Psychophysiology*, 77, 3 (2010), 238.

72. Patton, M. Decision Support for Rapid Assessment of Truth and Deception Using Automated Assessment Technologies and Kiosk-Based Embodied Conversational Agents, dissertation, 2009, University of Arizona.

73. Pavlidis, I.; Eberhardt, N.L.; and Levine, J.A. Seeing through the face of deception. *Nature*, 415, 6867 (2002), 35.

74. Pavlidis, I., and Levine, J. Thermal facial screening for deception detection. In *24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference*. Houston, Texas, 2002, pp. 1143–1144.

75. Porter, S.; Brinke, L. ten; and Wallace, B. Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior*, 36, 1 (2012), 23–37.

76. Recio, G.; Shmuilovich, O.; and Sommer, W. Should I smile or should I frown? An ERP study on the voluntary control of emotion-related facial expressions. *Psychophysiology*, 51, 8 (2014), 789–799.

77. Rockwell, P.; Buller, D.B.; and Burgoon, J.K. Measurement of deceptive voices: Comparing acoustic and perceptual data.*Applied Psycholinguistics*, 18, 4 (1997), 471–484.

78. Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1–2 (2010), 1–39.

79. Rosenfeld, J., and Labkovsky, E. New P300-based protocol to detect concealed information: Resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures. *Psychophysiology*, 47, 6 (2010), 1002–1010.

80. Rosenfeld, J.; Labkovsky, E.; Winograd, M.; Lui, M.A.; Vandenboom, C.; and Chedid, E. The complex trial protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, 45, 6 (2008), 906–919.

81. Rosenfeld, J.; Soskins, M.; Bosh, G.; and Ryan, A. Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, 41, 2 (2004), 205–219.

82. Ryan, J.D.; Hannula, D.E.; and Cohen, N.J. The obligatory effects of memory on eye movements. *Memory*, 15, 5 (2007), 508–525.

83. Scherer, K.R.; Johnstone, T.; and Klasmeyer, G. Vocal expression of emotion. In R.J. Davidson, K.R. Scherer, and H.H. Goldsmith (eds.), *Handbook of Affective Sciences*. New York: Oxford University Press, 2003, pp. 433–456.

84. Schwedes, C., and Wentura, D. The revealing glance: Eye gaze behavior to concealed information. *Memory and Cognition*, 40, 4 (2012), 642–651.

85. Sokolovsky, A.; Rothenberg, J.; Labkovsky, E.; Meixner, J.; and Rosenfeld, J. A novel countermeasure against the reaction time index of countermeasure use in the P300-based complex trial protocol for detection of concealed information.*International Journal of Psychophysiology*, 81, 1 (2011), 60–63.

86. Titze, I.R., and Martin, D.W. Principles of voice production. *Journal of the Acoustical Society of America*, 104, 3 (1998), 1148.

87. Trovillo, P.V. A history of lie detection. *Journal of Criminal Law and Criminology*, 29, (1939), 848–881.

88. Twyman, N.W.; Elkins, A.C.; Burgoon, J.K.; and Nunamaker, J.F., Jr. A rigidity detection system for automated credibility assessment. *Journal of Management Information Systems*, 31, 1 (2014), 173–201.

89. Twyman, N.W.; Lowry, P.B.; Burgoon, J.K.; and Nunamaker, J.F., Jr. Autonomous scientifically controlled screening systems for detecting information purposefully concealed by individuals. *Journal of Management Information Systems*, 31, 3 (2014), 106–137.

90. Verschuere, B.; Crombez, G.; De Clercq, A.; and Koster, E.H.W. Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology*, 41, 3 (2004), 461–466.

91. Verschuere, B.; Meijer, E.; and De Clercq, A. Concealed information under stress: A test of the orienting theory in real-life police interrogations. *Legal and Criminological Psychology*, 16, 2 (2011), 348–356.

92. Vizer, L.M.; Zhou, L.N.; and Sears, A. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, 67, 10 (2009), 870–886.

93. Vrij, A., and Mann, S. Telling and detecting lies in a high-stakes situation: The case of a convicted murderer. *Applied Cognitive Psychology*, 15, 2 (2001), 187–203.

94. Vrij, A.; Semin, G.R.; and Bull, R. Insight into behavior displayed during deception. *Human Communication Research*, 22, 4 (1996), 544–562.

95. Zuckerman, M.; DePaulo, B.M.; and Rosenthal, R. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, 14, 1 (1981), 1–59.

96. Zuckerman, M.; Koestner, R.; Colella, M.J.; and Alton, A.O. Anchoring in the detection of deception and leakage. *Journal of Personality and Social Psychology*, 47, 2 (1984), 301–311