Psychology Theses & Dissertations                                    Psychology

Summer 2009

# The Effects of Automation Expertise, System Confidence, and Image Quality on Trust, Compliance, and Performance

Randall D. Spain
*Old Dominion University*

# THE EFFECTS OF AUTOMATION EXPERTISE, SYSTEM CONFIDENCE, AND IMAGE QUALITY ON TRUST, COMPLIANCE, AND PERFORMANCE

by

Randall D. Spain
B.A. May 2003, Christopher Newport University
M.S. December 2006, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
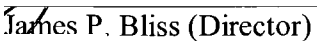Requirement for the Degree of
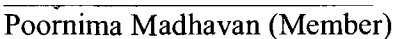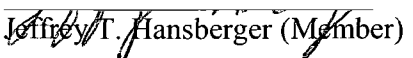
DOCTOR OF PHILOSOPHY

HUMAN FACTORS PSYCHOLOGY

OLD DOMINION UNIVERSITY
August 2009

Approved by:

James P. Bliss (Director)

Poornima Madhavan (Member)

Jeffrey T. Hansberger (Member)

# ABSTRACT

## THE EFFECTS OF AUTOMATION EXPERTISE, SYSTEM CONFIDENCE, AND IMAGE QUALITY ON TRUST, COMPLIANCE, AND PERFORMANCE

Randall D. Spain
Old Dominion University, 2009
Director: Dr. James P. Bliss

This study examined the effects of automation expertise, system confidence, and image quality on automation trust, compliance, and detection performance. One hundred and fifteen participants completed a simulated military target detection task while receiving advice from an imperfect diagnostic aid that varied in expertise (expert vs. novice) and confidence (75% vs. 50% vs. 25% vs. no aid). The task required participants to detect covert enemy targets in simulated synthetic aperture radar (SAR) images. Participants reported whether a target was present or absent, their decision-confidence, and their trust in the diagnostic system's advice. Results indicated that system confidence and automation expertise influenced automation trust, compliance, and measures of detection performance, particularly when image quality was poor. Results also highlighted several incurred costs of system confidence and automation expertise. Participants were more apt to generate false alarms as system confidence increased and when receiving diagnostic advice from the expert system. Data also suggest participants adopted an analogical trust tuning strategy rather than an analytical strategy when evaluating system confidence ratings. This resulted in inappropriate trust when system confidence was low. Theoretical and practical implications regarding the effects of system confidence and automation expertise on automation trust and the design of diagnostic automation are discussed.

This dissertation is dedicated to my loving wife, Kate. Without your unconditional love, support, and patience this would not be possible.

# ACKNOWLEDGMENTS

First and foremost, I would to thank God for providing me with the strength, motivation, people, and support that guided me through my five years of graduate school.

I would also like to thank Dr. James Bliss for his patience and advice. As an academic mentor, he challenged me to grow and accomplish things I never imagined I would. I could not have completed this research without his insightful comments, suggestions, and commitment.

Special thanks also go to my dissertation committee members, Drs. Poornima Madhavan and Jeffrey Hansberger, for their efforts and advice on this dissertation and the additional projects we have completed. Furthermore, I would like to acknowledge the tremendous support I received from Peggy Kinard and Mary Boswell; your help was invaluable, as were the treats that were always stocked in MGB 250.

Finally, I would like to think my family and friends for supporting me and providing the much needed stress relief from graduate school. Mom and Dad your strength, generosity, and guidance are irreplaceable, thank you! To Travis, Christy, Nanny, and the rest of my family, thanks for your unconditional support. To my friends at home, thanks for the weekend jams - they kept me sane. To my fellow lab members and ODU colleagues, thanks for the friendship and support! I acknowledge that each of you played a significant role in helping me complete my dissertation and Ph.D.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

Technological advancements have allowed engineers to introduce automation into complex technical systems. Automation is technology that gathers, filters, and organizes information, makes decision, and carries out actions that a human would otherwise execute (Parasuraman & Riley, 1997). Parasuraman, Sheridan, and Wickens (2000) have identified four stages of automation that parallel the stages of human information processing. These stages are information acquisition, diagnosis, action selection, and execution. Information acquisition automation assists operators by selecting, organizing, highlighting, and filtering information that needs to be processed by the operator. Examples include information filtering and prioritization, cueing, and highlighting (Wickens & Hollands, 2000). Diagnostic automation assists operators by performing cognitive operations such as integration and diagnosis. Examples include alerts, alarms and decision support systems. Action selection and execution automation assist operators by generating decision alternatives and executing actions on behalf of the operator (Wickens & Hollands, 2000).

Despite assumptions that automation can replace the human element, there is consensus that the human operator must remain "in-the-loop" as an integral part of the system (Parasuraman & Wickens, 2008). Therefore, it has become increasingly important to understand how humans interact with automation. One variable that influences the joint performance of the human machine "team" is trust. Trust is an attitude that guides automation reliance and compliance (Muir, 1989; Lee & Moray, 1992; Lee & See, 2004; Parasuraman & Riley, 1997).

This dissertation adheres to the format of *Human Factors*

Diagnostic automation, such as decision support systems, possesses several facets that may affect automation trust. First, diagnostic systems are based on imperfect algorithms that function in an uncertain world. Consequently, automation failures are likely to occur. These failures can take two forms: automation misses and automation false alarms. There is considerable evidence to suggest that false alarms are more damaging than misses (Bliss, 2003) and that the two types of errors affect trust related behaviors differently. Automation false alarms tend to affect operator compliance, whereas automation misses tend to affect operator reliance (Rice, *in press*). Second, operators may or may not have insight into the raw data the system is diagnosing. Having access to raw data allows the operator to determine the validity of the diagnosis, rather than blindly accepting the system's recommendation (Sorkin & Woods, 1985). When raw data are not available, or when the data are too difficult to interpret, the operator has only the automation's recommendation to base his or her judgments. In these situations, an operator's decision to rely on the system's advice will likely depend on a number of factors including his or her perception of automation capability and preconceived cognitive biases (Sheridan & Parasuraman, 2006).

The purpose of this study was to empirically determine how information pertaining to automation capability, specifically automation expertise and system confidence, affected trust and compliance with an automated diagnostic system during a simulated target detection task. The ensuing sections provide a summary of the pertinent literature, including a review of the theoretical and empirical literature concerning trust in automation, followed by a detailed overview of the research domain and the experimental methodology.

**Defining Automation Trust**

Researchers have defined trust with respect to automation in various ways. Muir (1989) defines trust as a generalized expectation related to the occurrence of a future event. Sheridan (2002) defines trust as a cause and effect that is based on the judged reliability, perceived robustness, and familiarity of automation. In Sheridan's expanded definition, he states that trust affects the interaction with automation and the interaction with automation affects trust. Lee and See (2004) describe trust as "an attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability (pg 54)." In this definition, the agent refers to a human or machine.

**Historical Theories of Automation Trust**

The role of trust in human-computer interaction has been the focus of much research over the past two decades (see Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Madhavan & Wiegmann, 2007; Lee & Moray, 1992, 1994; Lee & See, 2004; Muir, 1989; Muir & Moray, 1996). From these efforts, researchers have developed various theories that describe the development of automation trust. Muir (1989) developed the first theory of automation trust. Her Machine Trust Theory states that trust is contingent upon a machine's predictability, dependability, and an operator's faith that the machine will function for his or her best interest. Muir's theory proposes that in the first stage of trust development, predictability, operators observe system performance and make judgments concerning a machine's reliability. If the operator observers inconsistencies in performance, trust will diminish. As the relationship matures, trust becomes dependent on attributions of performance, such as dependability. These attributes can be influenced by perceptions of performance and hearsay information regarding the machine's capabilities.

The final stage of trust development requires the establishment of faith in future performance. That is, the operator must have faith that the machine will reliably perform the tasks it was designed to perform.

Lee and Moray (1992) extended Muir's (1989) theory of automation trust. Contrary to Muir's framework, Lee and Moray hypothesized that automation trust consisted of four dimensions that were complimentary in nature rather than orthogonal. The first dimension, *the foundation of trust*, represents the fundamental assumption of natural order that makes the other levels of trust possible. The second dimension, *performance*, reflects the expectation of consistent, stable, and desirable performance. This dimension is similar Muir's concept of predictability. The third dimension, *process*, represents the underlying functioning that guide automation performance. This dimension corresponds with Muir's notion of dependability. The fourth dimension, *purpose*, reflects the underlying motives or intent of the machine. Purpose corresponds to faith and benevolence, and reflects a positive orientation regarding a machine's future performance.

Lee and Moray (1992) empirically tested their theory by conducting a study in which participants completed sixty trials of a simulated orange juice pasteurization task. Participants were required to monitor system performance and allocate control of plant pump rates and heater settings to manual or automated control. Under manual control, participants manually controlled pump rates and settings, whereas under automated control automation controlled these settings. To investigate the effect of faults on trust and performance, one of the pumps periodically malfunctioned and disrupted juice production. The size of the fault ranged in magnitude. Lee and Moray measured trust in

the automated system by having participants complete a questionnaire at the close of every trial. Results indicated that participants' trust was sensitive to automation failures and that the loss of trust was proportional to the magnitude of the automated fault. Furthermore, for severe faults, Lee and Moray found that trust did not recover instantaneously.

Using these results, Lee and Moray (1992) computed a mathematical model of trust using an autoregressive moving average vector (ARMAV) time series analysis. Contrary to Muir's (1989) theory, Lee and Moray found that faith was most closely associated with trust, followed by dependability and predictability. Later testing indicated that trust alone did not predict whether participants relied on manual or automated control. Rather, participants' decision to use automation was best predicted by assessing the difference between their trust in automated control and their self-confidence in manual control (Lee & Moray, 1994).

These early theories of automation trust were significant landmarks. They proposed that human machine interaction was governed by principles similar human-human interaction. Over the years, researchers have tested, critiqued, and modified these theories to accommodate additional factors that influence automation trust. Next, the discussion reviews modern theories of automation trust. For the sake of brevity, two prominent trust theories are reviewed: Dzindolet, Pierce, Beck, Dawe, and Anderson's (2001) Utility Theory of Automation Trust, and Lee and See's (2004) Appropriate Trust Framework. These theories are most relevant to this review because they describe the cognitive and behavioral dimensions of automation trust, and emphasize the dynamic

relationship between an operator's expectations, biases, and cognitive processes that guide automation interaction and automation trust.

## Modern Theories of Automation Trust

*Utility theory of automation trust.* Dzindolet et al.'s (2001) Utility Theory of Automation Trust describes the cognitive, social, and motivational processes that influence automation trust. This theory states that, when formulating trust judgments, human operators compare the perceived reliability of automation to the perceived reliability of manual performance. The outcome of this comparison, called the perceived utility of automation, determines the level of automation trust. If the perceived utility is high, trust in automation will be high and automation dependence is expected. If the perceived utility is low, trust will be low and self-reliance is expected. Dzindolet and colleagues argue that the perceived utility will only be accurate if an operator knows the true reliability of automation and manual performance. Unfortunately, operators seldom know the reliability of either. Furthermore, biases can distort reliability estimates. One common bias that occurs when operators compare the utility of automation to manual control is the perfect automation schema (Dzindolet, Pierce, Beck, & Dawe, 2002). The perfect automation schema refers to the expectation that automation will perform perfectly. This bias causes operators to readily remember automation errors. As a result, operators underestimate the true reliability of automation and disuse it.

Dzindolet et al.'s (2001) model depicts two important phenomena. First, it notes that trust is appropriate when an operator knows the reliability of automated and manual performance. Therefore, communicating reliability information to an operator may foster appropriate utility estimates. Second, it states that preconceived expectations and

cognitive biases can affect trust judgments. Thus, perceptions of performance, as opposed to observed performance, may account for variance in trust judgments.

*Lee and See's appropriate trust framework.* Lee and See's (2004) Appropriate Trust Framework builds upon the theoretical work of Muir (1989) and Lee and Moray (1992, 1994) and describes the facets, dynamics, antecedents, and cognitive components of automation trust. Appropriate trust refers to the degree of match between the operator's trust and the true capability of automation (Lee & See, 2004).

Lee and See (2004) conceptualize appropriate automation trust according to three facets: calibration, resolution, and specificity. *Trust calibration* refers to the match between the operator's level of trust in the automated aid and the automated aid's accuracy. Trust calibration is essential for achieving appropriate dependence. If an operator's trust is not calibrated to the true accuracy of the system, he or she may misuse or disuse the system (Parasuraman & Riley, 1997). Automation misuse refers to over-reliance on automation, or automation complacency. This "overtrust" occurs because operators believe the automated system to be reliable and accurate than human performance. Automation disuse refers to under-reliance on automation. This "distrust" occurs because operators believe manual performance to be more reliable than automated performance. Disuse is associated with the cry-wolf effect, a phenomenon that results from frequent exposure to false alarms (see Bliss, 1993; Breznitz, 1984).

*Trust resolution* describes how precisely a judgment of trust corresponds to the automation's capabilities. Proper resolution is reflected when a range of system capabilities maps onto the same range of trust (Lee & See, 2004).

*Trust specificity* refers to trust in a particular component of automation. Trust specificity can be both functional and temporal. Functional specificity describes trust in sub-functions or modes of automation. An individual with high functional specificity will trust specific components of automation, whereas an individual with low functional specificity will only trust the capabilities of the overall system. Temporal specificity describes changes in trust as a function of changing situations or contexts (Lee & See, 2004). An individual with high temporal specificity will adjust his or her trust to match the capabilities of a system in different contexts and situations.

*Cognitive components of trust.* Lee and See (2004) note that trust calibration, resolution, and specificity are based on an accurate understanding of an automated aid's performance, process, and purpose. Information that forms the basis of these dimensions can be assimilated through three different cognitive processes, or tuning methods: analytical, analogical and affective methods. Each method uses different cues and cognitive processes to formulate trust judgments.

Trust developed via the *analytic method* involves observing system performance and deducing when the system performs reliably. This method is the most cognitively demanding because it depends on human reasoning and an accurate understanding of the automation's underlying motives and functions. Furthermore, it requires users to formulate accurate reliability estimates across a variety of situations and build trust according to these estimates (Cohen, Parasuraman, & Freeman, 1998).

Trust developed via the *analogical method* involves using observable cues, such as brand name, to infer a broad categorical membership and using these stereotypes to calibrate trust. This method is less cognitively demanding because it uses dispositional

features of automation, such as expertise and credibility, rather than performance features to assimilate trust. Trust based on consumer reviews or hearsay information is an example of tuning trust with analogical information (Lee & See, 2004).

*Affective methods* for trust development focus on emotional responses to automation rather than logic. For instance, Kim and Moon (1998) found that trust in online banking systems was influenced by surface level features of the website that produced positive affect such as coloring and text, rather than its actual banking capability. Affective methods for trust tuning are the least demanding because they use affect as a short cut to bypass cognitively demanding appraisal processes. The affective method also acts as a barrier because, if the user does not like the system, he or she may not use it enough to develop appropriate trust (Lee & See, 2004).

In review, Lee and See's (2004) framework indicates that appropriate trust is based on an accurate understanding of an automated system's performance, process and purpose. Information that forms the basis of these dimensions can be assimilated through three different methods. Each method focuses on different cues and requires a different level of cognitive processing. Lee and See's framework also offers a typology for conceptualizing appropriate trust; it conceptualizes trust according to calibration, resolution, and specificity. Theoretically, providing operators with information about automation capability should promote appropriate trust calibration, resolution, and specificity. However, few studies have empirically determined the manner in which operators use such information to calibrate trust.

The next section will review the empirical research concerning the effects of system confidence and automation expertise on trust and dependence. These two forms of

trust supporting information are of particular interest to this review because they each can influence automation trust and trust related behaviors through different cognitive processes.

**System Confidence and Trust**

Trust is theorized to be appropriate when an operator is cognizant of the capabilities and limitations of automation (Cohen et al., 1998). An operator can learn these parameters in a number of ways. For instance, experience with automation can provide operators with insight regarding automation reliability and performance (Parasuraman & Riley, 1997). Alternatively, if an individual is unfamiliar with a system, displaying information related to system capability can inform operators of system performance (Dzindolet et al., 2003). Indeed, several studies have found that explicitly providing operators with analytic information, such as system reliability, or displaying system confidence ratings can facilitate appropriate compliance. One notable example is Sorkin, Kantowitz, and Kantowitz's (1988) research on likelihood alarm displays (LAD). LADs use multi-level diagnostic signals to express the degree of certainty associated with a signal event. Essentially, LADs provide operators with insight regarding the likelihood that a signal event is true. To test the effectiveness of LADs, Sorkin et al. designed a study that required participants to concurrently perform a monitoring and a tracking task. Participants received support for the monitoring task from either an LAD or a traditional binary alarm system. The LAD generated graded alarms that were associated with different likelihoods of a true signal "signal" event. The binary alarm system generated alarms only when a "signal" event occurred. Sorkin et al found that participants who

interacted with the LAD performed better and allocated attention more appropriately than participants who interacted with the traditional binary display.

Other researchers have found similar results concerning the utility of LADs. For instance, St. John and Manes (2002) found that providing users with analytic information through a likelihood cuing system facilitated target search performance, even when the system was imperfect. More recently, McGuirl and Sarter (2006) examined the effect of dynamic system confidence information on trust and performance during a simulated aviation task. Their experiment required pilots to complete 28 simulated flight trials in icy conditions. The pilots' goal was to complete each trial without experiencing ice induced stalling. To help monitor ice conditions, selected participants received dynamic confidence information from an automated Smart Icing System (SIS). This system assessed reductions in aircraft stability and performance due to ice build-up. The system confidence display was modeled after Sorkin et al.'s (1988) LAD and provided pilots with confidence ratings in its own diagnostic capability. To assess the usefulness of this display, the researchers created two conditions: a fixed accuracy condition in which the reliability of the SIS system was 70% accurate, and an updated condition in which the system's confidence fluctuated among three levels: 89%, 50% and 25%. Results indicated that pilots who had access to dynamic confidence information made more effective flight decisions, more accurate estimates of system accuracy, and had fewer ice related stalls than pilots in the fixed condition. Further results demonstrated that compliance rates varied as a function of experimental condition. Specifically, participants in the fixed condition tended to over-comply with the aid's recommendation, whereas participants in the variable condition tended to comply more appropriately. McGuirl and

Sarter concluded that the availability of dynamic confidence information led to a significant improvement in trust calibration, which in turn, increased appropriate compliance and flight performance.

The results of the aforementioned studies suggest that communicating system confidence can facilitate appropriate compliance. However, to date, researchers have not explicitly tested the effects of system confidence on automation trust. McGuirl and Sarter (2006) assumed that compliance, a behavioral index of trust, was indicative of operators' trust in the system. As previously noted, trust is an attitude with behavioral consequences. Compliance is a behavioral state that frequently, but not always, relates to trust. Because compliance is not a direct measure of trust, it is not appropriate to assume causation in this case. A goal of this study was to empirically answer this question and assess the effects of system confidence on automation trust.

System confidence could influence automation trust in two ways. First, operators could observe system performance and deduce the system's accuracy for each level of system confidence. This resembles an analytic tuning method (Lee & See, 2004). Applying this tuning method to the interpretation of system confidence suggests operators would reason that (a) when system confidence is high, a signal event is likely and (b) when system confidence is low, a signal event is unlikely. In both instances, system confidence accurately portrays the state of the world. If operators appropriately trust both levels of confidence, then their compliance rates should match the system's level of confidence. This type of trust tuning strategy mimics the results reported by McGuirl and Sarter (2006). In their study, participants calibrated their compliance rates to

the system's level of confidence because, assumingly, they trusted system confidence

ratings appropriately.

Alternatively, operators could associate system confidence with the system's

diagnostic ability and base trust on the perceived ability of the system. This resembles an

analogical tuning strategy (Lee & See, 2004). Applying this tuning strategy to the

interpretation of system confidence suggests operators would deduce that high system

confidence reflects high diagnostic ability and low system confidence reflects poor

diagnostic ability. Assumingly, operators would be more likely to trust and comply with

the system when system confidence was high than when system confidence was low.

Such findings have been reported in the Social Psychology literature. Sniezek and Van

Swol (2001) had participants perform a decision-making task in which they solicited

advice from an expert advisor. Results indicated that advisor confidence had a positive

effect on trust and the tendency to follow the advisor's advice. Specifically, participants

complied and trusted the advice when advisor confidence was high.

**Automation Expertise and Trust**

As previously discussed, knowledge about automation capability can influence

trust and compliance. Similarly, expectations concerning system performance can also

influence automation trust. For instance, receiving hearsay information from a co-worker

that a system is error prone or reviewing a briefing about the capability of a new system

can effect expectations of system performance (Lee & See, 2004). One variable related to

user expectations that has generated considerable attention in the empirical literature is

advisor expertise. Expertise refers to the level of knowledge an advisor has about a topic

(Rhine & Severance, 1970). Social psychology research indicates that advisor expertise

influences relationships in two ways. First, individuals trust expert advisors more than novice advisors (Nan, 2007). Second, individuals are more likely to rely on advice from expert advisors than novice advisors (Sniezek & Von Swol, 2001). Of interest to this study was whether expertise demonstrates the same degree of persuasion and trustworthiness in human-machine relations as it does in interpersonal relationships.

Several researchers have empirically documented the effects of automation expertise on automation trust and dependence. A decade ago, Dijkstra (1999) found that individuals tended to over-rely on advice from expert systems. In his experiment, participants reviewed three law cases and made decisions concerning each suspect's sentencing. To help with each ruling decision, participants had access to the attorney's notes and advice from an expert computer system. The computer system analyzed the facts from each case and determined the best sentence. Unbeknownst to participants, the attorney's notes always contained the correct sentencing, whereas the computer system always gave incorrect advice. Dijkstra found that when making their sentencing decisions, participants over-relied on the advice from the expert computer system. He concluded that the advertised expertise of the system accelerated users' trust in the system, which led to over-reliance. Dijkstra also found that participants who relied on the expert system exerted less mental effort than participants who relied on the attorney's notes. These results suggest that participants used the expert system's advice as a means to bypass expending mental effort.

Dijkstra (1999) explained these results using the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1981). The ELM claims that advice seekers use two different paths when evaluating advice: the central route and the peripheral route. Individuals who

are highly motivated and confident in their ability to analyze the content of the advice use the central route. Individuals who are less motivated to attend to the content of the message use the peripheral route. Individuals using the peripheral route base their reliance decisions on surface level cues, like the advisor's presumed expertise, whereas individuals using the central route base their decision on more detailed cues, like message details and the situation. Dijkstra suggested that individuals who relied on counsel from the expert system used the peripheral route. Thus, he concluded that expertise influenced trust and reliance on the system. Unfortunately, Dijkstra (1999) did not compare advice acceptance across different levels of expertise. Therefore, it is hard to determine whether automation expertise truly influenced trust and advice acceptance. Nevertheless, Dijkstra's study highlights the persuasive role of analogical information, like expertise, in human-machine teams.

More recently, Madhavan and Wiegmann (2007) resolved the limitations of Dijkstra's (1999) study by examining the effects of source expertise on trust and dependence. In their study, participants completed 200 trials of a simulated luggage-screening task in which they were responsible for stopping baggage that contained contraband. To help with the task, participants received diagnostic advice from an advisor. The source (human vs. automation), expertise (novice vs. expert), and reliability (70% vs. 90%) of the advice varied across experimental groups. Madhavan and Wiegmann found participants trusted the expert advisor more than they trusted the novice advisor. Additional analyses revealed that when the automated advisor was 70% reliable, the expertise significantly influenced compliance and reliance. At the beginning of the task, participants' complied with the expert system more than the novice system.

However, these effects faded as participants learned the situational accuracy of the system. By the end of the task, participants actually complied with the novice system more than they complied with the expert system. This decline in expert system compliance reflects a rapid breakdown of the perfect automation schema (Dzindolet et al., 2002). Furthermore, Madhavan & Wiegmann (2007) found that automation expertise and reliability interacted to influence the participants' perceptions of system reliability. When the system was 90% reliable, participants perceived the expert aid to be more reliable than the novice system. However, when the system was 70% reliable, individuals actually perceived the novice system to be more reliable than the expert system. These results suggest that participants seemed to be more forgiving of novice errors than expert errors. These results also suggest that incorporating expertise characteristics into the design of automated systems actually heightens expectations and hinders trust calibration.

In a different study, Mayer (2008) examined the role of user expectations on automation trust, reliance, and compliance during a simulated warehouse management task. Expectations were manipulated by providing participants with a written description about an automated warehouse management system that framed likely performance of automation in terms of high, low, or standard performance. Mayer found that participants' preconceived expectations influenced automation trust and dependence; operators who expected the system to perform well trusted and depended on the system more than operators who expected the system to perform poorly. However, Mayer noted that the differences in automation dependence lasted only through the first session. At the conclusion of the experimental session, there were no differences in dependence between

participants in the high and low expectation conditions. This rapid decline in dependence also illustrates the breakdown of the perfect automation schema.

Merritt and Ilgen (2008) have also empirically documented the effect of automation expertise on trust and dependence. In their study, participants performed a simulated luggage-screening task while receiving diagnostic advice from an imperfect automation detection system. Prior to interacting with the system, participants read detailed instructions that described the system's competence, predictability, and dependability. Participants in the "high machine function" group were informed that the system was very accurate, performed predictability, and was very dependable. Participants in the "low machine function" group were informed that the system was inaccurate, unpredictable, and was prone to breakdowns. After participants finished the detection task, they completed a questionnaire that assessed subjective trust in the diagnostic system. Merritt and Ilgen found that trust predicted how often participants used the system. Furthermore, Merritt and Ilgen found that participants' perceptions of machine characteristics influenced trust more than actual machine characteristics. These findings represent major strides in automation trust research and collectively indicate how biases and perceptions about system characteristics can influence automation trust.

In summary, the literature cited above indicates that automation expertise affects trust and dependence. These results can be explained by Lee and See's (2004) appropriate trust framework. Analogical methods to trust development use categorical membership as a basis for trust. Information, such as automation expertise, influences operators' expectancies about likely system performance. Because expert systems are expected to perform more reliably and accurately than novice systems, operators calibrate their trust

to presumed levels of performance rather than actual performance capabilities. Furthermore, these expectances may bias information processing; causing operators to activity seek out information that confirms their expectations. Consequently, operators may judge imperfect expert systems more harshly than imperfect novice systems, particularly because expert errors violate users' expectation of perfect performance.

**Task Difficulty: The Effects of Image Quality on Automation Trust**

There is also evidence that automation trust and dependence are moderated by environmental variables, such as task difficulty. One of the fundamental reasons for introducing automation into complex task environments is to reduce workload and improve performance (Parasuraman & Riley, 1997). Therefore, one could assume that as task difficulty increases automation dependence will increase. Indeed, Wickens and colleagues have found that operators are more likely to depend on an automated cueing system when a task is difficult as opposed to when it is easy (see Wickens, Conejo, & Gempler, 1999; Yeh & Wickens, 2000).

Evidence also suggests that task difficulty can affect the conspicuity of automation errors and, in turn, affect automation trust. Madhavan, Wiegmann, and Lacson (2006) found that automation errors on tasks easily performed by humans were more damaging to automation trust than automation errors on difficult tasks. Maltz and Shinar (2003) found similar results in their study, in which participants performed a simulated target detection task with the aid of an imperfect automated cueing system. Specifically, these researchers found that automated cueing facilitated performance for difficult tasks, but hindered performance for easy tasks. In their experiment, task difficulty was manipulated by controlling image quality. Matlz and Shinar hypothesized

that automation impaired performance on easy tasks because it increased workload. However, there could be another explanation for these results. In the easy condition, automation errors were more salient because targets and automation miscues were easily recognized. After observing repeated automation errors, participants distrusted and disused the system (i.e., they relied more on manual performance even though automated performance was more reliable). Conversely, in the difficult condition, automation errors were not salient. Therefore, participants relied on the system to perform the task. Unfortunately, Matlz and Shinar did not measure operator trust; therefore, it is impossible to determine if image quality affected trust. A goal of this study was to address this limitation.

**Measuring Automation Trust and Dependence**

*Subjective measurement techniques.* To appropriately understand the relationship between trust in automated systems and use of these systems, researchers must be able to effectively measure trust (Jian, Bisantz, & Drury, 2000). Many researchers have used self-report scales to measure automation trust. For instance, Singh, Molloy and Parasuraman (1993) developed the Complacency Potential Rating Scale (CPRS) to ascertain people's attitudes toward automation. Lee and Moray (1994), as well as Muir and Moray (1996), have measured automation trust with multi-item questionnaires. Other researchers, such as Sanchez (2006) and Brown and Galster (2004) have measured trust in automated systems with a single item indicator.

Despite the wide use of subjective measures in automation trust research, their validity has rarely been the subject of focused investigation and is often not assessed beyond internal consistency reliability. One of the more validated measures for assessing

automation trust is the System Trust Scale (Jian et al., 2000). The System Trust Scale was developed over the course of a three-phase experiment comprised of a word elicitation study, a questionnaire study, and a paired-comparison study (Jian et al., 2000). The word elicitation study required participants to provide written descriptions of their understandings of trust and distrust with respect to trust in people, automation, and trust in general. In the second phase, participants rated the extent to which trust and distrust were similar with respect to trust in people, automation, and general trust. Results from the questionnaire study indicated that trust and distrust were correlated. In the third phase, participants completed a paired comparison study. The results of these efforts produced a 12-item scale with two subscales for trust and distrust.

Since its initial development, several researchers have used revised or abbreviated versions of the System Trust Scale. For instance, Fallon, Bustamante, Ely, and Bliss (2005) used a 10-item modified version of the scale to assess operator trust in alarm systems. Results showed that the internal consistency of the modified scale was slightly higher than original scale, yielding an internal consistency reliability of $\alpha = .93$. Safar and Turner (2005) used a revised 12-item scale to measure trust in two different Internet banking websites. Their evaluation found that the System Trust Scale demonstrated high internal consistency and convergent validity. Spain and Bliss (2008) used a revised 12-item System Trust Scale to measure trust in sonification systems. Similar to Fallon et al. (2005), their evaluation found that two of the original items did not offer sound psychometric fit with the underlying factor structure. The resulting 10-item scale yielded a high internal consistency reliability ($\alpha = .91$). More recently, Spain, Bustamante, and Bliss (2008) performed a psychometric evaluation of the System Trust Scale to assess its

construct validity. Using structural equation modeling, these researchers found that the scale accurately measured trust and distrust, and that these constructs were distinct, yet related, factors.

*Behavioral measurement techniques.* In addition to subjective measurement techniques, many researchers use behavioral indices of dependence to gauge automation trust. The assumption is that trust is an attitude with behavioral consequences; thus, if an operator trusts automation he or she will rely on it.

Meyer (2004) suggests that researchers should dichotomize automation dependence into *reliance* and *compliance*. Reliance refers to the action an operator takes when automation identifies the state of the world as being "all is well." Compliance refers to the action the operator takes when automation identifies a "signal event" in the world. In target detection paradigms, compliance refers to an operator responding 'target present' when the automation says 'target present'. Conversely, reliance is demonstrated when an operator responds 'target absent' when automation indicates 'target absent'. Ideally, an operator's strategy to comply and rely on automation will match the accuracy of the automated aid. Compliance and reliance are appropriate strategies when automation generates a correct response (hit or correct rejection). Conversely, compliance and reliance are poor strategies when automation errs.

The distinction between reliance and compliance is important for several reasons. First, evidence suggests that false alarm prone automation will influence an individual's decision to comply with automation whereas miss prone automation will influence an operators decision to rely on automation (Rice, *in press*). Thus, system errors have

differential effects on strategy adoption. Second, reliance and compliance are behavioral indices that may depend on features of automation such as confidence or expertise.

*The timing of subjective trust assessment.* One concern regarding the assessment of automation trust is the timing of trust measurement. Many researchers choose to measure trust subjectively, either prior to, or after, participants interact with a system. Though recording trust in this manner provides an accurate account of overall system trust, it does not fare well for measuring the dynamics of trust. Other researchers assume that compliance and reliance are indicative of trust and can therefore be used as a measure of trust. However, as stated above, reliance and compliance are behaviors that are guided by, but not always related to, trust. Therefore, it is inappropriate to assume that compliance and reliance are indicative of trust.

To examine the dynamics of trust, researchers must measure trust intermittently over the course of the task. To date, few researchers have measured trust subjectively on a trial-by-trial basis, thus providing a micro-scale assessment of automation trust. Part of the reason so few researchers have examined trust in this manner is because it is difficult and impractical to administer a 12-item measure such as the System Trust Scale after each experimental trial. Furthermore, many researchers fail to differentiate between trust as an attitude and trust as a behavior. One of the goals of this study was to document trust, subjectively and behaviorally, across time.

## Review and Limitations in Previous Research

In review, trust is an important psychological construct that mediates human interaction with automation. Trust largely depends on perceptions of automation capability (Sheridan & Parasuraman, 2006). Therefore, trust and dependence are more

likely to be appropriate when operators have information about an automated system's capability. Lee and See (2004) identify three methods by which operators can assimilate trust related information: analytic, analogical, and affective methods. Empirical evidence suggests that displaying system confidence can promote appropriate dependence (McGuirl & Sarter, 2006). Evidence also suggests that automation expertise can influence operators' expectancies regarding system performance, which ultimately can affect trust and dependence (Madhavan & Wiegmann, 2007; Mayer, 2008; Merritt & Ilgen, 2008).

To date, few studies have examined the influence of system confidence on subjective trust. Such research is important for several reasons. If system confidence ratings are proposed to indicate situational accuracy of the automated aid's capability, one could postulate that communicating this information to an operator would promote appropriate trust and dependence (Lee & See, 2004). Furthermore, preconceived biases could influence the manner in which operators interpret confidence ratings from expert and novice systems. In such a case, the expertise and confidence of a system could influence trust and compliance. For example, in the context of military target detection, a soldier may trust and depend on a highly confident expert targeting system more than a highly confident novice targeting system, even though the actual reliabilities of the two systems are the same. Unfortunately, the combined effects of system confidence and automation expertise on trust and compliance are unknown. Theoretically, understanding these effects would fill a void in the existing literature and provide valuable insight for developing a model of automation trust and decision-making in complex environments.

A second limitation with current research is that few studies have examined the influence of system confidence and automation expertise on target detection

performance. Such research is important for several reasons. There are currently several automated systems, such as automated target recognition (ATR) systems, that provide operators with diagnostic confidence ratings (Sterling & Jacobson, 2006). Ideally, providing operators with system confidence would minimize false alarms and increase detection performance. However, few studies have empirically examined the benefit of system confidence from a performance standpoint. Understanding the benefits or drawbacks of incorporating system confidence ratings into automated detection systems is critical for the refinement and development of future systems. Furthermore, expectations regarding system performance could influence an operator's decision bias. That is, an operator may be more willing to comply with an expert system than a novice system because of preconceived biases concerning the performance standard of an expert system. Consequently, operators may be more prone to false alarms when interacting with expert systems. Currently, the design of expert systems and their effects on performance is not under study. This is a limitation because it is not only the quality of the underlying algorithms that guides human-automation performance (Sorkin & Woods, 1985), but how the information is rendered and communicated to the operator. A goal of this study was to empirically document the effects of automation expertise and system confidence on task performance.

A third limitation with existing research relates to the manner in which researchers have measured trust. Existing studies have assumed that compliance, a behavioral index of trust, is indicative of operator trust. However, trust is an attitude that influences, but does not completely determine, compliance. Furthermore, the few studies that have assessed subjective trust have measured trust either prior to, or after, interacting

with the automated system. While this measurement technique may provide valuable information concerning trust calibration, it fails to capture the dynamics of trust. To capture the dynamics of trust, researchers must measure trust intermittently. Recording trust on a trial-by-trial basis should allow valuable insight into the development and maintenance of trust. Furthermore, measuring trust on a trial-by-trial basis would allow the assessment of system confidence and automation expertise on automation trust over time.

Finally, existing research has not determined how system confidence and automation expertise support trust across different levels of image quality. There may be instances in which automation expertise or system confidence emerges as an important factor that affects automation trust and compliance. For instance, in the context of military target detection, soldiers may blindly follow the advice of a diagnostic aid if the aid has a reputation for being state-of-the-art, especially when target detection is difficult. Such heightened expectations may also have a dramatic effect on automation usage decisions, especially when the aid errs. Dzindolet et al. (2002) reports that dissonance between alleged expertise and actual performance standards may cause operators to abandon automation, even when aided performance is statistically more accurate than manual performance. Therefore, it is important to determine how image quality influences trust and compliance.

**Purpose of Current Study**

The purpose of the current study was to determine the effects of system confidence, automation expertise, and image quality on trust, compliance, and performance. Previous research on automation trust suggests a conceptual model where

information pertaining to automation capability, such as automation expertise and system confidence, influences trust and compliance (see Figure 1). In that model, operators appraise system and task information and filter it through cognitive biases. The resultant appraisal of automation capability and manual capability influences automation trust and compliance. Based on this model and previous research, the following hypotheses were tested.
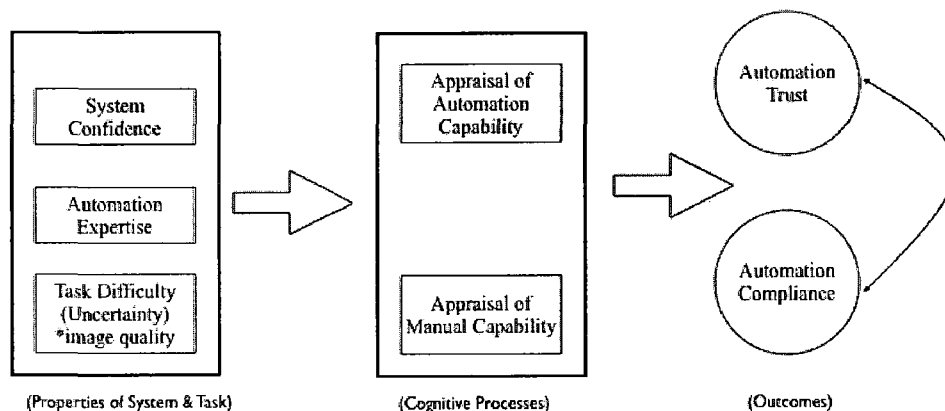


*Figure 1.* Model depicting the effects of system confidence, automation expertise and image quality on trust and compliance.

**Trust Hypotheses**

*Main effect of system confidence on trust (Hypothesis 1).* Trust is theorized to be influenced by perceptions of automation capability (Lee & See, 2004). Thus, cues such as system confidence, may affect trust. As such, I hypothesized that system confidence would positively impact trust ratings. This effect was predicted to manifest itself in a significant main effect of system confidence for diagnostic trust.

*Interaction between system confidence, automation expertise, and image quality on trust (Hypothesis 2).* System confidence and automation expertise reflect two sources of trust supporting information. Because these two processes do not exist in a vacuum, I expected a significant interaction. Specifically, I hypothesized that participants would weigh diagnostic confidence from expert and novice systems differently (Sniezek & Von Swol, 2001). Moreover, these differential weighting effects were predicted to be more salient when image quality was low than when image quality was high. These effects were predicted to manifest in a significant interaction of system confidence, automation expertise and image quality on trust.

*Interaction between automation expertise, system confidence and time on trust (Hypothesis 3).* Dzindolet et al.'s (2002) notion of the breakdown of the perfect automation schema suggests that automation errors significantly decrease perceptions of system accuracy, which can cause mistrust. This loss of trust may be more evident in "expert systems" because of the heightened expectation surrounding their performance standards. Indeed, previous research has indicated that compliance decreases as operators learn a system's true accuracy and that this decrease is more evident for expert systems than novice systems (Madhavan & Wiegmann, 2007; Mayer, 2008). However, researchers have yet to explicitly measure this change in trust. Based on previous research, I expected trust in the expert system to decrease more rapidly than trust in the novice system. However, I also expected system confidence ratings to moderate this effect. Specifically, I expected the system confidence ratings to mitigate the loss of trust associated with expert errors particularly when system confidence was high. This hypothesis was rooted in previous research by McGuirl and Sarter (2006) and Bliss and

Dunn and Fuller (1995) who demonstrated that confidence information can induce compliance with diagnostic system advice and mitigate the loss of trust associated with automation errors. This effect was predicted to manifest in a significant interaction of system confidence, automation expertise and session on trust

**Compliance Hypotheses**

*Interaction between automation expertise, system confidence, and image quality on compliance (Hypothesis 4).* Based on the literature and similar to Hypothesis 2, I expected participants to calibrate their compliance to the system's level of confidence (McGuirl & Sarter, 2006). I also expected an interaction between system confidence, automation expertise, and image quality. Specifically, I predicted that participants would comply more often with the expert system than the novice system when system confidence was high. I further hypothesized that this interaction would be greater when image quality was low than when image quality was high.

*Interaction between automation expertise, system confidence, and time on compliance (Hypothesis 5).* Previous research indicates that compliance with advice from expert systems decreases when the aid errs (Madhavan & Wiegmann, 2007). This change is assumed to be a direct result of a breakdown in participants' perfect automation schema (Dzindolet et al., 2002). Therefore, I expected to find a significant interaction between session and automation expertise on compliance. Specifically, I expected compliance with the expert system to decline more rapidly over the course of the experiment than compliance with the novice system. Similar to the trust hypothesis, I also expected system confidence to mitigate the loss of compliance with expert errors.

**Performance Hypotheses**

*Detection sensitivity hypotheses.* I expected participants to be more accurate when they received automated assistance than when they performed the task manually. I also expected system confidence and image quality to interact and significantly influence detection sensitivity. Specifically, I expected participants to correctly identify more targets when image quality was high and the automated system was highly confident that a target was present than when image quality was high and the system was not confident that a target was present.

*Response bias hypotheses.* With regard to detection bias, I expected participants to more bias (i.e., indicate that a target was present more often) when they interacted with the expert system. Conversely, I expected participants to be more conservative (i.e., indicate that a target was present less often) when they interacted with the novice system. I also expected participants to be more liberal when image quality was low as compared to when image quality was high. Furthermore, I expected participants to adopt a liberal bias on trials in which the system was highly confident a target was present, and a conservative bias on trials in which the system was not confident that a target was present.

# METHOD

## Experimental Design

A 2 (Automation Expertise: expert, novice) x 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Image Quality: high, low) x 3 (Session: 1, 2, 3) mixed subjects factorial design was used for this experiment. In addition, a single control group was added to the experimental design. Automation expertise was manipulated between subjects so that operators interacted with a single system over the course of the experiment. System confidence, image quality, and session were modeled as within subject variables. System confidence varied randomly on a trial-by-trial basis within each experimental session. The experimental design was crossed using a Latin-square design. Dependent variables included initial trust, diagnosis trust, overall system trust, perceived system reliability, automation compliance, and detection performance. Detection performance was assessed using signal detection theory measures of sensitivity ($d'$ *prime*) and bias ($c$).

## Participants

One hundred and fifteen undergraduate students from Old Dominion University, selected through convenience sampling, participated in this study. According to Keppel and Wickens' (2004) sample size calculation formula, this sample size achieved an experimental power of .80 with a medium effect size of .25. Participants received one and a half extra credit points for participating. Participants were at least 18 years of age ($M$ = 22.22, $SD$ = 5.96) and had normal or corrected-to-normal vision. Of the 115 participants, 42 were male and 73 were female. All participants were treated in accordance with the American Psychological Association's "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002).

**Materials and Apparatus**

*Informative flyer.* A flyer that described the purpose of the study was posted on Old Dominion University's online research participation system, SONA™, to advertise the study (see Appendix A).

*Participant background information form.* Participants completed a background information form that assessed demographic information such as age, class status, computer usage, and opinions concerning automation. The form also asked questions pertaining to participants' visual deficiencies (see Appendix B).

*Experiment instructions.* The experiment instructions described the nature of the study and provided participants with a description of the automated system that assisted them during the experiment (see Appendix C).

*Post instruction questionnaire.* A post instruction questionnaire was used to ensure that participants understood the expertise, purpose, development history, and performance standards of the automated system that assisted them during the experiment (see Appendix D). Participants were required to correctly answer each question before proceeding to the data collection phase.

*Trust questionnaire.* A modified version Jian et al.'s (2000) System Trust Scale (see Appendices E & F) was used to assess the level of trust participants maintained in the automated diagnostic system during the experiment. The System Trust Scale contained 12 items, seven of which were intended to assess trust and five of which were intended to assess distrust in automated systems. Each item was measured on a 7-point Likert-type scale.

*Diagnosis trust.* Diagnosis trust was operationally defined as participants' trust in the system's diagnosis. This variable was measured after each trial using a single item indicator that ranged from 1 (not at all) to 5 (very much so) on a Likert-type scale. Specifically, the item asked, "Do you trust the system's diagnosis?" Because participants did not receive diagnostic assistance on "no aid" trials, diagnostic trust was not assessed during those trials.

*Opinion questionnaire.* Upon finishing the experiment, participants completed an opinion questionnaire. The opinion questionnaire assessed subjective levels of stress, image quality, and task stimulation. Participants also reported strategies they adopted for completing the detection task (see Appendix G).

*Apparatus.* A simulated military target detection scenario served as the primary task. The scenario was created using Visual Basic 6.0™. The simulation was hosted on IBM compatible 3.20 GHz Intel Pentium D computers. Each computer had a 17-inch monitor. The scenario required participants to search for covert enemy targets in two types of images, high quality and low quality images. During the task, participants received diagnostic advice from an automated target detection system that varied in expertise (expert vs. novice) and confidence (75%, 50%, 25%, no aid). As shown in Figure 2, targets (i.e., foes) always faced left; friendly objects always faced right. Approximately half of the images contained an enemy target. The type (i.e., Hum-V, Soldier, or Tank) and placement of the target randomly varied within the images.
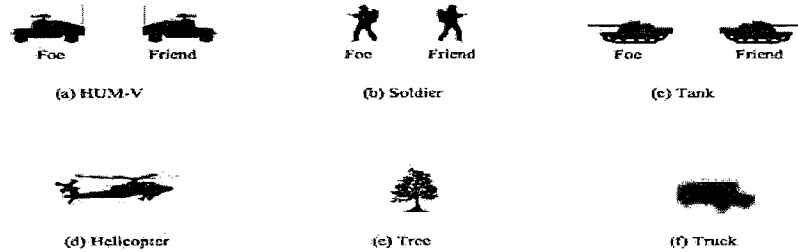
*Figure 2.* Experimental stimuli: (a) HUM-V, (b) soldier, (c) tank, (d) helicopter, (e) tree, (f) truck.

## Experimental Manipulations

*Automation expertise.* Automation expertise was manipulated by providing participants with written descriptions concerning the automated system's technology, development history, system specifications, testing, and performance capability (see Table 1). Previous researchers have used a similar technique to manipulate automation perceptions of automation capability (see Madhavan & Wiegmann, 2007; Mayer, 2008). To ensure that the descriptions were comparable in construction and clarity, twenty participants were recruited to read each dimension (i.e., technology history, system specification, etc.) and rated the comparability of the language, grammar, sentence construction, and length between both systems. Ratings ranged from 1 (not comparable) to 5 (completely comparable). Then, participants read each description in its entirety and

rated the "credibility" of each system using a seven-point scale ranging from 1 (not at all credible) to 7 (very credible).

**Table 1:** System characteristics for novice and expert systems

| Dimension | SYSTEM 1 | SYSTEM 2 |
|---|---|---|
| Introduction | Today you will perform a target detection task and will receive advice from CONTRAST DETECTOR. | Today you will perform a target detection task and will receive advice from SUPER CONTRAST DETECTOR. |
| System technology | CONTRAST DETECTOR is an automated diagnostic aid that has been designed to identify military targets in synthetic aperture radar (SAR) images. CONTRAST DETECTOR is based upon technology used in military target detection over the past 10 years. | SUPER CONTRAST DETECTOR is an automated diagnostic aid that has been designed to identify military targets in synthetic aperture radar (SAR) images. SUPER CONTRAST DETECTOR is based upon, but far exceeds, technology used in military target detection over the past 10 years. |
| Development history | CONTRAST DETECTOR was designed and developed at a small technical college in the Midwest that contains a small department in military target detection. | SUPER CONTRAST DETECTOR was designed and developed by the nations top military research firm in Washington D.C. that contains a highly specialized department in military target detection. |
| System specifications | CONTRAST DETECTOR currently possesses a limited database of the types of modern weapons and targets commonly found in today's military operations. Its algorithms are relatively ineffective in their attempts to detect enemy targets. | SUPER CONTRAST DETECTOR currently possesses an extensive database of the types of modern weapons found in today's military operations. Its algorithms are highly effective in their attempts to detect enemy targets. |
| System testing | Recent testing indicates that the accuracy, dependability, and robustness of CONTRAST DETECTOR do not meet the standard for military target detection systems. | Recent testing indicates that the accuracy, dependability, and robustness of SUPER CONTRAST DETECTOR set the standard for military target detection systems. |
| Expected system performance | The U.S. Department of Defense is considering whether to conduct limited field-testing using CONTRAST DETECTOR. | The U.S. Department of Defense is currently using SUPER CONTRAST DETECTOR in its Middle Eastern military operations. |

Pilot study results indicated that all comparisons received average comparability ratings of 4.0 or higher. The only exception was the dimension 'System Technology' which received an average comparability rating of 3.1. Consequently, this description was modified. Following the modification, ten additional participants read and rated the system descriptions. Analyses indicated that each dimension averaged at least 4.0 on comparability. Because each system varied only in their description, not their performance capability, automation expertise provided analogical information regarding automation capability.

*System confidence.* System confidence was manipulated by providing participants with four levels of diagnostic confidence: 75%, 50%, 25% and no aid. Participants were informed both verbally and in the written instructions that the confidence estimates were based on how well the information collected from the system's detection algorithms matched the enemy template located in the system's target database. They were also informed that higher confidence estimates were associated with a higher probability of a target being present. System confidence ratings were presented numerically and graphically. A 75% rating was displayed with a red bar three-fourths the size of the horizontal indicator with the rating superimposed in black font (Figure 3). A 50% rating was displayed with an orange bar, one-half the size of the horizontal indicator, with the rating superimposed in black font. A 25% rating was displayed with a yellow bar, one-fourth the size of a horizontal indicator, with the rating superimposed in black font. On trials with no aid, participants did not receive diagnostic advice from the system. Each participant received 24 high confidence, 24 neutral confidence, 24 low confidence trials,

and 24 no aid trials. In these four conditions, a target was present on 75%, 50%, 25%, and

50% of the trials, respectively. These levels were chosen for two reasons.
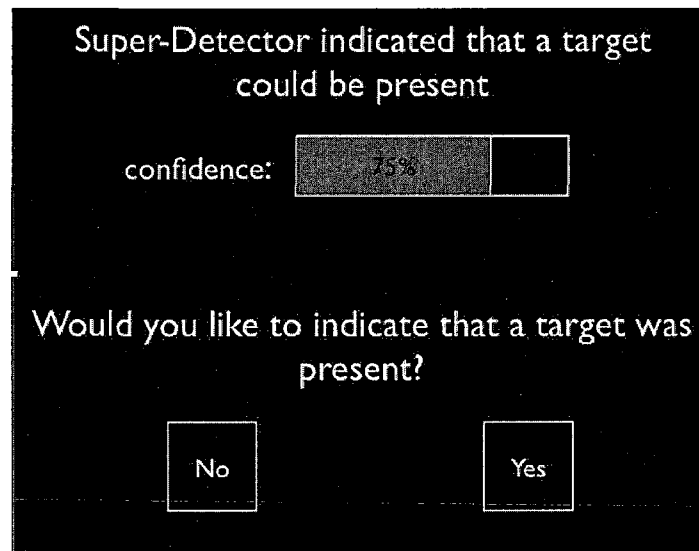


*Figure 3.* Screen-shot of simulation interface.

First, as McGuirl and Sarter (2006) noted, using a range of confidence estimates

assures that there is a perceptually distinct difference in performance between each level

of confidence. Second, current technologies, such as automated target recognition (ATR)

systems, provide operators with a range of confidence estimates. Therefore, the levels in

the current study were chosen to increase the ecological validity of the simulation.

Because each level of system confidence was associated with a unique probability of a

target being present, system confidence provided analytical information regarding the

capability of the detection system.

*Image quality.* Image quality was manipulated as an independent variable by introducing random pixel noise into the simulated SAR images using Adobe® Photoshop® CS3 Expanded version 10.0.1 for Macintosh. This program allows users to increase image distortion from 10% to 400% in magnitude. Two types of images were created, moderately distorted and severely distorted, which coincided with noise levels of 100% and 200%, respectively. Entin, Entin, MacMillan, and Serfaty (1995) and MacMillan, Entin, and Serfaty (1994) each used a similar technique to control image noise in their target detection research. Additionally, the distortion levels replicated those of real SAR images, thus promoting the ecological validity of the simulation. Figures 4a and 4b show an image with distortion rates of 100% and 200%, respectively.



(a) High image quality                    (b) Low image quality

*Figure 4.* Image distortion percentages (a) 100% and (b) 200%.

Results from pilot testing indicated that target detection was more difficult for the severely distorted images ($M = 105.92$, $SD = 13.01$) than the moderately distorted images
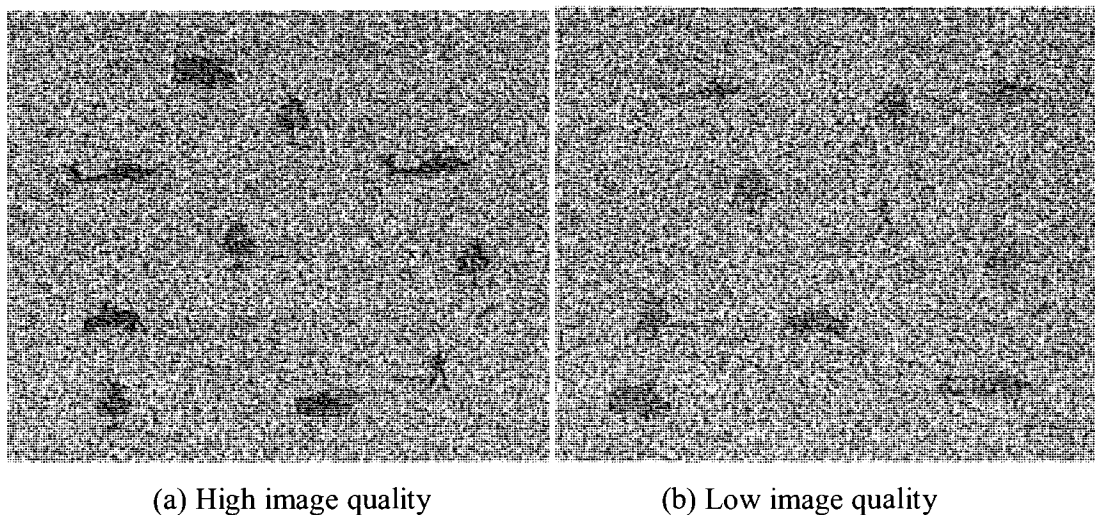
($M$ = 115.92, $SD$ =19.88), $F(1, 25)$ = 5.09, $p$ < .033. Therefore, I categorized the two

picture types as being representative of two levels of image quality; classifying the

severely distorted pictures as the "low image quality" and the moderately distorted

pictures as the "high image quality".

**Tasks and Procedure**

During the scenario, participants played the role of a military inspector who had

to search simulated SAR images for enemy targets. At the onset of each trial, the SAR

image appeared on the screen for 1000ms. After the image disappeared, aided

participants received diagnostic advice from the detection system in the form of a text

message. This message included the system's diagnosis and confidence estimate

concerning the presence of an enemy target. As previously noted, participants were

informed that the confidence estimates were based on how well the information collected

from the system's detection algorithms matched the enemy template located in the

system's target database. Higher confidence estimates were associated with a higher

probability of a target being present. It is important to note that the system always

indicated that a target could be present; only the system's confidence concerning the

likelihood of a target varied. After reviewing the system's diagnosis, participants

indicated whether they thought a target was present. Participants clicked the "Yes, Target

Present" icon or the "No, Target Not Present" icon. Then, participants reported their

decision confidence on a Likert-type scale that ranged from 1 (no confidence) to 5 (very

confident). Participants also reported their trust in the system's diagnosis using Likert-

scale that ranged from 1 (not at all) to 5 (very much). After making their ratings,

participants received feedback concerning the accuracy of their decision. Detection

performance was also scored. Participants received one point if they correctly identified a true target or correctly rejected a false target. Conversely, participants lost one point if they missed a true target or if they responded to a false target. Pilot testing and past research has indicated that participants are highly motivated to respond quickly, accurately, and appropriately when presented with this payoff matrix (Bliss, 2000).

When participants arrived, they were randomly assigned to an experimental condition. Each participant completed a background information form and signed the informed consent. Then, they logged onto separate computers hosting the experimental simulation and read the task instructions. The instructions included an introduction to the target detection task, a description of the automated detection system, and a description of how the system generated its confidence estimates. Next, participants completed the post-instruction questionnaire to ensure they understood the forthcoming task and the system that was assisting them. Participants also completed a modified version of the System Trust Scale to assess their initial trust in the diagnostic system. Afterwards, participants completed several practice trials. Then, experimental testing commenced.

After completing the first session (96 trials), participants completed the System Trust Scale to report their trust in and their perceived reliability of the automated system. Then, participants took a short comfort break. The second session followed the same procedures as the first session; however, the image quality changed. Participants who viewed high quality images in the first session, viewed low quality images in the second session, and vice versa. At the end of the second session, participants completed the System Trust Scale and an opinion questionnaire. Then they were debriefed. Participants completed both sessions (i.e., 192 trials) in approximately one hour.

# RESULTS

Prior to computing inferential statistical analyses, I screened the data set for missing data, unequal sample sizes, and outliers. Nine cases were removed because the data were deemed invalid. The response patterns indicated that the participants responded inappropriately; they always indicated that a target was present or absent. These nine cases were removed resulting in a sample of 106 participants. Descriptive statistics for each variable were computed to ensure that the statistical assumptions for each analysis were not violated. Means and standard deviations for trust, compliance, sensitivity and bias are presented in Table 2.

Prior to calculating inferential analyses I computed an analysis to ensure that target base rate did not confound the effects of system confidence. Results indicated that target base rate did not influence response patterns. Greater details of these results are reported in Appendix H.

Hypotheses were tested via planned comparisons and by building custom analysis of variance (ANOVA) models in SPSS version 17.0 for Macintosh. Building custom models simplifies the experimental design and preserves power by isolating relevant portions of the data and testing specific family comparisons (Keppel & Wickens, 2004). I addressed violations of homogeneity of variance by using a more stringent alpha level. Violations of sphericity were addressed by using Greenhouse Giesser $F$ value (Tabachnick & Fidell, 2001). After testing the hypotheses, exploratory analyses were calculated to examine addition relationships among variables. Post hoc analyses for quantitative within subjects variables were addressed via trend analyses. Post hoc analyses for qualitative within subjects variables were addressed via simple contrast analyses. Unless otherwise noted, all analyses were computed using a critical value of $\alpha$ = .05.

**Table 2:** Means and Standard Deviation for Trust, Compliance, Sensitivity, and Bias as a Function of Image Quality, System Confidence, and Automation Expertise

| | High Image Quality | | | | | | Low Image Quality | | | | | |
| | 25% | | 50% | | 75% | | 25% | | 50% | | 75% | |
| | M | SD | M | SD | M | SD | M | SD | M | SD | M | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Diagnostic Trust** | | | | | | | | | | | | |
| Expert | 2.73 | 0.67 | 2.79 | 0.59 | 3.40 | 0.61 | 2.61 | 0.69 | 2.74 | 0.67 | 3.24 | 0.85 |
| Novice | 2.41 | 0.69 | 2.47 | 0.63 | 2.91 | 0.75 | 2.21 | 0.69 | 2.28 | 0.66 | 2.74 | 0.76 |
| Total | 2.57 | 0.70 | 2.63 | 0.63 | 3.15 | 0.72 | 2.41 | 0.71 | 2.51 | 0.70 | 2.99 | 0.84 |
| **Compliance** | | | | | | | | | | | | |
| Expert | 0.36 | 0.17 | 0.51 | 0.15 | 0.73 | 0.21 | 0.57 | 0.19 | 0.41 | 0.22 | 0.64 | 0.18 |
| Novice | 0.30 | 0.11 | 0.48 | 0.15 | 0.68 | 0.16 | 0.53 | 0.17 | 0.39 | 0.19 | 0.51 | 0.15 |
| Total | 0.33 | 0.15 | 0.50 | 0.15 | 0.71 | 0.18 | 0.55 | 0.18 | 0.40 | 0.20 | 0.58 | 0.18 |
| **Sensitivity ($d'$ *prime*)** | | | | | | | | | | | | |
| Expert | 0.63 | 0.75 | 0.49 | 0.67 | 0.43 | 0.64 | -0.02 | 0.63 | -0.02 | 0.53 | 0.05 | 0.45 |
| Novice | 0.95 | 0.81 | 0.79 | 0.73 | 0.48 | 0.71 | -0.18 | 0.53 | 0.04 | 0.54 | 0.07 | 0.62 |
| Total | 0.79 | 0.79 | 0.64 | 0.71 | 0.45 | 0.67 | -0.10 | 0.59 | 0.01 | 0.53 | 0.06 | 0.54 |
| **Bias ($C$)** | | | | | | | | | | | | |
| Expert | 0.26 | 0.45 | -0.02 | 0.41 | -0.61 | 0.71 | -0.94 | 0.54 | -0.40 | 0.53 | 0.26 | 0.63 |
| Novice | 0.35 | 0.31 | 0.08 | 0.44 | -0.42 | 0.56 | -0.71 | 0.59 | -0.04 | 0.38 | 0.34 | 0.52 |
| Total | 0.30 | 0.39 | 0.03 | 0.43 | -0.51 | 0.65 | -0.83 | 0.57 | -0.22 | 0.49 | 0.30 | 0.58 |

Note: N = 80

**Testing the Predicted Effects**

*Hypothesis 1*. I hypothesized that system confidence would significantly impact

trust ratings. To discern this effect, I computed a repeated contrast analysis using system

confidence as the predictor and diagnostic trust as the criterion. Because the comparisons

were not mutually orthogonal, I used a bonferroni correction to control for Type I family

wise error rates (Keppel & Wickens, 2004; p. 115). All interferences were made using an

$\alpha$ = .025. As shown in Figure 5, participants trusted the system significantly more when it

was 75% confident ($M$ = 3.07, $SD$ = .85) than when it was 50% confident ($M$ = 2.57, $SD$

= .72), $F(1, 79)$ = 107.15, $p$ = .001, partial $\eta^2$ = .58. The difference between trust ratings

when the system was 50% confident ($M$ = 2.57, $SD$ = .72) and 25% confident ($M$ = 2.49,

$SD$ = .77) failed to reach significance, $F(1, 79)$ = 4.72, $p$ > .025.
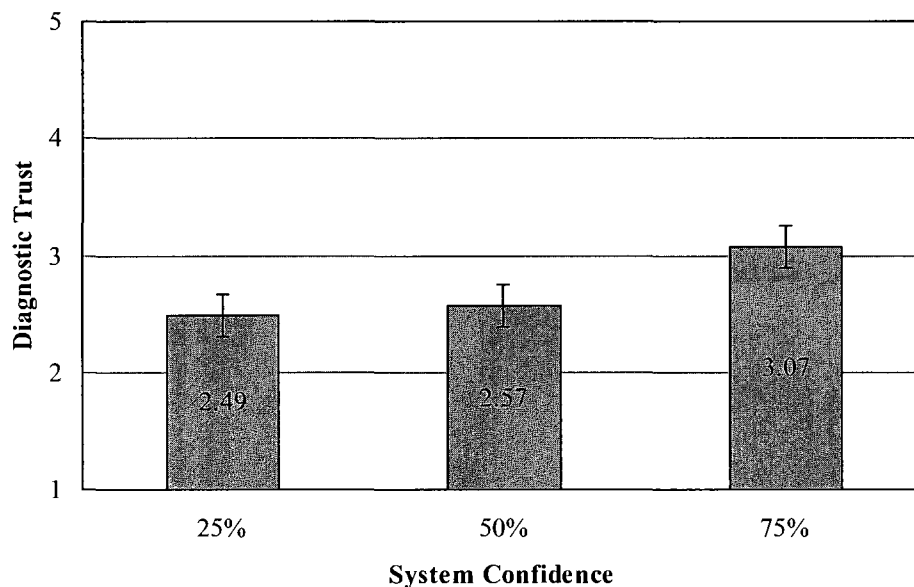


*Figure 5*. Diagnostic trust as a function of system confidence.

*Hypothesis 2.* A custom 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation

Expertise: expert, novice) x 2 (Image Quality: high, low) mixed factorial ANOVA was

calculated to test the hypothesized interaction of system confidence, automation

expertise, and image quality on trust (note: the variable 'session' was included in the

model, but mathematically operated as a control variable). The assumption of sphericity

was violated for several of the within-subjects variables, therefore all interpretations were

made using the Greenhouse-Geisser $F$ value. The predicted interaction failed to reach

significance, $F(1.78, 139.01) = .91, p > .05$. However, the main effects for system

confidence, $F(1.77, 138.05) = 91.45, p = .001$, partial $\eta^2 = .54$, image quality, $F(1, 78) =$

$10.32, p = .002$, partial $\eta^2 = .12$, and automation expertise, $F(1, 78) = 9.49, p = .003$,

partial $\eta^2 = .11$, were statistically significant. Post hoc analyses indicated a significant

linear trend for system confidence, $F(1, 78) = 122.88, p = .001$; trust increased linearly as

system confidence increased (refer to Figure 5). The main effect of image quality

indicated that participants exhibited more trust in the system when image quality was

high ($M = 2.78, SD = .76$) than when image quality was low ($M = 2.63, SD = .81$). The

main effect of automation expertise indicated that participants trusted the expert system

($M = 2.92, SD = .76$) more than the novice system ($M = 2.50, SD = .76$).

*Hypothesis 3.* A custom 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation

Expertise: expert, novice) x 3 (Session: 1, 2, 3) mixed factorial ANOVA was calculated

to test the hypothesized interaction of system confidence, automation expertise, and time

on trust (note: the variable 'image quality' was included in the model, but mathematically

operated as a control variable). The assumption of sphericity was violated for the within-

subjects variables, therefore all interpretations were made using the Greenhouse-Geisser

$F$ value. The predicted interaction failed to reach statistical significance, $F(3.75, 292.7) =$ 1.50, $p > .05$. However, the omnibus ANOVA for session was statistically significant, $F(1.79, 139.84) = 3.91$, $p = .02$, partial $\eta^2 = .05$. Post hoc trend analysis indicated a significant linear trend, $F(1,78,) = 5.56$, $p = .02$, partial $\eta^2 = .07$; trust declined over the course of the experiment.

*Hypothesis 4.* A custom 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation Expertise: expert, novice) x 2 (Image Quality: high, low) mixed factorial ANOVA was calculated to test the hypothesized interaction of system confidence, automation expertise, and image quality on compliance. Compliance was defined as the portion of times participants reported that a target was present. Because participants did not receive diagnostic assistance on 'no aid' trials, responses for these trials were excluded from this analysis. The predicted interaction was statistically significant, $F(2, 156) = 3.74$, $p = .026$, partial $\eta^2 = .05$. Simple effects analyses indicated that the interaction between system confidence and image quality differed for the levels of automation expertise. Therefore, interactions were compared separately for the expert and novice systems. Only for the *expert* system was the interaction statistically significant, $F(2, 78) = 3.24$, $p = .044$, partial $\eta^2 = .08$. Participants who interacted with the *expert* system complied more often when image quality was low than when image quality was high across all levels of system confidence, except when the system was 25% confident (Figure 6). When the system was 25% confident, the mean compliance rates for the low and high image quality conditions were ($M = .41$) and ($M = .37$), respectively, $F(1, 39) = 2.77$, $p > .05$.
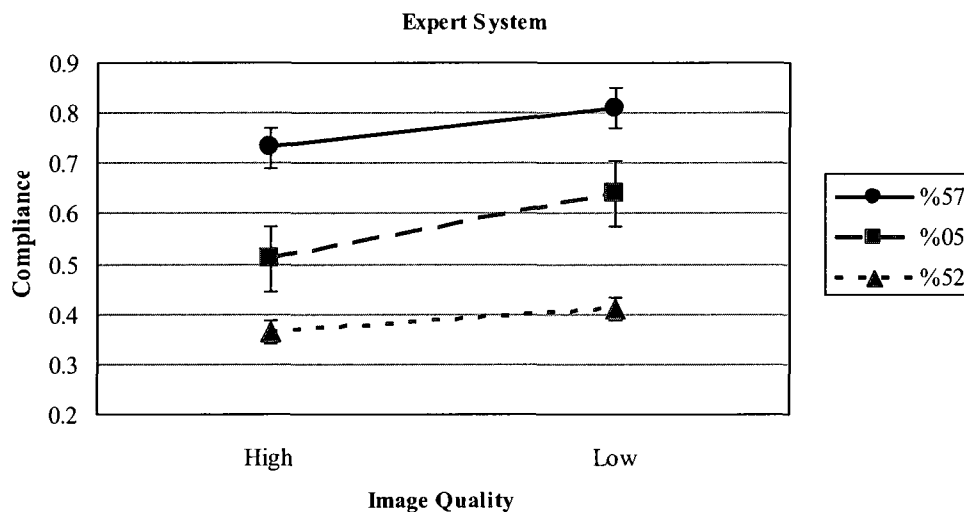
**Expert System**



*Figure 6.* Compliance as a function of system confidence and image quality for expert system condition.

*Hypothesis 5.* A custom 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation

Expertise: expert, novice) x 3 (Session: 1, 2, 3) mixed factorial ANOVA was calculated

to test the hypothesized interaction of automation expertise, system confidence, and

session on compliance. Results indicated that the predicted interaction failed to reach

statistical significance, $F(4, 312) = .36, p > .05$. However, the omnibus ANOVA for

system confidence, $F(2, 156) = 194.37, p = .001$, partial $\eta^2 = .71$, session, $F(2, 156) =$

$21.75, p = .001$, partial $\eta^2 = .22$, and automation expertise, $F(1, 78) = 6.51, p = .013$,

partial $\eta^2 = .08$, were statistically significant. Post hoc analysis indicated a significant

linear trend for system confidence, $F(1, 78) = 242.27, p = .001$; compliance increased as

system confidence increased (see Figure 7). Mean compliance rates for the 25%, 50%

and 75% confident conditions were $M = .36$ ($SD = .22$), $M = .54$ ($SD = .22$) and $M = .75$

($SD = .22$), respectively. Trend analyses also indicated a significant linear trend for

session, $F(1, 78) = 33.40$, p = .001; compliance decreased over the course of the

experiment (see Figure 8). The main effect of automation expertise indicated that

participants complied with the expert system ($M = .58$, $SD = .22$) more than the novice
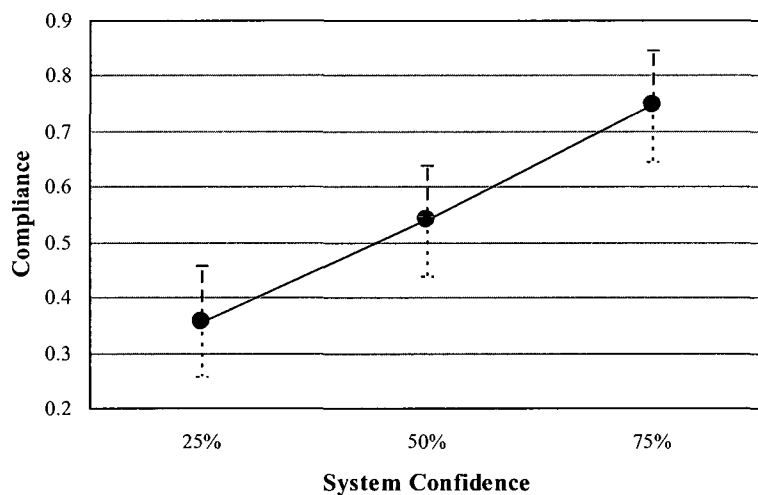
system ($M = .52$, $SD = .20$).



*Figure 7.* Compliance as a function of system confidence.
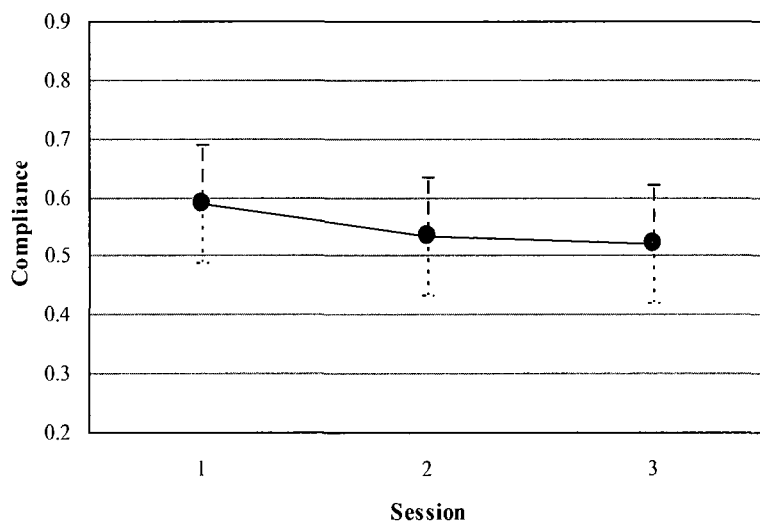


*Figure 8.* Compliance as a function of session.

*Detection sensitivity hypotheses.* A custom contrast analysis with group

(experimental vs. control) as the predictor and detection sensitivity as the criterion was

calculated to determine if sensitive was higher when participants received automated

assistance than when they performed the task manually. Detection sensitivity was

operationally defined as participants' decision-making accuracy and was calculated using

the Signal Detection Theory metric $d'$ prime (Green & Swets, 1966). Results failed to

support the predicted relationship, $F(1, 104) = .16, p > .05$; aided participants ($M = .29$,

$SD = .35$) were not significantly more sensitive than unaided participants ($M = .31, SD = .36$).

Next, a custom 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Image

Quality: high, low) ANOVA was calculated to test the predicted interaction between

system confidence and image quality. Results revealed a significant interaction, $F(3, 243) = 7.84, p = .001$, partial $\eta^2 = .09$ (see Figure 9). When image quality was *high* sensitivity
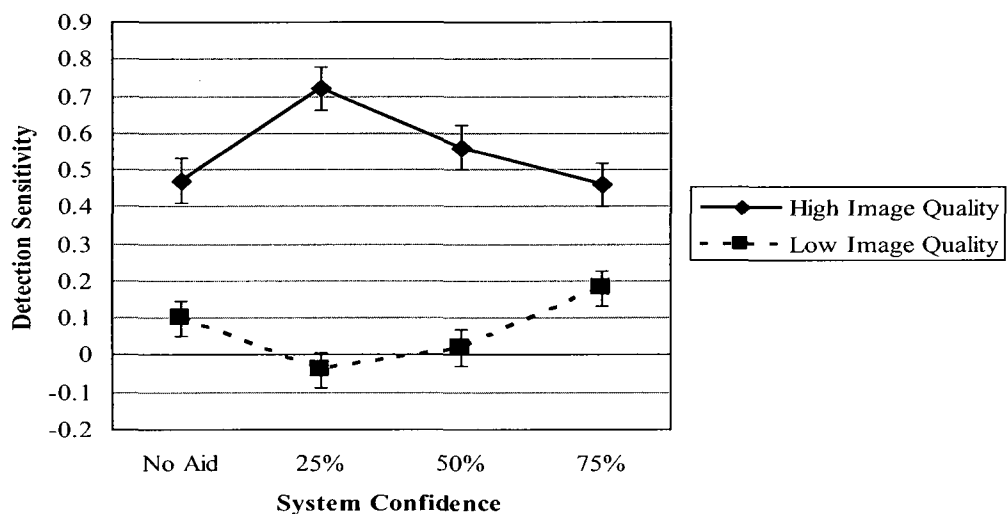


*Figure 9.* Detection sensitivity as a function of system confidence and image quality.

decreased as system confidence increased. Conversely, when image quality was *low*, sensitivity increased as system confidence increased.

*Detection bias hypotheses.* A custom contrast analysis was calculated to determine if automation expertise influenced response bias. Response bias was operationally defined as participants' response criterion, or willingness to respond, and was calculated using Signal Detection Theory metric *C*. As expected, results showed that participants who interacted with the expert system ($M = -.17$, $SD = .43$) adopted a more liberal response strategy than participants who interacted wit the novice system ($M = -.03$, $SD = .37$), $F(1, 78) = 5.61$, $p = .02$, partial $\eta^2 = .07$ (see Figure 10).
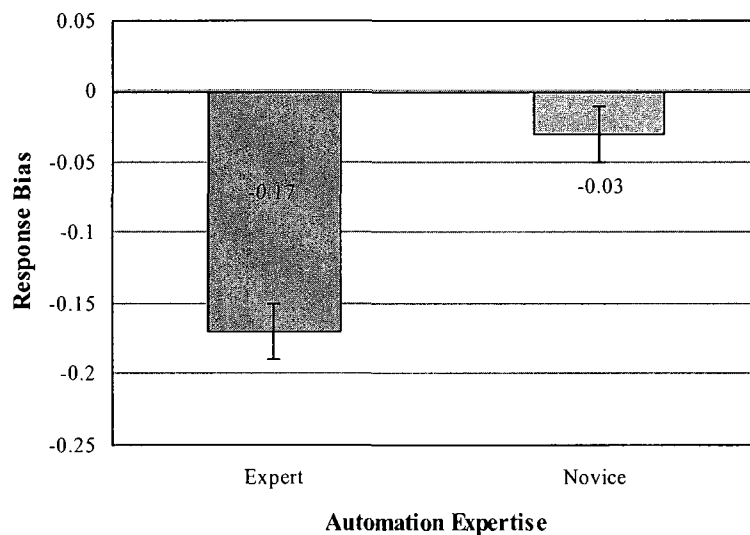


*Figure 10.* Response bias as a function of automation expertise.

Next, a simple contrast analysis was calculated to determine if image quality influenced response bias. Results revealed a significant trend, $F(1, 78) = 27.56$, $p = .001$,

partial $\eta^2$ = .26. Contrary to expectations, participants were more liberal when image

quality was low ($M$ = -.18, $SD$ = .43) than when image quality was high ($M$ = -.01, $SD$ =

.38) throughout the task (see Figure 11).
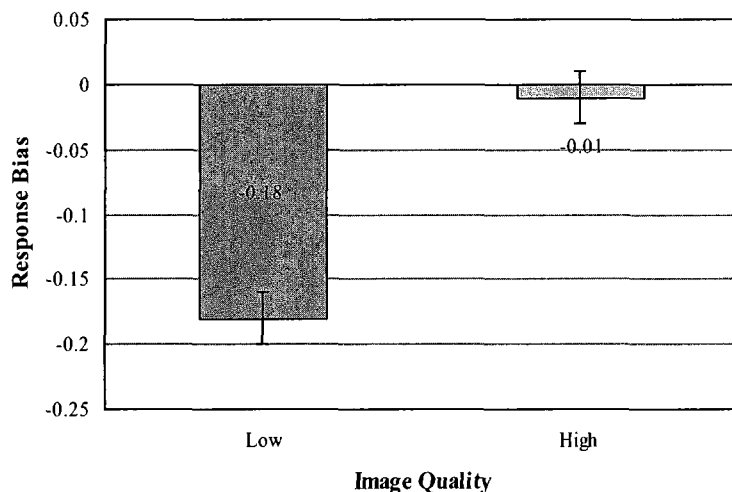


*Figure 11*. Response bias as a function of image quality.

Finally, a linear trend analysis was calculated to determine if participants'

response criterion was positively influenced by system confidence. Results revealed a

significant linear ($F(1, 78)$ = 123.32, $p$ = .001, partial $\eta^2$ = .61) and quadratic trend ($F(1$,

78) = 169.98, $p$ = .001, partial $\eta^2$ = .69). As shown in Figure 12, when no aid was

available participants adopted a neutral response strategy. However, the introduction of

system confidence ratings influenced participants' response criterion such that

participants became more liberal as system confidence increased. Participants' mean

response criterion rates for the no aid, 25%, 50%, and 75% confident trials were $M$ = -

.01 $(SD = .39)$, $M = .26$ $(SD = .40)$, $M = -.09$ $(SD = .40)$, and $M = -.54$ $(SD = .45)$.
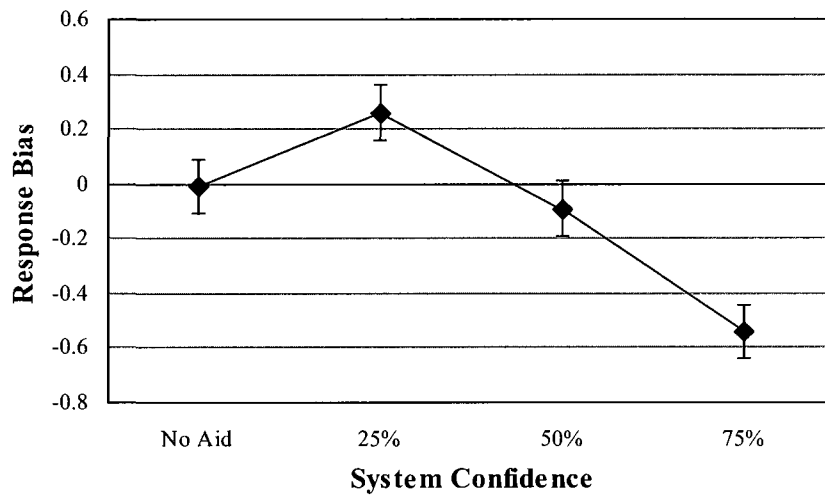


*Figure 12.* Response bias as a function of system confidence.

## Exploratory analyses

Next, a series of exploratory analyses were calculated to examine additional relationships among the variables. The results of these analyses are grouped according to each analysis' criterion (i.e., dependent variable). For brevity, interpretations exclude previously discussed main and interaction effects; only unique relationships are reported.

*Initial trust.* A one-way between subjects ANOVA was computed to examine the effects of system expertise on initial trust. Initial trust was measured using a modified version of the System Trust Scale (Jian et al., 2000). Though no hypotheses were made regarding participants' initial trust in the diagnostic system, empirical data suggest that automation expertise can impact perceptions of system accuracy. Therefore, it is reasonable to assume that participants would perceive the expert system to be more trustworthy than the novice system. Prior to running the ANOVA, Pearson's correlation

analyses were calculated to determine if the scale items were related. Results indicated

that the distrust and trust items were significantly negatively correlated. However,

because previous research suggests that trust and distrust are separate but related

constructs, the distrust items were excluded from the analysis (Spain, Bustamante, et al.,

2008). Thus, only the items that pertained to operator trust were aggregated to arrive at a

single score of initial trust. Results confirmed a significant main effect for automation

expertise, $F(1, 72) = 23.76$, $p = .001$, partial $\eta^2 = .25$. As illustrated in Figure 13,

participants perceived the expert system ($M = 4.82$, $SD = 1.08$) as being more trustworthy

than the novice system ($M = 3.64$, $SD = .98$).



*Figure 13.* Initial trust ratings as a function of automation expertise.
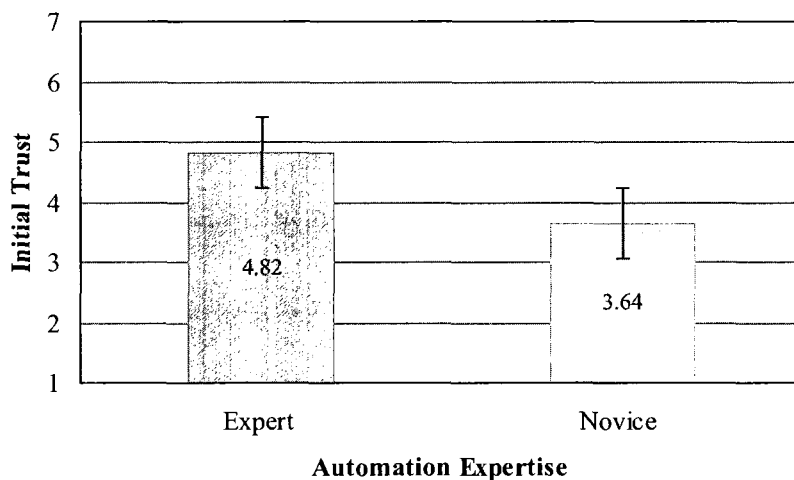
*Overall system trust.* A 2 (Automation Expertise: expert, novice) x 2 (Image

Quality: high, low) mixed factorial ANOVA was calculated to assess the effects for

automation expertise and image quality on overall system trust. Overall system trust was

assessed at the end of each experimental session using the System Trust Scale. Results indicated that the interaction between image quality and expertise, $F(1, 78) = .92, p > .05$, and the main effect for image quality, $F(1, 78) = .20, p > .05$, failed to reach statistical significance. However, a significant main effect for automation expertise indicated that participants trusted the expert system ($M = 3.31$, $SD = 1.16$) more than the novice system ($M = 2.84$, $SD = 1.05$), $F(1, 78) = 5.11, p = .026$, partial $\eta^2 = .06$.

*Perceived reliability.* A 2 (Automation Expertise: expert, novice) x 2 (Image Quality: high, low) mixed factorial ANOVA was calculated to assess main and interaction effects of automation expertise and image quality on perceived system reliability. Perceived reliability was assessed at the end of each experimental session. The interaction between image quality and automation expertise, $F(1, 72) = .14, p > .05$, and the main effect for image quality, $F(1, 72) = .31, p > .05$, failed to reach statistical significance. However, the main effect for automation expertise approached significance, $F(1, 72) = 3.70, p = .06$, partial $\eta^2 = .05$. Data suggest participants perceived the expert system ($M = 56.55$, $SD = 16.00$) as being more reliable than the novice system ($M = 50.44$, $SD = 14.58$).

*Diagnosis trust.* A 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation Expertise: expert, novice) x 2 (Image Quality: high, low) x 3 (Session: 1, 2, 3) mixed factorial ANOVA was calculated to explore the effects of automation expertise, system confidence, image quality, and session. The purpose of this analysis was to determine if any of these factors jointly contributed to the temporal variability of trust scores. Analyses indicated that the assumption of sphericity was violated for several within-subjects variables, therefore all interpretations were made using the Greenhouse-Geisser

*F* value. Results revealed significant interactions of image quality and session, $F(1.85,$ 141.6) = 3.99, $p$ = .020, partial $\eta^2$ = .05, and image quality, system confidence, and session $F(3.69, 288.27)$ = 3.12, $p$ = .015, partial $\eta^2$ = .04. The significant three-way interaction suggested that image quality and system confidence affected diagnostic trust differently over the course of the experiment. Therefore, I conducted a follow-up *post hoc* analysis to examine these differences.

Simple effects interaction analyses indicated that the interaction differed for the two image quality conditions. Only for the *high* image quality condition was the interaction of system confidence and session statically significant, $F(3.64, 287.37)$ = 2.68, $p$ = .032, partial $\eta^2$ = .03. As shown in Figure 14, when image quality was high, trust changed over the course of the experiment and the rate of change varied as a function of system confidence. When the system was 75% confident, participants sustained a high level of trust during the first ($M$ = 3.21, $SD$ = .72) and second session ($M$ = 3.23, $SD$ = .86), but during the third session ($M$ = 3.02, $SD$ = .86) trust dropped significantly, $F(1, 79)$ = 9.99, $p$ = .002, partial $\eta^2$ = .11. A different trend was evident when the system was 50% confident. Trust dropped significantly from the first session ($M$ = 2.76, $SD$ = .66) to the second session ($M$ = 2.57, $SD$ = .71), $F(1, 79)$ = 9.50, $p$ = .002, partial $\eta^2$ = .11, but no change was evident from the second to third session (p > .05). Conversely, when the automated system was 25% confident, trust declined linearly from the first ($M$ = 2.65, $SD$ = .71) to the third session ($M$ = 2.47, $SD$ = .79), $F(1, 79)$ = 6.41, $p$ = .013, partial $\eta^2$ = .08.

**High Image Quality Condition**
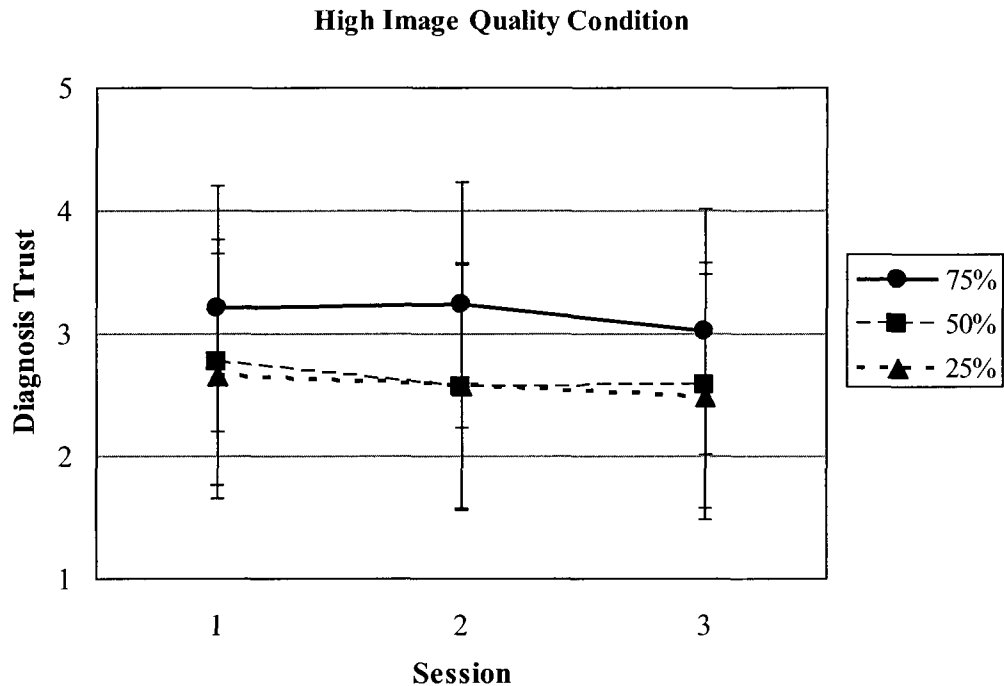


*Figure 14.* Diagnosis trust as a function of system confidence and session for the high image quality condition.

*Compliance.* A 3 (System Confidence: 75%, 50%, 25%) x 2 (Automation

Expertise: expert, novice) x 2 (Image Quality: high, low) x 3 (Session: 1, 2, 3) mixed

factorial ANOVA was calculated to explore the effects of automation expertise, system

confidence, image quality, and session on compliance that were not accounted for in the

original hypotheses. The purpose of this analysis was to determine if any factors

contributed to the temporal variability of compliance. Results revealed interactions

among the following variables: system confidence and session, $F(4, 312) = 3.74$, $p =$

.005, partial $\eta^2 = .05$; image quality, session, and automation expertise, $F(2, 156) = 3.99$,

$p < .020$, partial $\eta^2 = .05$, and image quality, system confidence, and session, $F(4, 312) =$

$3.44$, $p = .009$, partial $\eta^2 = .05$.

A post hoc analysis was conducted to examine the interaction of system confidence, image quality, and session on compliance. I was specifically interested in determining if participants' adopted different compliance strategies based on image difficulty and system confidence. Results showed participants' compliance differed for the levels of image quality. Therefore, simple effects interactions were compared separately for the high and low image quality conditions.

Only for the *high* image quality condition was the interaction between system confidence and session statistically significant, $F(4, 316) = 6.07, p = .001$, partial $\eta^2 = .07$. As shown in Figure 15, and similar to the diagnostic trust results, compliance changed and the rate of change varied a function of system confidence. To discern this change, I computed repeated contrasts to look at deviations in compliance across consecutive sessions (i.e., session 1 to session 2, and session 2 to session 3). A bonferroni correction was used to control for type I error. When the system was 75% confident, participants sustained a high level of compliance during the first ($M = .74, SD = .21$) and second session ($M = .75, SD = .20$), but at the third session compliance dropped significantly ($M = .63, SD = .25$), $F(1, 79) = 21.62, p = .001$, partial $\eta^2 = .22$. A different trend was evident when the system was 50% confident. In this condition, compliance mimicked a negative linear trend, $F(1, 79) = 10.71, p = .002$, partial $\eta^2 = .11$. Mean compliance rates for the first, second, and third session were $M = .55$ ($SD = .24$), $M = .50$ ($SD = .20$), and $M = .44$ ($SD = .44$), respectively. When the system was 25% confident, compliance dropped significantly from the first session ($M = .36, SD = .20$), to the second session ($M = .28, SD = .21$), $F(1, 79) = 10.53, p = .003$, partial $\eta^2 = .12$.

**High Image Quality Conidtion**
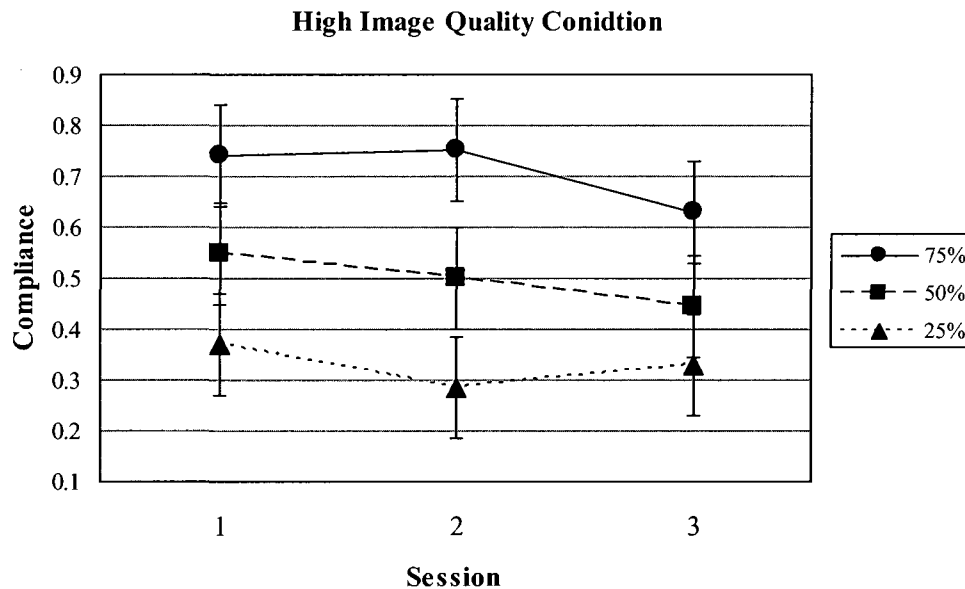


*Figure 15*. Compliance as a function of system confidence and session for the high image quality condition.

*Observed compliance versus optimal compliance.* Next, a series of exploratory

analyses were conducted to compare participants' observed compliance to optimal

compliance. Because each level of system confidence was associated with a unique

probability of a target being present, optimal compliance was defined as the system's

accuracy given its level of confidence. Thus, optimal compliance when the system was

75% confident was .75, optimal compliance when the system was 50% confident was .50,

and optimal compliance when the system was 25% confident was .25. Group means were

compared using one sample *t*-tests.

As shown in Figure 16, on trials in which image quality was *low* and the system

was 75% confident, participants over-complied with the *expert* system ($M$ = .82, $SD$ =

.16), $t(39)$ = 2.69, $p$ = .010. Similarly, when image quality was *low* and the system was

50% confident, participants also over-complied with the *expert* system ($M = .64$, $SD = .18$), $t(39) = 5.10$, $p = .001$. On trials in which image quality was *high* and the system was 75% confident, participants under-complied with the *novice* system ($M = .68$, $SD = .16$), $t(39) = -2.62$, $p = .010$. Finally, participants always over-complied with the system when it was 25% confident, regardless of image quality or automation expertise. However, when image quality was *high* and the system was 25% confident, participants were more likely to comply with the expert system ($M = .41$, $SD = .22$), $t(39) = 4.69$, $p = .001$) than the novice system ($M = .30$, $SD = .11$), $t(39) = 2.76$, $p = .010$.
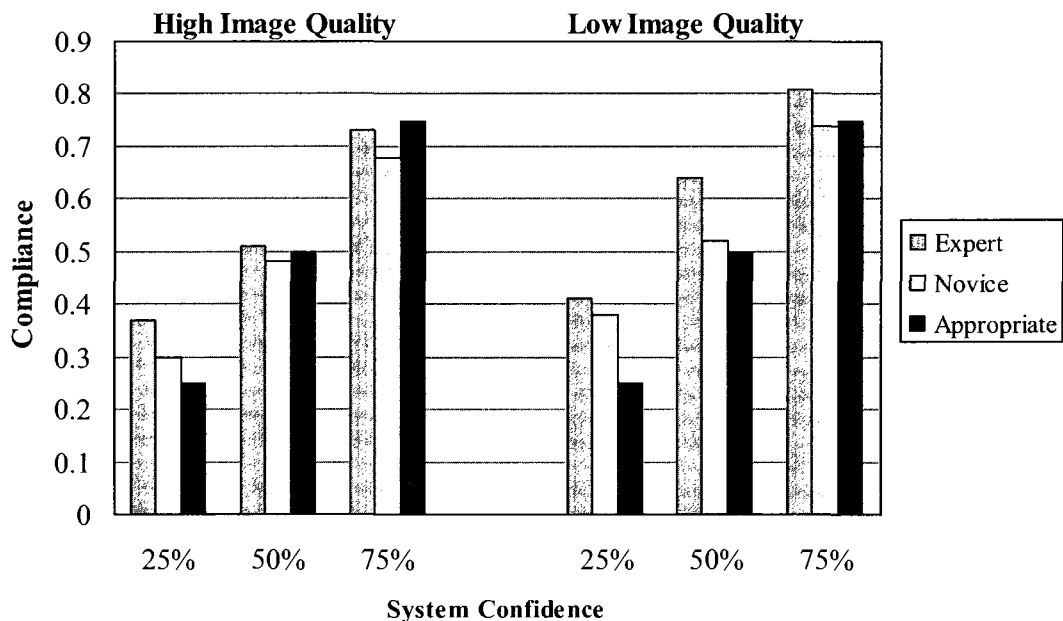


*Figure 16.* Observed compliance versus optimal compliance.

*Response bias.* A 2 (Group: experimental, control) x 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Image Quality: high, low) mixed ANOVA was calculated to determine if there were differences in the response bias between participants who

received automated assistance and participants in the control condition. Results revealed

a significant interaction between group and confidence, $F(3, 312) = 29.01, p = .001$,

partial $\eta^2 = .87$. As shown in Figure 17, on trials in which the system was 75% confident,

aided participants ($M = -.66, SD = .54$) were more liberal than unaided participants ($M = -.02, SD = .31$), $F(1, 104) = 31.78, p = .001$, partial $\eta^2 = .23$. Conversely, on trials in which

the system was 25% confident, aided participants ($M = .30, SD = .36$) were more

conservative than unaided participants ($M = .01, SD = .41$), $F(1, 104) = 10.63, p = .002$,

partial $\eta^2 = .09$.



*Figure 17.* Response bias as a function of group and system confidence.

A 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Automation Expertise:

expert, novice) x 2 (Image Quality: high, low) x 3 (Session: 1, 2, 3) mixed factorial

ANOVA was calculated to explore the effects of automation expertise, system

confidence, image quality, and session on response bias for aided participants.

Participants in the control condition were excluded from this analysis. Results revealed a

significant interaction between image quality, system confidence, and automation

expertise, $F(3, 234) = 2.89$, $p = .036$, partial $\eta^2 = .03$. Simple effects analysis indicated

that participants' response bias differed for the levels of automation expertise. Only for

the *expert* system was the interaction between image quality and system confidence

statistically significant, $F(3, 117) = 5.32$, $p = .01$, partial $\eta^2 = .12$. As shown in Figure 18,

on trials in which the *expert system* was 75% confident, participants were more liberal

when image quality was low ($M = -.75$, $SD = .38$) than when image quality was high ($M$

$= -.49$, $SD = .52$), $F(1, 39) = 20.83$, $p = .001$, partial $\eta^2 = .35$.
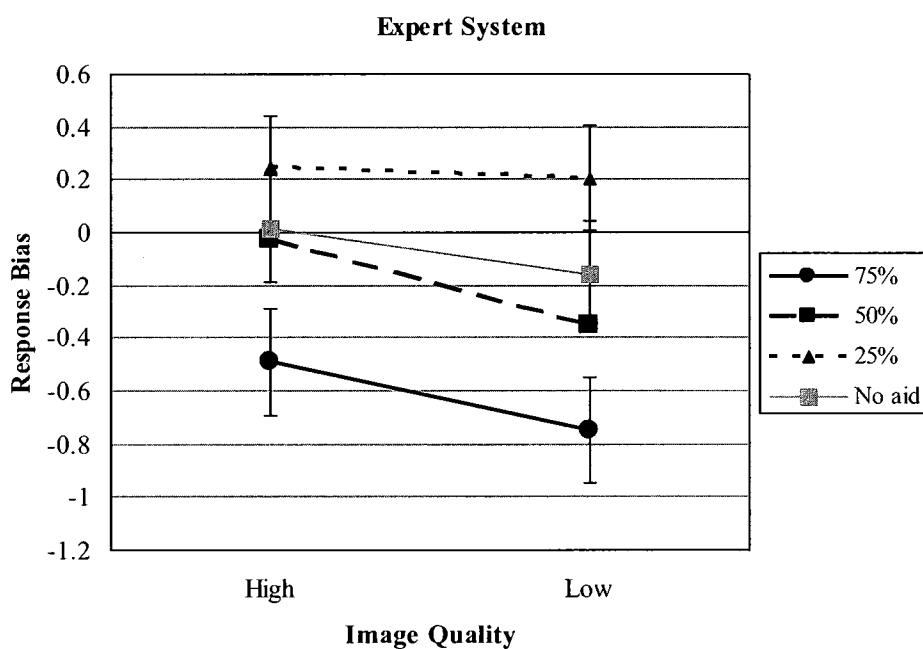
**Expert System**



*Figure 18.* Response bias as a function of system confidence and image quality for the expert system condition.

A similar trend occurred on trials in which the *expert system* was 50% confident; participants were more liberal when image quality was low ($M$ = -.35, $SD$ = .44) than when image quality was high ($M$ = -.02, $SD$ = .36), $F(1, 39)$ = 19.12, $p$ < .001, partial $\eta^2$ = .33. Conversely, on trials in which the expert system was 25% confident, there was no difference in participants' response bias between the low ($M$ = .21, $SD$ = .51) and high ($M$ = .24, $SD$ = .38) image quality condition, $F(1, 39)$ = .23, $p$ > .05.

*False alarm rates.* Based on the performance results, an exploratory 2 (Group: experimental, control) x 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Image Quality: high, low) mixed ANOVA was calculated to determine if there were differences in false alarm rates between participants who received automated assistance and participants in the control group. False alarm rate was defined as the proportion of false targets reported. Because participants in the control condition did not receive automated assistance, this analysis tested if participants were more apt to generate false alarms when they received automated assistance.

A comparison of false alarm rates on 75%, 50%, 25%, and no aid trials between the control condition and experimental condition revealed a statistically significance interaction, $F(3, 312)$ = 25.83, $p$ = .001 partial $\eta^2$ = .20. As shown in Figure 19, aided participants' false alarm rates increased as system confidence increased. Conversely, participants in the control group maintained a consistent rate of false alarms. Results also revealed a significant main effect for image quality, $F(1, 104)$ = 50.76, $p$ = .001, partial $\eta^2$ = .33. Participants were more likely to commit a false alarm when image quality was low ($M$ = .54, $SD$ = .20) than when image quality was high ($M$ = .40, $SD$ = .21). The omnibus

ANOVA for system confidence was also significant, $F(3, 302) = 27.03$, $p = .001$, partial $\eta^2 = .21$. Post hoc trend analysis indicated significant linear and quadratic trends.
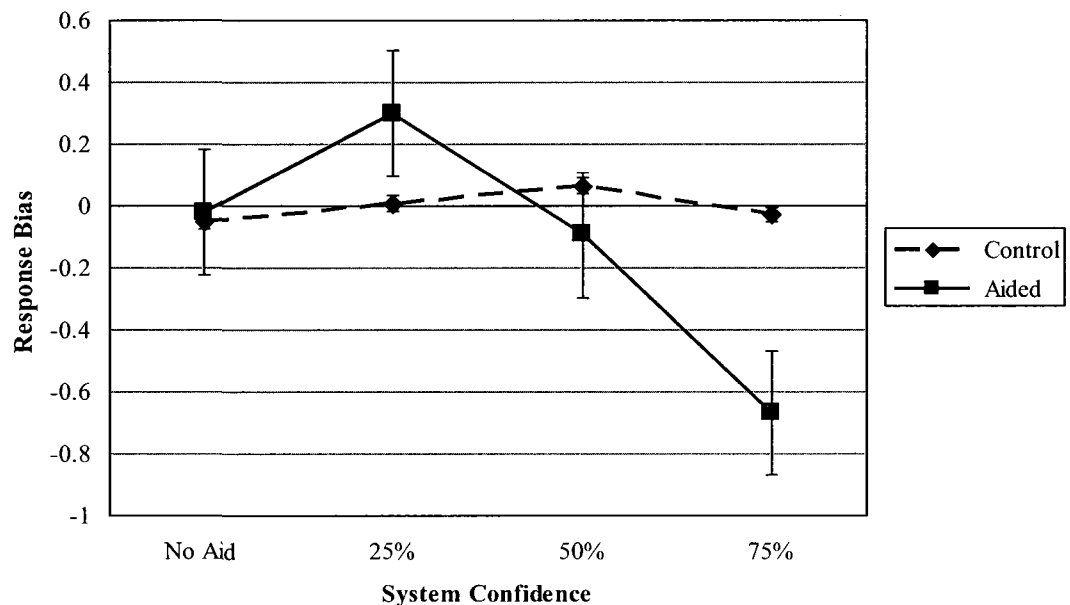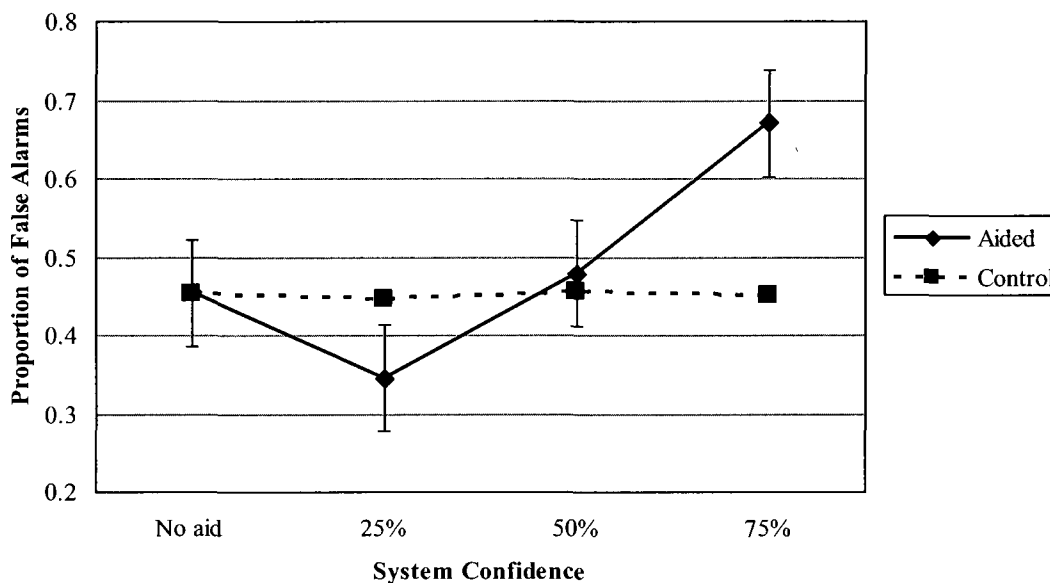


*Figure 19.* False alarm rates as a function of group and system confidence.

Next, a 4 (System Confidence: 75%, 50%, 25%, no aid) x 2 (Automation Expertise: expert, novice) x 2 (Image Quality: high, low) x 3 (Session: 1, 2, 3) mixed factorial ANOVA was calculated to assess the effects of automation expertise, system confidence, image quality, and session on participants' false alarm rates. Participants in the control condition were excluded from this analysis. Though no specific hypotheses were made regarding false alarm rates, it is reasonable to assume that automation expertise, system confidence, and image quality would significantly impact false alarm rates.

As expected, results revealed significant main effects for image quality, system confidence, automation expertise, and session. Participants were more likely to commit a false alarm when image quality was low ($M$ = .55, $SD$ = .23) than when it was high ($M$ = .42, $SD$ = .24), $F(1, 78)$ = 73.10, $p$ = .001 partial $\eta^2$ = .48. The main effect of automation expertise indicated that participants who interacted with the expert aid ($M$ = .51, $SD$ = .23) committed more false alarms than participants who interacted with the novice aid ($M$ = .45, $SD$ = .22), $F(1, 78)$ = 6.11, $p$ = .016, partial $\eta^2$ = .07. The omnibus ANOVA for system confidence was significant, $F(3, 234)$ = 93.61, $p$ = .001 partial $\eta^2$ = .55. Post hoc analyses revealed a significant linear trend; false alarm rates increased as system confidence increased. Mean false alarm rates for the 25%, 50%, 75% trials were $M$ = .36 ($SD$ = .21), $M$ = .48 ($SD$ = .24), and $M$ = .63 ($SD$ = .23), respectively. With respect to the effect of session, post hoc linear trend analysis indicated that false alarm rates decreased over the course of the experiment, $F(2, 156)$ = 15.90, $p$ = .001 partial $\eta^2$ = .16. Mean false alarm rates for the first, second, and third session were, $M$ = .52, ($SD$ = .22), $M$ = .48 ($SD$ = .23), and $M$ = .45 ($SD$ = .23), respectively. All other effects failed to reach significance ($p$ > .05).

*Overall detection performance.* A 2 (Automation Expertise: expert, novice) x 2 (Image Quality: high, low) mixed factorial ANOVA was calculated to assess main and interaction effects for automation expertise and image quality on overall detection performance scores. Results revealed a significant main effect for image quality, $F(1, 78)$ = 65.27, $p$ = .001, partial $\eta^2$ = .46; performance was significantly impaired in the low image quality condition ($M$ = 108.93, $SD$ = 10.98) compared to the high image quality

condition ($M = 126.92$, $SD = 16.99$). All other effects failed to reach significance ($p > .05$).

# Results Summary

- Trust
  - System confidence affected automation trust.
  - Participants trusted the expert system more than the novice system.
  - Participants had greater trust in the diagnostic system when image quality was high than when image quality was low.
  - System confidence influenced the temporal variability of diagnostic trust, particularly when image quality was high.

- Compliance
  - Participants complied with the expert system more than the novice system
  - Participants weighed confidence information from expert systems. differently when image quality was low, which manifested itself in different compliance strategies.
  - System confidence influenced the temporal variability of compliance, particularly when image quality was high.
  - Participants demonstrated overmatching behavior, particularly when image quality was low.

- Sensitivity
  - Aided participants were not more sensitive than unaided participants.
  - When image quality was low, participants used system confidence to increase performance accuracy. Conversely, when image quality was high, performance suffered as system confidence increased.

- Bias
  - Automation expertise, system confidence, and image quality influenced response bias.
  - Participants who interacted with the expert system were more likely to indicate that a target was present when image quality was low than when image quality was high.

- False Alarm Rates
  - False alarm rates increased linearly with system confidence.
  - Participants who interacted with the expert aid were more apt to generate false alarms than participants who interacted with the novice aid.
  - Participants generated more false alarms while viewing low quality images.

# DISCUSSION

The present study had four main objectives. The first objective was to assess the effects of system confidence on operator trust and compliance. The second objective was to determine if automation expertise moderated the effects of system confidence on trust and compliance. The third objective was to determine if image quality influenced the relationship of automation expertise and system confidence on trust and compliance. The final objective was to assess the performance effects of system confidence, automation expertise, and image quality. Research concerning trust in automation is not new; a unique contribution of this research is that the present experiment examined the joint influences of two different sources of trust supporting information, system confidence and automation expertise, on trust and compliance. Furthermore, unlike previous research, this study measured trust on a trial-by-trial basis, thus providing greater insight into the temporal variability of automation trust and compliance. The results from this study have theoretical and practical implications. The results are revisited below, followed by their theoretical and practical implications.

## Automation Trust and Compliance

*System confidence and trust.* Data from the present study indicated that participants exhibited the least amount of trust in the 25% confident system and the greatest amount of trust in the 75% confident system. In reviewing comments from the opinion questionnaire, many participants stated that they only relied on the system when it was 75% confident because they thought the system was wrong when it was 50% and 25% confident. These comments are intriguing considering that the diagnostic aid's

confidence levels were associated with a unique likelihood of a target being present. Specifically, when the aid was 75% confident there was a 75% likelihood that a target was present, when the aid was 50% confident there was a 50% likelihood that a target was present, and when the aid was 25% confident there was a 25% likelihood that a target was present. Thus, the aid was generally accurate in its diagnoses. Still, trust varied as a function of system confidence. There are several possible explanations for this effect.

First, participants may have interpreted system confidence ratings using an analogical trust tuning strategy rather than an analytic tuning strategy. Analytic methods to trust development are cognitively demanding (Lee & See, 2004). In the present study, adopting this strategy would have required participants to discern the system's level accuracy for each level of system confidence. Analogical methods, on the other hand, are a less demanding and imply that trust can be based on an entity's dispositional characteristics. Applying this tuning strategy to the interpretation of system confidence suggests participants may have reasoned that high system confidence reflected high system diagnostic ability and low system confidence reflected poor diagnostic ability. Participant comments support this interpretation as many reported that they associated the 25% confidence rating with poor diagnostic ability and the 75% confidence rating with reliable diagnostic ability.

Second, the framing of the system's diagnosis and confidence estimate may have hindered participants' ability to determine the system's accuracy for each level of system confidence. In the current experiment, the system stated that a target *could* be present and provided a confidence estimate regarding the likelihood that a target was present. This may not have been the optimal framing. Research indicates that humans interpret

probability information poorly and that the framing of such information can influence decision-making (see Dzindolet et al., 2002; Wickens & Hollands, 2000). Furthermore, research on decision-making and over-confidence for dichotomous choice tasks suggests it is best only to present confidence information between 50% and 100% for a given outcome because decision makers inappropriately interpret confidence information when it is below 50% (Yates, Lee, Shiotsuka, Patalano, & Sieck, 1998). Applying these results to the current study suggests that participants may have distrusted the system when it was 25% confident because they were not able to associate low confidence with high levels of accuracy. Perhaps changing the framing of the system's diagnosis and confidence estimate would have mitigated this effect. Rather than indicating that the system was 25% confident that a target *could be present*, the interface could have indicated that the system was 75% confident that a target was *not present*. Communicating system confidence in this manner may have better facilitated appropriate trust. Future research concerning best practices for displaying system confidence information to operators is needed.

*System confidence and compliance.* The present study also observed the effects of system confidence on compliance. Results indicated that participants' compliance rates matched the system's level of confidence. These results are interesting considering participants distrusted the system's diagnostic capability when system confidence was below 50%. Based on the observed trust data, participants should have complied with the system when it was 75% confident and relied on their own intuition when system confidence was 50% and below. However, the compliance data failed to confirm this pattern.

Figure 20 shows the relation between trust and compliance found in previous research and in the current study. In McGuirl and Sarter's (2006) study, participant's compliance matched the system's level of confidence because, assumingly, participants trusted the system. That is, they trusted that when confidence was high, medium, and low the probability of a problem occurring was also high, medium, and low, respectively. In the present study, participants did not trust each level of system confidence equivalently. Yet, their compliance matched the system's level of confidence. This differs from previous research and highlights disconnect between trust and compliance. There are several plausible explanations for these results.
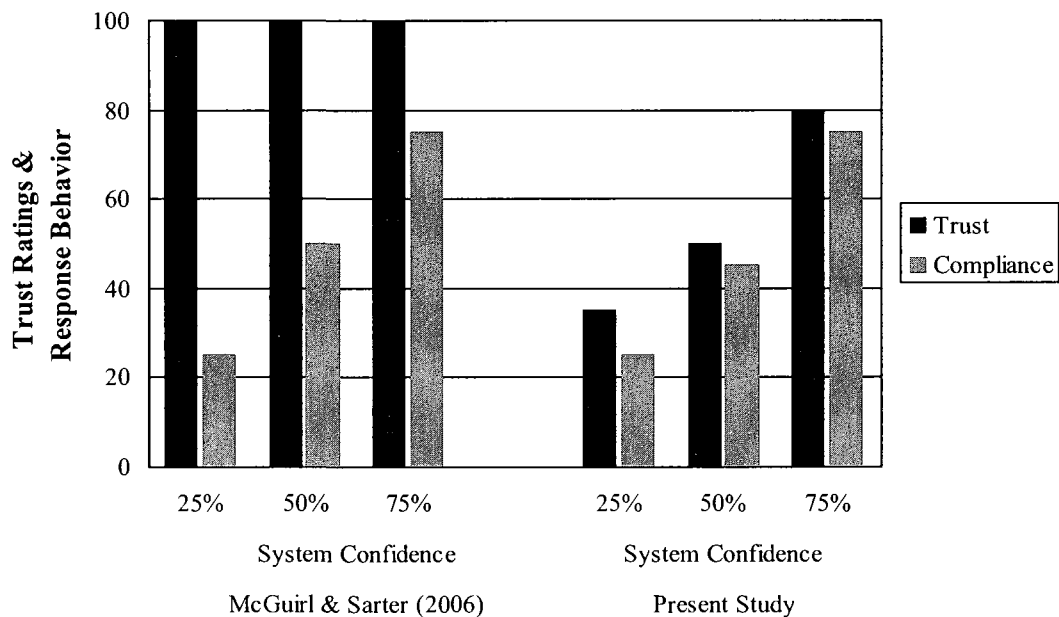


*Figure 20.* Comparative figure: Trust in system confidence and corresponding compliance rates.

First, participants may have calibrated their compliance and diagnostic trust to the base rate of a target occurrence for each level of system confidence. In the present study, each level of confidence was associated with a unique likelihood that a target was present. Given this set up, it is reasonable that participants used the base rate as a heuristic to calibrate trust and compliance. This would explain why participants' diagnostic trust matched compliance patterns.

Alternatively, participants' concurrence strategy may have hindered their ability to accurately evaluate system performance; which led to inappropriate levels of system trust when system confidence was low. As shown by the compliance data, many participants engaged in a probability matching behavior; they matched their response frequency to the accuracy of the aid. This strategy may have caused participants to confuse the reliability of the system with the reliability of manual performance. Indeed, the trust data support this interpretation, as participants were not able to estimate the true reliability of the system when it was 50% and 25% confident. A similar phenomenon has been reported in automation trust literature before. Specifically, Wiegmann (2002) found that participants who adapted a probability matching strategy were not able to accurately estimate the reliability of a diagnostic aid. He attributed these results to the cognitive load associated with simultaneously keeping track of manual and automated performance.

Finally, the observed relation between the trust and compliance data could be attributed to the present study's payoff matrix. The extent to which real world users of diagnostic aids will engage in probability matching is likely a function of the costs associated with correct and incorrect decision. The present experiment used a payoff strategy that encouraged participants to consider the costs and benefits associated with

correct and incorrect decisions. Specifically, participants earned one point for a correct decision and lost one point for an incorrect decision. Participants reported that the payoff system was motivating. Nonetheless, earning and losing points does not approximate the circumstances that might confront a military soldier who must decide whether an object in a SAR image is friendly of hostile, and thus whether to engage on the target. Because the payoff for inappropriately responding was not too costly, participants may have approximated their response rates to the systems confidence to maximize the probability of detecting a target despite their distrust in the system's diagnosis.

*Automation expertise and trust.* A secondary goal of this research was to determine the effects of automation expertise on trust. As expected, participants perceived the expert system as being more trustworthy and reliable than the novice system. These results support Lee and See's (2004) theoretical model and provide further evidence that hearsay information concerning automation performance can influence automation trust. An important contribution of this research relates to the convergence of evidence concerning the effects of automation expertise on automation trust. Three different measurements (i.e., initial trust, diagnostic trust, and system trust) collected at three different time points (i.e., prior to, during, and after interacting with the system) converged to indicate that participants trusted expert systems more than novice systems. These results provide strong empirical evidence that automation expertise influenced automation trust. As discussed by Madhavan and Wiegmann (2007), the source of diagnostic information plays a significant role in the development and maintenance of automation trust.

*Factors affecting the appraisal of diagnostic information.* System confidence, automation expertise, and image quality were expected to influence automation trust and compliance. As such, the present study tested several interaction effects. Results partially supported the hypotheses concerning the interaction of system confidence, automation expertise, and image quality on trust and compliance. Individuals who interacted with the expert system weighed confidence information differently when image quality was low than when image quality was high. No differences were observed for the novice system. These results suggest that the perceived capability of an expert system affects user compliance strategies, particularly when uncertainty is high (i.e. the task is difficult). The data for automation trust trended in a similar pattern but failed to reach statistical significance.

*Temporal variability of trust and compliance.* The effects of automation expertise, system confidence, and image quality were also tested over time. The predicted interaction between system confidence, automation expertise, and session on trust failed to reach significance. Rather, results showed that image quality, not automation expertise, interacted with system confidence to influence the temporal variability of trust and compliance. Specifically, trust and compliance declined over time, particularly when image quality was high, and the rate of decline varied as a function of system confidence. Participants sustained trust the longest when the system was 75% confident; when system confidence fell below 75% trust declined rapidly over time. These results suggest that participants developed a greater trust for the system when it was 75% confident.

The fact that participants' trust and compliance varied only in the high image quality condition could be attributed to the saliency of automation errors. According to

the easy errors hypothesis, automation errors on task easily performed by humans undermines trust and compliance with automated aids (see Dzindolet et al., 2003, Madhavan et al., 2006). In the present experiment, participants may have been more likely to notice "easy errors" when image quality was high than when image quality was low. Indeed, many participants reported that it was easier to detect targets when image quality was high. Thus, it seems likely that participants were also more likely to observe system errors in this condition.

It's also important to note that participants trusted and complied with the expert system more than the novice system over the entire experiment. These results contradict previous research that has shown compliance with imperfect expert systems to decrease more rapidly than compliance with imperfect novice systems (Mayer, 2008). This discrepancy could be attributed to the type of automated system used in the current and previous research. In Madhavan and Wiegmann's (2007) and Mayer's (2008) research, participants interacted with a traditional binary automated diagnostic system. Binary diagnostic systems provide two forms of diagnostic information about problems: "present" and "absent". This type of design philosophy, though needed in some instances, can be limiting because it does not allow insight into the system's decision-making process. As previously discussed, the best way to provide insight into a system's algorithm is to use a design philosophy similar to Sorkin et al.'s (1988) Likelihood Alarm display (LADs). LADs use multi-level diagnostic signals to express the degree of certainty associated with a signal event. The diagnostic aid used in the present study was modeled similar to an LAD. The system indicated the likelihood that a target was present. Participants may have judged the performance of the current system less severely than

they would judge a traditional binary diagnostic system gave them more diagnostic information. This slight difference in the design of the diagnostic aid may have mitigated the loss of trust compliance associated with the breakdown of the perfect automation schema observed in previous research. Further research is needed to substantiate these possibilities.

*Observed compliance vs. optimal compliance.* When comparing observed compliance against optimal compliance, participants over-complied with the expert system across all levels of system confidence, particularly when image quality was low. This "over-matching" behavior suggests participants used the expertise of the aid as a cue to reduce uncertainty. Dijkstra (1999) found similar results and explained them with the Elaborate Likelihood Model (ELM; Petty & Cacioppo, 1981). The ELM states that individuals use two routes when evaluating advice: the central route and the peripheral route. Individuals using the central route are highly confident in their ability to analyze the content of advice, whereas individuals using the peripheral route are not, and therefore base their compliance decisions on surface level cues such as the advisor's presumed expertise. In the present study, participants used the peripheral route when image quality was low to combat uncertainty, and therefore complied with the expert system more often than they complied with the novice system.

It is also important to note that on trials in which the system was 25% confident, participants always over complied with the system. That is, they reported that a target was present too often. This response strategy resembles an estimation bias associated with low likelihood events. Similar to the trust results, these findings have implications

for designing human-machine interfaces that support appropriate trust. These implications will be discussed later.

**Performance Data**

*Detection Sensitivity.* Data from the current experiment partially supported expectations concerning the effects of image quality and system confidence on detection sensitivity. Participants made more correct decisions when image quality was high than when image quality was low. However, compared to the control group, aided participants were not able to use system confidence to improve detection performance. Jamieson and Wang (2007) have found similar results. One reason the data from the present experiment did not show a performance effect for aided participants could be related to their false alarm rates. Aided participants' committed more false alarms as system confidence increased than unaided participants committed. An increase in false alarm rates can reflect poorly on detection sensitivity (Wickens & Hollands, 2000).

The significant interaction between image quality and system confidence suggests that when faced with high levels of uncertainty, providing confidence information can be beneficial to performance. However, there is a cost of presenting confidence information when the detection task is easy. Maltz and Shinar (2003) found similar costs of using automation when a detection task was easy; specifically they found that automated cuing facilitated performance for difficult tasks and impaired performance for easy tasks.

*Response bias.* Similar to the compliance results, automation expertise and system confidence significantly influenced response bias, particularly when image quality was low. These data indicate that participants attempted to maximize the number of targets found when the expert system was moderately and highly confident, particularly when

the task was difficult. These results are in accordance with previous research (see Madhavan & Wiegmann, 2007) and suggest that shifts in criterion setting are strongly influenced by automation expertise, particularly when uncertainty is high.

*Additional costs.* Data from the current experiment also indicated several incurred cost of system confidence and automation expertise. Participants were more apt to generate false alarms as system confidence increased. Furthermore, automation expertise influenced false alarm rates. Participants were more likely to generate false alarms when the expert aid was highly and moderately confident in its diagnosis. This behavior resembles a form of automation bias in which operators rely on automation rather than processing task related information manually (Mosier & Skitka, 1996). The increase in false alarm rates was also evident in participants' response bias. Individuals who adopted a liberal response strategy were likely to commit more false alarms than participants who adopted a conservative strategy.

**Theoretical Contributions**

Lee and See's (2004) Appropriate Trust Framework suggests that operators use analytic and analogical methods to calibrate automation trust. Calibrating trust via an analytic method can be cognitively demanding because it requires human reasoning and the ability to deduce when a system is performing reliably. Calibrating trust via an analogical method is less demanding because it involves using cues to infer how automation will perform. Results from the present study suggest that rather than using logical reasoning to deduce how accurate the diagnostic system was for each level of confidence (i.e. an analytic approach), participants adopted an analogical trust tuning strategy. That is, participants used system confidence as a cue to infer how the system

was performing, rather than its original intent: to provide likelihood information regarding the presence of an enemy target. As a result, trust varied as a function of system confidence.

Data from the current experiment could be used to update Lee and See's (2004) Appropriate Trust Framework. Currently, the framework does not address the manner in which operators use analytic, analogical, methods to guide trust calibration. The model suggests only that these are three methods operators use to tune trust. The present experiment's results suggest that operators may be more likely to adopt a less cognitively demanding trust-tuning strategy. That is, operators may be more likely to use analogical information to serve as a bridge that facilitates trust until operators acquire enough analytic information to guide trust. This calibration strategy has considerable design implications, especially considering the current trend to display automation confidence and reliability information to system users (Jamieson & Wang, 2007).

The results from the present study could also be used to update utility models of automation trust. Dzindolet et al.'s (1999) model describes the manner in which users appraise automation and manual capability, but it does not address how information pertaining to automation capability influences these utility assessments. The present study's results suggest that preconceived biases can influence the interpretation of system confidence information from automation of varying expertise, particularly when uncertainty is high. This appraisal affects trust and compliance.

**Practical Implications**

With regard to interface design, the results of this research have implications for presenting system confidence feedback to operators. Participants calibrated their trust

levels and compliance rates to the system's level of confidence. However, this led to mistrust for the 25% confident system. This could be due to low perceptions of automation capability. Applying the results in the context of a target detection task suggests that a diagnostic system should only provide system confidence feedback for the state of the world that is above chance. That is, rather than indicating that the system is 25% confident that a target *is present*, the interface should indicate that the system is 75% confident that a target is *not present*. However, there is a notable limitation with this type of interface: humans are notoriously bad at interpreting negatively phrased information (see Dzindolet et al. 2002). Future research should focus on best practices for designing human-machine interfaces that display confidence or reliability information.

Results from the current experiment failed to reflect immense performance benefit associated with system confidence ratings. The availability of system confidence ratings improved performance when the task was difficult. However, participants were more apt to generate false alarms as system confidence increased. Furthermore, system confidence ratings impaired detection performance when the task was easy. It is possible that performance suffered as system confidence increased in the easy task condition because automation errors were more obvious and participants stopped relying on the system because it was imperfect. Further testing and refinement are needed to validate the utility of incorporating confidence estimates into diagnostic automation.

Results from the current experiment also indicate that it is important to consider the influence of automation expertise on trust and performance. The analogical process of trust may play an important part in real-world interactions with automation. The U.S. Armed Forces currently have many different forms of aided target recognition technology

(Boyd et al., 2006). Clearly, hearsay information concerning the aid's history and functionality can affect trust and dependence (Bliss, Dunn, et al., 1995). This type of information will likely affect an operator's decision bias to comply with aid advice. In the current experiment, participants were more liberal, and consequently generated more false alarms, when they interacted with the expert system. In an operational environment, operators who interact with an "expert" aid may also adopt a liberal response criterion until they learn the accuracy of the system. In this case, the "expertise" of the system may serve as a bridge that facilitates trust until operators acquire enough information to achieve at least moderate levels of analytic trust. This may cause operators to over-rely on automation, and consequently generate more false alarms. Research needs to be conducted to explore the generalizability of the present findings to higher-risk scenarios and specifically determine if users of expert systems adopt liberal or conservative response strategies.

There are several strategies and practices that designers and practitioners could implement to combat the observed cost associated with system confidence and expert systems. First, practitioners could train operators to recognize situations or signal patterns that correlate with automation capability and reliability. This type of training may enhance users' temporal specificity of trust. Second, practitioners could inform operators about the capabilities and limitations of expert systems. Learning the performance standards of so-called "expert systems" may reduce preconceived cognitive biases and facilitate appropriate trust and compliance. Third, practitioners and training agencies should institute training events that provide real time feedback regarding the performance of automated and manual performance during training events. This way, operators will

learn the situational accuracy of manual and automated performance. This too may reduce biases about automated and manual performance.

A final intervention strategy relates to the design philosophy driving the automation. Researchers acknowledge that it is important that operators have the ability to observe the system's decision and the information it uses to reach that decision (Beck, Dzindolet, & Pierce, 2007; Lee & See, 2004; Sheridan, 2002). Thus, automation should be designed to allow users to "reach back" and validate automated decisions. This mimics a "trust but verify" design strategy. Incorporating such a level of transparency into the design of automated systems may enhance trust calibration.

**Funding Opportunities and Directions for Future Research**

Because recent concern over reducing battlefield fratricide has focused interest on developing advanced battlefield combat identification systems, results from the current experiment could lead to funding opportunities for the research and development of automatic target recognition (ATR) systems. Automatic target recognition systems use advanced algorithms and sensor data to detect, recognize, and identify battlefield targets. ATR systems were originally envisioned to operate autonomously, detecting, locating, and classifying targets with little or no human intervention (MacMillian et al., 1994). However, completely autonomous performance remains well beyond current ATR capability. Under current performance levels, human operators play a critical role: screening ATR interpretations, rejecting false alarms, and searching for additional targets. Consequently, it has become increasingly important to understand how humans interact with their automated counterparts.

Future research should focus on best practices for conveying confidence information to operators and methods for training operators to appropriately use confidence information. Research concerning group decision-making indicates that, for a dichotomous choice task, decision makers should receive confidence information only when it is above 50% for a given outcome (Yates et al., 1998). To date, there is no empirical evidence regarding human machine interface configurations for communicating confidence information to system operators. This seems to be a fruitful avenue of research considering the proliferation of automated systems in complex task environments.

Future studies should also focus on the best presentation format (numerical vs. text), modality (verbal vs. visual), and timing (prior vs. concurrent vs. after presenting the image) for presenting system confidence information. Additionally, it may be important to examine if system confidence information should be supplemented with additional visual or verbal cues. Providing operators with confidence ratings along with a referent image of the supposed target may influence detection sensitivity and response bias.

Future research should also focus on the continued refinement of trust measurement techniques. One of goals of this study was to measure trust on a trial-by-trial basis. To achieve this goal, I used a single item indicator. Participants were instructed to rate their trust in the system's diagnosis using a Likert type scale. One limitation with this approach pertains to the validity of the measurement item. Though the data indicate that system confidence did influence trust ratings, participants may have rated their trust in the systems overall diagnosis capability, rather than their trust for in the system's diagnosis that specific trial. Additionally, participants may not have coupled

the system's confidence rating and diagnosis when making their trust rating. Future research may consider measuring trust for each level of system confidence during and after the experiment to obtain convergent validity. The measurement of trust needs continual refinement to ensure researchers can make appropriate assumptions from their research.

# CONCLUSIONS

The present study sought to determine the effects of system confidence, automation expertise, and image quality on trust, compliance, and performance. As expected, these sources of information significantly influenced automation trust and participants' response strategies. Specifically, results indicated that

- Participants matched their trust and compliance to the system's level of diagnostic confidence.

- Participants were more likely to trust and comply with the expert system than the novice system.

- Participants weighed confidence information from expert and novice system differently, especially when uncertainty was high. This resulted in different compliance strategies.

- System confidence affected the temporal variability of automation trust and compliance.

- Participants generated more false alarms as system confidence increased and when interacting with the expert system.

These results may prove to be beneficial for designing automated aids and updating training modules for interacting with automated aids. Future research should focus on best practices for conveying system confidence information to operators and methods for training users to appropriately interpret system confidence feedback from expert systems.

# REFERENCES

American Psychological Association. (2002). Ethical principles of psychologists and

code of conduct. *American Psychologist, 47*, 1597-1611.

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decisions:

Controlling intent and appraisal errors in a target detection task. *Human Factors,*

*49*, 429-437.

Bliss, J. P. (1993). *The cry-wolf phenomenon and its effect on operator responses.*

Unpublished doctoral dissertation, University of Central Florida, Orlando.

Bliss, J. P. (2000). Investigations of alarm mistrust under conditions of varying alarm and

ongoing task criticality. In *Human Factors in Auditory Warnings*. N. Stanton & J.

Edworthy (Eds.), pp. 173-199 (Aldershot: Ashgate).

Bliss, J. P. (2003). Collective mistrust of alarms. *International Journal of Applied*

*Aviation Studies, 3*(1), 13-38.

Bliss, J., Dunn, M., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An

investigation of two methods to increase alarm response rates. *Perceptual and*

*Motor Skills*, 80, 1231-1242.

Boyd, C. S., Collyer, R. S., Skinner, D. J., Smeaton, A. E., Wilson, S. A., Krause, D. W.,

et al. (2005). Characterization of combat identification technologies. In *IEEE*

*International Region 10 Conference,* Melbourne, Australia. pp. 568-573.

Breznitz, S. (1984). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence

Erlbaum.

Brown, R. D., & Galster, S. M. (2004). Effects of reliable and unreliable automation on

subjective measures of mental workload, situation awareness, trust, and

confidence in a dynamic flight task. In *Proceedings of the 48<sup>th</sup> Annual Meeting of the Human Factors and Ergonomics Society,* New Orleans; LA, pp 147-151.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: What is it and how can it be improved? *Proceedings of the 1998 Command and Control Research and Technology Symposium,* Monterey, CA. pp 1- 28.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology, 18,* 399-411.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58,* 697-781.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79–94.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147–164.

Entin, E. B., Entin, E. E., MacMillan, J., & Serfaty, D. (1995). Situation awareness and human performance in target recognition. *Intelligent Systems for the 21<sup>st</sup> Century. IEEE International Conferences, 4,* 3833-3837.

Fallon, C. K., Bustamante, E. A., Ely, K. M., & Bliss, J. P. (2005). Improving user trust with a likelihood alarm display. In *Proceedings: 11th International Conference on Human-Computer Interaction,* pp 1-10.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Jamieson, G. A., & Wang, L. (2007). *Developing human-machine interfaces to support appropriate trust and reliance on automated combat identification systems* (Tech. Rep. No. 1.). Toronto, Canada: University of Toronto.

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*, 53-71.

Kim, J., & Moon, J. Y. (1998). Designing towards emotional usability in customer interfaces – Trustworthiness of cyber-banking system interfaces. *Interacting with Computers, 10,* 1- 29.

Keppel, G., & Wickens, T. D. (2004). *Design and Analysis: A Researcher's Handbook*, 4th Edition, Prentice Hall, Upper Saddle River, NJ.

Lee, J. D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153-84.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46,* 50-80.

MacMillian, J., Entin, E. B., & Serfaty, D. (1994). Operator reliance on automated support for target detection. In *Proceedings: 38th Annual Meeting of the Human Factors and Ergonomics Society,* 1285-128.

Madhavan, P., & Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*, 241-256.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, reliability, and pedigree on operator interaction with decision support systems. *Human Factors*, *47*(2), 332-341.

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, *45*(2), 281-295.

Mayer, A. (2008). *The manipulation of user expectancies: Effects on reliance, compliance, and trust using an automated system.* Unpublished master's thesis, Georgia Institute of Technology, Atlanta.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48,* 656-665.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history based trust in human-automation interactions. *Human Factors, 50,* 194-210.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46,* 196-204.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds). *Automation and human performance: Theory and applications. Human Factors and Transportation* (pp. 201-220). Mahwah, NJ: Lawrence Erlbaum Associates.

Muir, B.M. (1989). *Operators' trust in and percentage of time spent using the automatic controllers in a supervisory process control task.* Unpublished doctorial dissertation. University of Toronto, Ontario, Canada.

Muir, B. M., & Moray, N. (1996). Trust in automation II: Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

Nan, X. (2007). The effect of perceived source credibility on persuasion: Moderators and mechanisms. Paper presented at the *Annual Meeting of the International Communication Association,* San Francsisco CA, Online Retrieved 2008-7-17 from http://www.allacademic.com/meta/p168951_index.html.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30*(3), 286-297.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50,* 511-520.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches.* Dubuque, IA: William C. Brown.

Rice, S. (*in press*). Examining single and multiple-process theories of trust in automation. *Journal of General Psychology.*

Rhine, R. J., & Severance, L. J. (1970). Ego-involvement, discrepancy, source credibility and attitude change. *Journal of Personality and Social Psychology, 16,* 175-190.

Safar, J. A., & Turner, C. W. (2005). Validation of a two factor structure for system trust. In *Proceedings: 49th Annual Meeting of the Human Factors and Ergonomics Society*, 497 – 501.

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid.* Unpublished doctoral dissertation. Georgia Institute of Technology, Atlanta.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993) Automation-induced complacency – Development of the complacency potential rating scale. *The International Journal of Aviation Psychology, 3*, 111-122.

Sheridan, T. B. (2002). *Humans and Automation: System Design and Research Issues.* Wiley Interscience: Santa Monica, CA.

Sheridan T., & Parasuraman, R. (2006). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*, 89-129.

Sniezek, J. A., & Von Swol, L. M. (2001). Trust, confidence and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes, 84*, 288-307.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30*(4), 445-459.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human Computer Interaction, 1*, 49-75.

Spain, R. D., & Bliss, J. P. (2008). The effect of sonification display pulse rate and reliability on operator trust and perceived workload during a simulated patient monitoring task, *Ergonomics, 51*, 1320-1337.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. In *Proceedings: 52nd Annual Meeting of the Human Factors and Ergonomics Society,* 1335-1339.

Sterling, B. S., & Jacobson, C. N. (2006). *A human factors analysis of aided target recognition technology.* (Technical Memorandum No. ARL-TR-3959). Aberdeen Proving Ground, MD: Army Research Laboratory.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings: 46th Annual Meeting of the Human Factors and Ergonomics Society,* 332-336.

Tabachnick, B. G., & Fidell, L. F. (2001). *Computer-Assisted research design and analysis.* Needham Heights, MA: Allyan & Bacon.

Wickens, C. D., Conejo, R., & Gempler, K. (1999). Unreliable automated attention cueing for air-ground targeting and traffic maneuvering. In *Proceedings: 43rd Annual Meeting of the Human Factors and Ergonomics Society.* 21-25.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance.* 3rd Edition. New Jersey, NJ: Prentice Hall.

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies, *Human Factors, 44,* 44-50.

Yeh, M., & Wickens, C. D. (2001). Examination of explicit and implicit display signaling on attention allocation and trust calibration. *Human Factors, 42,* 455-465.

Yates, J. F., Lee, J., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge

overconfidence. *Organizational Behavior and Human Decision Processes, 74,* 89-177.

# APPENDIX A

## FLYER FOR PROJECT TARGET DETECTION (IRB # 08 087)

**Description:**
This project is a laboratory experiment studying human interaction with computer advisors. The study is interested in how automation credibility and automation confidence ratings affect detection performance, decision-making accuracy, trust, and the perceived reliability of computer advisors.

**Eligibility:**
You must be 18 years or older to participate. You must have normal hearing. If you require corrective lenses, you must wear them during the experiment.

**Incentives:**
Participation in this study will earn you two Psychology Department research credits.

**Location and Time:**
This study will take place in Mills Godwin Building room 328. The study will take approximately 1 hour and 30 minutes. You may sign up for the experiment using SONA.

**Researchers:**
Principal Researcher: Dr. James P. Bliss, Ph.D.
Graduate Researcher: Randall D. Spain, M.S.

**Contact Information:**
rspain@odu.edu

## APPENDIX B

## PARTICIPANT BACKGROUND INFORMATION FORM

Participant #_____ Date: _____ Time: _____

The purpose of this questionnaire is to collect background information for participants in this experiment. This information will be used strictly for this experiment and for research purposes only. Please complete each item to the best of your ability.

1. Age _____

2. Sex (circle one)        Male            Female

3. Status (circle one) Undergrad     Grad     Faculty     Staff     N/A

4. Department / Major _____

5. How often do you use a computer? (circle one)

   5-7 days/week   2-4 days/week   1 day/week   2-3 days/month   1 day/month   less

6. Have you ever been diagnosed as color blink or color deficient?

   0 = No

   1 = Yes

7. Have you ever been diagnosed as having a deficiency in your vision?

   0 = No

   1 = Yes

   a. If yes, do you have correction with you (i.e., glasses, contact lenses, etc.)?

8. Which statement below best describes your attitudes towards computers and other automated devices in general? (check one)

   a. _____ Computers and automated devices are generally reliable until they prove otherwise.

   b. _____ Computers and automated devices are unreliable until they prove otherwise.

# APPENDIX C

## EXPERIMENT INSTRUCTIONS

Welcome to the experiment!

Today, you will pretend to be a military analyst, looking for covert enemy targets in intelligence images. Your job will be to search these images and report whether a target is present. Click on the "View Targets" button to familiarize yourself with the potential targets. Then, click on the "View SAR Images" button to familiarize yourself with the intelligence images.

You will complete 96 trials. On each trial, you will view an image for about 1 second. After that, you will make several responses. First, you will click on a button to indicate whether you wish to report an enemy target. Click on the YES button if you wish to report that an enemy target is present. Click on the NO button if you do not think that an enemy target is present. Next, you will indicate how CONFIDENT you are in your response. You will rate confidence on a 1 to 5 scale, where 1 indicates "Not Confident" and 5 indicates "Very Confident". In addition to reporting your decision confidence, on some trials you will be asked to indicate how much you trust a computer aid's diagnosis. You will rate your trust on a 1 to 5 scale where 1 indicates that you "Not at all" trust the diagnosis and 5 indicates you "Very much" trust the aid's diagnosis. After you make your responses, you will receive immediate feedback regarding the accuracy of your decision, and your score will be updated.

You will start with 100 points. You will receive +1 point if you correctly CONFIRM a target, or correctly DISMISS a false diagnosis. Conversely, you will be deducted a -1 point if you wrongfully DISMISS a true target or if you CONFIRM when a target is not present.

Please take a moment to become familiar with the enemy targets. To do so, please press the "targets" icon. Enemy targets will always face towards the left, while friendly targets will always face towards the right, like one of the many pictured in front of you right now. Now take a moment to familiarize yourself with how these images look in radar photos. Look at this sample synthetic aperture radar (SAR) image. Can you spot the enemy target? If not, let the experimenter know and s/he will help you spot it.

**APPENDIX C (CONTINUED)**

[NOVICE SYSTEM DESCRIPTION]

Because of the difficult nature of the task, you will have some help detecting enemy targets from a computer system called CONTRAST DETECTOR.

What is CONTRAST DETECTOR?

CONTRAST DETECTOR is a novice automated diagnostic aid that has been designed to detect enemy military targets in intelligence images. CONTRAST DETECTOR is based upon technology used in military target detection over the past 10 years. CONTRAST DETECTOR was designed and developed at a small technical college in the Midwest that contains a small department in military target detection. CONTRAST DETECTOR currently possesses a limited database of the types of modern weapons and targets commonly found in today's military operations. Its algorithms are relatively ineffective in their attempts to detect enemy targets. Recent testing indicates that the accuracy, dependability, and robustness of CONTRAST DETECTOR are not up to military standards for military target detection. The U.S. Department of Defense (DOD) is considering whether to conduct limited field-testing using CONTRAST DETECTOR.

If present, CONTRAST DETECTOR will indicate how confident it is that the image contains an enemy target. Note: CONTRAST DETECTOR's confidence estimates are based on how well the information collected from CONTRAST DETECTOR's algorithms match enemy templates located in its target database.

A 75% confidence estimate indicates that CONTRAST DETECTOR has considerable evidence that a target is present.

A 50% confidence estimate indicates that CONTRAST DETECTOR has variable evidence that a target is present.

A 25% confidence estimate indicates that CONTRAST DETECTOR has little evidence that a target is present.

Remember, on some trials CONTRAST DETECTOR may not be present. When it is present, the use of CONTRAST DETECTOR is completely optional. The responsibility of the final decision is ultimately your own; you can choose to either accept the aid's diagnosis or ignore it. Because of the "fog of war" CONTRAST DETECTOR may not always be correct.

Now, you must answer several questions to make sure that you understand the task and the background of the diagnostic aid that will assist you in the detection task.

## APPENDIX C (CONTINUED)

## [EXPERT SYSTEM DESCRIPTION]

Because of the difficult nature of the task, you will have some help detecting enemy targets from a computer system called SUPER CONTRAST DETECTOR.

What is SUPER CONTRAST DETECTOR?

SUPER CONTRAST DETECTOR is an expert automated diagnostic aid that has been designed to detect military targets in intelligence images. SUPER CONTRAST DETECTOR is based upon, but far exceeds, technology that the U.S. Military has used in military target detection over the past 10 years. SUPER CONTRAST DETECTOR was designed and developed by the nations top military research firm in Washington D.C. that contains a highly specialized department in military target detection. SUPER CONTRAST DETECTOR possesses an extensive database of the types of modern weapons and targets found in today's military operations. Its algorithms are highly effective in their attempts to detect enemy targets. Recent testing indicates that the accuracy, dependability, and robustness of SUPER CONTRAST DETECTOR set the standard for military target detection systems. The U.S. Department of Defense (DOD) is currently using SUPER CONTRAST DETECTOR in its Middle Eastern military operations.

If present, SUPER CONTRAST DETECTOR will indicate how confident it is that the image contains an enemy target. Note: SUPER CONTRAST DETECTOR's confidence estimates are based on how well the information collected from SUPER CONTRAST DETECTOR's algorithms match enemy templates located in its target database.

A 75% confidence estimate indicates that SUPER CONTRAST DETECTOR has considerable evidence that a target is present.

A 50% confidence estimate indicates that SUPER CONTRAST DETECTOR has variable evidence that a target is present.

A 25% confidence estimate indicates that SUPER CONTRAST DETECTOR has little evidence that a target is present.

Remember, on some trials SUPER CONTRAST DETECTOR may not be present. When it is present, the use of SUPER CONTRAST DETECTOR is completely optional. The responsibility of the final decision is ultimately your own; you can choose to either accept the aid's diagnosis or ignore it. Because of the "fog of war" SUPER CONTRAST DETECTOR may not always be correct.

Now, you must answer several questions to make sure that you understand the task and the background of the diagnostic aid that will assist you in the detection task.

## APPENDIX D

## POST INSTRUCTION QUESTIONNAIRE

Instructions: Please answer the following questions about the detection aid whose decisions will assist you in the target detection task. You must answer each question correctly before continuing. You may refer to the information you just read if you need to.

1) What is the name of the computerized detection aid?
   a. Contrast Detector
   b. Super Contrast Detector

2) The computerized detection aid is an/a
   a. Novice
   b. Expert

3) The detection aid's knowledge / database of modern military weapons and targets in target detection is:
   a. Limited
   b. Extensive

4) The detection aid is being used in <u>current</u> military efforts in the Middle East.
   a. True
   b. False

5) The detection aid's confidence estimates are based on the degree of match between the data its algorithms collect and the target templates contained in its database.
   a. True
   b. False

# APPENDIX E

## INITIAL TRUST QUESTIONNAIRE

### [NOVICE SYSTEM]

**INSTRUCTIONS:** Please respond to the following statements addressing your impressions of SUPER CONTRAST DETECTOR on a scale of 1 to 7 with 1 being "strongly disagree" and 7 being "strongly agree."

1) CONTRAST DETECTOR is likely to
   i. Use underhanded tactics (i.e. random guesswork) to arrive at diagnosis _____
   ii. Behave in a deceptive manner _____

2) I am suspicious of CONTRAST DETECTOR's diagnostic potential _____

3) I have little or no confidence in CONTRAST DETECTOR's ability to formulate accurate diagnoses _____

4) CONTRAST DETECTOR comes across as having integrity _____

5) CONTRAST DETECTOR'S decisions are likely to be consistent _____

6) CONTRAST DETECTOR is likely to be dependable _____

7) CONTRAST DETECTOR is likely to be reliable _____

8) I have faith in CONTRAST DETECTOR's ability to generate correct diagnoses _____

9) I will feel comfortable and familiar using CONTRAST DETECTOR ___

10) I can trust CONTRAST DETECTOR _____

**APPENDIX E (CONTINUED)**

# INITIAL TRUST QUESTIONNAIRE

## [EXPERT SYSTEM]

**INSTRUCTIONS:** Please respond to the following statements addressing your impressions of SUPER CONTRAST DETECTOR on a scale of 1 to 7 with 1 being "strongly disagree" and 7 being "strongly agree."

1) SUPER CONTRAST DETECTOR is likely to
   i. Use underhanded tactics (i.e. random guesswork) to arrive at diagnosis _____
   ii. Behave in a deceptive manner _____

2) I am suspicious of SUPER CONTRAST DETECTOR's diagnostic potential _____

3) I have little or no confidence in SUPER CONTRAST DETECTOR's ability to formulate accurate diagnoses _____

4) SUPER CONTRAST DETECTOR comes across as having integrity _____

5) SUPER CONTRAST DETECTOR'S decisions are likely to be consistent_____

6) SUPER CONTRAST DETECTOR is likely to be dependable _____

7) SUPER CONTRAST DETECTOR is likely to be reliable _____

8) I have faith in SUPER CONTRAST DETECTOR's ability to generate correct diagnoses _____

9) I will feel comfortable and familiar using SUPER CONTRAST DETECTOR___

10) I can trust SUPER CONTRAST DETECTOR _____

**APPENDIX F**

# OVERALL TRUST QUESTIONNAIRE

## [NOVICE SYSTEM]

**Instructions**: Please respond to the following statements addressing your impressions of CONTRAST DETECTOR on a scale of 1 to 7 with 1 being "strongly disagree" and 7 being "strongly agree."

1) CONTRAST DETECTOR is likely to use underhanded tactics (i.e. random guesswork) when diagnosing targets. _____

2) CONTRAST DETECTOR behaves in a deceptive manner _____

3) I am suspicious of CONTRAST DETECTOR's target diagnoses _____

4) I am wary of CONTRAST DETECTOR's target diagnoses _____

5) CONTRAST DETECTOR's diagnoses are likely to have a harmful outcome _____

6) CONTRAST DETECTOR is a dependable target detection aid _____

7) CONTRAST DETECTOR is a competent target detection aid _____

8) CONTRAST DETECTOR is a reliable target detection aid _____

9) I have faith in CONTRAST DETECTOR's diagnoses _____

10) The answer provided by CONTRAST DETECTOR is predictable _____

11) I feel comfortable and familiar using CONTRAST DETECTOR _____

12) I can trust CONTRAST DETECTOR _____

13) CONTRAST DETECTOR is credible _____

14) One a scale from 0% -100%, please indicate how reliable you think CONTRAST DETECTOR was at identifying targets: _____

## APPENDIX F (CONTINUED)

## OVERALL TRUST QUESTIONNAIRE

## [EXPERT SYSTEM]

**Instructions**: Please respond to the following statements addressing your impressions of SUPER CONTRAST DETECTOR on a scale of 1 to 7 with 1 being "strongly disagree" and 7 being "strongly agree."

1) SUPER CONTRAST DETECTOR is likely to use underhanded tactics (i.e. random guesswork) when diagnosing targets. _____

2) SUPER CONTRAST DETECTOR behaves in a deceptive manner _____

3) I am suspicious of SUPER CONTRAST DETECTOR's target diagnoses _____

4) I am wary of SUPER CONTRAST DETECTOR's target diagnoses _____

5) SUPER CONTRAST DETECTOR's diagnoses are likely to have a harmful outcome _____

6) SUPER CONTRAST DETECTOR is a dependable target detection aid _____

7) SUPER CONTRAST DETECTOR is a competent target detection aid _____

8) SUPER CONTRAST DETECTOR is a reliable target detection aid _____

9) I have faith in SUPER CONTRAST DETECTOR's diagnoses _____

10) The answer provided by SUPER CONTRAST DETECTOR is predictable _____

11) I feel comfortable and familiar using SUPER CONTRAST DETECTOR _____

12) I can trust SUPER CONTRAST DETECTOR _____

13) SUPER CONTRAST DETECTOR is credible _____

14) One a scale from 0% -100%, please indicate how reliable you think SUPER CONTRAST DETECTOR was at identifying targets: _____

# APPENDIX G

## OPINION QUESTIONNAIRE

Part. #: _____ Group: _____ Date: _____ Time: _____

Thank you for participating in this research project. Please complete the following items by entering the number of your choice on the answer sheet. As before, your answers are completely confidential.

**Please rate the target detection game on the following dimensions:**

1. Stress:

    1. Very Stressful
    2. Slightly Stressful
    3. Neither Stressful Nor Relaxing
    4. Slightly Relaxing
    5. Very Relaxing

2. Complexity:

    1. Very Understandable
    2. Slightly Understandable
    3. Neither Understandable Nor Complex
    4. Slightly Complex
    5. Very Complex

3. Simplicity:

    1. Very Challenging
    2. Slightly Challenging
    3. Neither Challenging Nor Simple
    4. Slightly Simple
    5. Very Simple

4. Stimulation:

    1. Very Stimulating
    2. Slightly Stimulating
    3. Neither Stimulating Nor Boring
    4. Slightly Boring
    5. Very Boring

5. Did you have a strategy for searching for enemy targets?_____
If so, what was it?_____

_____

_____

_____

_____

_____


6. Was is too difficult to spot enemy targets? If so, how did you determine whether a target was present?

_____

_____

_____

_____

_____

_____


7. Did you have a strategy for using the diagnostic aid's advice? Where there certain instances where you were more likely to rely on the aid's advice compared to others?

_____

_____

_____

_____

_____

_____


8. Do you have any other thoughts, feelings, or comments about this experiment?

_____

_____

_____

_____

_____

_____

# APPENDIX H

## MANIPULATION CHECK

A 2 (Group: experimental, control) x 4 (System Confidence: 75%, 50%, 25%, no aid) mixed ANOVA was calculated to ensure that system confidence, not event base rate, influenced participants' response frequency. Response frequency was defined as the proportion of times a participant reported that a target was present. Results revealed a significant interaction between group and system confidence, $F(3, 312) = 31.71$, $p < .001$, partial $\eta^2 = .23$. Participants who received automated assistance matched their response frequency to the system's level of confidence, whereas participants in the control condition reported a target being present roughly 50% of the time. These results suggest that system confidence, not target base rate, influenced participants' response frequency.
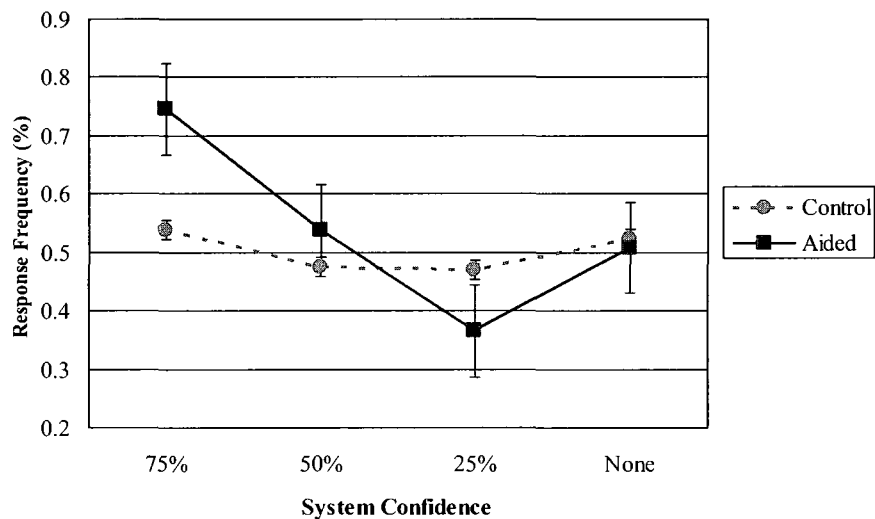
# VITA

**Randall D. Spain, Ph.D.**
**ODU Psychology Department**
**250 Mills-Godwin Building**
**Norfolk, VA 23529**

---

## EDUCATION

Old Dominion University, Norfolk, VA                                             August 2009
**Doctor of Philosophy** (Human Factors Psychology)

Old Dominion University, Norfolk, VA                                             December 2006
**Master of Science** (Experimental Psychology)

Christopher Newport University, Newport News, VA                                 May 2003
**Bachelor of Arts with Honors**, Magna Cum Laude (Psychology)

---

## EXPERIENCE

- **Research Scientist (Independent Contractor),** U.S. Army Research Laboratory, Suffolk VA, (2007 – 2009)
- **Graduate Research Assistant,** Research Environment for Alarms and Complex Task Simulation (REACTS), Norfolk VA, (2004 – 2009)
- **Graduate Research Assistant,** Virginia Modeling Analysis and Simulation Center (VMASC), Suffolk VA, (2005 – 2007)

---

## SELECTED PUBLICATIONS

Spain, R. D., & Bliss, J. B. (2008). The effects of sonification pulse rate and reliability on operator workload and trust during a simulated patient-monitoring task, *Ergonomics, 51*, 1320-1337.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. In *Proceedings of the 52th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1335-1339). New York, NY: Human Factors and Ergonomics Society.

Spain, R. D., Bliss, J. P., & Newlin, E. T. (2008). A multilevel analysis of operator trust in sonification systems [Abstract] *The 23rd Annual Meeting of the Society for Industrial and Organizational Psychologists. San Francisco, CA.*

Hansberger, J. T., Schreiber, C., & Spain, R. D. (2008). Analysis of C2 Distributed Cognition. In the Proceedings: *13th International Command and Control Research and Technology Symposium*, (Bellevue, WA).

Spain, R. D., Bliss, J. P., & Newlin, E. T. (2007). The effect of sonification pulse rate on perceived urgency and response behaviors. In the Proceedings: *51st Annual Meeting of the Human Factors and Ergonomics Society, pp 116-120.*

Bliss, J. P., and Spain, R. D. (2007). Sonification and reliability – implications for signal design. In *Proceedings of the 13th International Conference on Auditory Display,* (Montreal, Canada). pp. 154-159.