ABSTRACT

AIR TRAFFIC CONTROLLER TRUST IN AUTOMATION IN NEXTGEN

By

Tannaz Mirchi

August 2015

NextGen introduces new automated tools to help air traffic controllers (ATCos) manage the projected increase in air traffic over the next decades. The purpose of the current study was to assess the role of trust in automation for NextGen tools. Differences in sensitivity between three subjective trust in automation scales and the relationship of these trust metrics to ATCo trust behaviors were considered. Trust behaviors were measured using a behavioral measure of trust, the number of near-miss aircraft moved. Additionally, the relationship between trust levels and situation awareness was also investigated. Results indicated that the Modified Human-Automation Trust Scale (M-HAT) may be the most sensitive to changes in trust over the course of the internship, although there was no differences in trust behavior between low or high-trusting individuals. Trust questionnaires pertaining to an overall automated system (M-HAT) may able to detect changes in trust over time compared to a more specific trust scale. The results also suggest it may be more valuable to specifically train controllers to trust automation than provide general training.

AIR TRAFFIC CONTROLLER TRUST IN AUTOMATION IN NEXTGEN

A THESIS

Presented to the Department of Psychology

California State University, Long Beach

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Psychology

Option in Human Factors

Committee Members:

Thomas Z. Strybel, Ph.D. (Chair)
Kim-Phuong Vu, Ph.D.
James Miles, Ph.D.

College Designee:

Amy Bippus, Ph.D.

By Tannaz Mirchi

B.A., 2012, California State University, Long Beach

August 2015

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

TABLES Page

LIST OF FIGURES

FIGURES Page

CHAPTER 1

INTRODUCTION

The Next Generation Air Transportation System (NextGen) is a new automated system expected to accommodate the projected increase in air traffic in the National Airspace System (NAS; Joint Planning and Development Office [JDPO], 2010). With the introduction of NextGen into the NAS over the next decades, it is important to assess the role of trust in automation for air traffic controllers (ATCos). In order to ensure that the introduction of more automated tools for ATCos will be optimally used, it is essential to understand the individual differences of trust in automation, and its impact on training and controller performance. In this thesis I will review the empirical evidence on how trust in automation is defined, how trust in automation is measured, factors known to influence trust in automation, and effective training methods for developing proper trust in automation. I will also review the effects of trust on ATCo performance and situation awareness (SA). Lastly, I will describe an experiment that examined the differences between subjective trust scales and their relationship to trust behavior.

NextGen

NextGen is expected to accommodate the projected increase of air traffic in the NAS. The NextGen system should create a more efficient, safe, cost effective, and environmentally friendly approach to meet future aviation demands. To achieve these benefits, ATCos, and pilots will need new automation tools for managing traffic. New tools such as automated conflict detection and resolution will enable ATCos to resolve potential aircraft conflicts faster. Data Comm will reduce the number of controller-pilot verbal communications. These tools should increase safety through more accurate transfer of information and a reduction in operator workload (Kiken et al., 2011). With NextGen, airports should have less congestion and ATCos and pilots will have better tools to identify and mitigate potential hazards. NextGen tools should provide many benefits, but these benefits such as improving air traffic control SA and safety of the air traffic management (ATM) system, will only be seen if the tools are being used properly. For proper use of automation, the operator must be aware of and understand several characteristics of the automation tools such as their current reliability, false alarm rate, and miss rate. In other words, the operator's trust must be properly calibrated to the characteristics of the automation. Calibration is the connection between an operator's perception of the reliability of an automated system and its reliability (Lee & See, 2004). When the operator is not appropriately calibrated, negative consequences may result.

Trust in Automation

Trust can be defined as the predictability of another entity. Within the social psychology literature, trust is defined as the predictability of another person (Deutsch, 1958; Eckel & Wilson, 2004; Ergenli, Saglam, & Metin, 2007). In a general context, trust is defined as a dispositional viewpoint of people and the world (Rotter, 1967). In a more situation-specific context, interpersonal trust is defined as socially learned expectations dependent on social order (Barber, 1983) and trust as a collection of viewpoints of others (Pruitt & Rubin, 1986). These definitions focus on trust as a way of thinking, while others defined trust as a willingness to accept vulnerability (Mayer, Davis, & Schoorman, 1995). According to Rempel, Holmes, and Zanna (1985), there are three dimensions of interpersonal trust: predictability, dependability, and faith, which impact the willingness to believe any details provided by source. The predictability of an individual is reliant on his or her consistency and stability of actions over a period of time. Dependability is contingent upon the individual's internal behavioral characteristics, which influence one's confidence in them. Faith is dependent on expectations of future actions and accuracy of the individual. With these three dimensions, one can have better developed trust for individuals.

Various studies have looked at how these ideas of trust between humans also apply to automation. Automation is "technology that actively selects data, transforms information, makes decisions, or controls processes" (Lee & See, 2004, p. 50). Typically, automation is used for collecting and examining information, decision-making, performing actions, and supervising other systems. Many studies have shown the influence of automation in numerous fields such as aviation (Dixon & Wickens, 2006; Lee & Moray, 1994; Parasuraman, Molloy, & Singh, 1993; Parasuraman, Sheridan, & Wickens, 2000). Overall, researchers agree that proper usage of automation is positively correlated with operators trust in automation and with a higher level of trust comes more reliance on the automation (Geels-Blair, Rice, & Schwark, 2013). Hoff and Bashir's (2015) systematic empirical review on trust in automation described a common theme among the explanation of trust across various fields of research. Typically, interpretations of trust include three components. First, there must be a truster to give trust, a trustee to receive trust, and there must be some form risk to take. Second, the trustee must be motivated or enticed to perform the task either by some form of gain or their compassion. Related to technology, this usually means the benefit of using the system. Third, the trustee must be insecure about the possibility of failing the task. Although separate concepts, the components of trust mentioned above can be related to both interpersonal and human-machine relationships.

A number of studies have demonstrated the similarities between human-human relationships and human-machine interactions. One study carried out a series of experiments to show that humans perceive computer characteristics similar to the way they perceived human's interpersonal characteristics (Reeves & Nass, 1996). The experiment had participants take a quiz on a short factual presentation. Following the quiz, the participants were divided into two groups. All participants were provided with positive feedback, but one group was provided feedback on the computer they had just taken the quiz on while the other group was provided with feedback on a separate computer. As expected, the group who received feedback on the computer they had just taken the quiz on rated the computers more positively than those who had feedback from a separate computer. Previous research agrees that trust is a mediator in human-machine relationships just as it is in human-human relationships (Sheridan, 1975; Parasuraman et al., 1993). Therefore, unique social connections between humans are also seen with machines.

Human-machine trust is defined as a dynamic expectation that undergoes predictable changes as a result of experience with the system (Muir, 1987). Some common issues with human-machine interaction and trust can be overtrust and distrust. Overtrust is defined as trust exceeding the capabilities of the system, which may lead to misuse (Lee & See, 2004). On the other hand, distrust is defined as trust falling short of system capabilities, which can lead to disuse (Lee & See, 2004). Together these play a critical role in the appropriate calibration of an operator's amount of human-automation reliance.

Lee and Moray (1992) developed a model that is similar to Rempel et al.'s (1985) three dimensions of human trust in automation: performance, process, and purpose. Performance is in relation to the past, present, and future operation of a system in regards to predictability, reliability, and ability. If the automation performs as expected by the operator to achieve his or her goals, then the operator is more likely to trust the automation, a robust concept according to Sheridan (1992). Process refers to the automation algorithms ability to accomplish the operator's goals in an appropriate manner. And finally, purpose represents the designer's underlying motivations for the operator's use of the automation. Lee and Moray (1992) proposed performance was related to Rempel et al. (1985) predictability dimension, process related to dependability, and purpose related to faith. As noted by Madhavan and Wiegmann (2007), although performance and process refer to the dimensions of predictability and dependability in the Rempel et al. (1985) model, purpose does not completely relate to the third dimension of faith in the model. See Table 1 for a summary of the proposed relationship between the various dimensions of trust. While many definitions of trust exist, there are still inconsistencies concerning the true definition of trust. It is not surprising therefore, that there is little agreement on how to measure human trust in automation.

TABLE 1. Summary of Relationships of Various Trust Dimensions

	Barber (1983)	Rempel, Holmes, Zanna (1985)
Purpose	Fiduciary responsibility	Faith
Process		Dependability
Performance	Technically competent performance	Predictability
Foundation	Persistence of natural laws	

Trust Measures

Given the importance of trust, researchers have developed measures of the construct. Valid measures of trust in automation are essential in order to ensure that automation is used appropriately. Several methods for measuring trust in automation have been reported in the literature, most of which are subjective measures of trust. The Human-Automation Trust (HAT) Scale is an empirically based tool to measure people's trust in automated systems (Jian, Bisantz, Drury, & Llinas, 2000). Previously, theoretical scales had been developed which did not empirically consider the relationship between various types of trust such as general trust, human trust, and automation trust (see Table 2).

TABLE 2. Types of Trust

Trust Type	Definition
General	Overall propensity to be a trusting individual to various entities.
Human Trust	Trust for relationships with friends, families, and romantic partners.
Automation Trust	Trust for relationships between a human and the use of automated machines or systems.

Note. Various levels of trust defined (Jian et al., 2000).

It is important to consider if the relationships between these types of trust are similar because trust measures may be shared across domains.

Jian et al. (2000) tried to bridge this gap in trust scales by developing an empirical measurement of trust with a three-phased experiment. The first phase was a word elicitation study to create an extensive trust and distrust word bank. The second phase was a questionnaire study to narrow down the word bank, distinguish if trust and distrust have a negative relationship, and if these two concepts are the same or different for different types of trust (general trust, human trust, and automation trust). The final phase was a paired-comparison study aimed at developing a multi-dimensional trust measurement scale using a factor analysis. Results showed that strong negative correlations existed between trust and distrust ratings, therefore developing separate measures of high and low trust and distrust was not needed. Trust and distrust could be considered opposite ends of the same concepts and the measurement of operator's trust or distrust in automation can be assessed using the same rating scale. The trust words generated were similar for each of the three trust types previously mentioned (see Table 2). These findings imply that concepts of trust are not perceived differently across various types of relationships (Jian et al., 2000). Additionally, cluster analysis distinguished 12 possible factors that determine trust between humans and systems. The 12 factors were used to create the HAT Scale, which is specific to the ATM system as a whole. The scale includes a total of 12 items, using a 7-point Likert scale, with higher scores meaning high trust levels, and lower scores, low trust levels. Other researchers have used this original scale previously as an empirically validated measurement of trust in automation (Kunii, 2006). Kunii (2006) developed a modified version of the HAT to address negative attitudes towards the original wording of the questions during pilot testing. Some of the wording was changed without altering the meaning of the questions to create the Modified Human-Automation Trust (M-HAT) Scale, also containing 12 items (see Table 3). The present study used this Modified version of the HAT Scale (Kunii, 2006).

TABLE 3. Twelve Factors from Modified Human-Automation Trust Scale (M-HAT) and Example Questions

Modified Factor[1]	Words from Original Cluster Analysis[2]	Modified Scale[1]
Deceptive	Deception, Lie, Falsity, Betray, Misleading, Phony, Cheat	The system can be deceptive
Unpredictable	Sneaky, Steal	The system sometimes behaves in unpredictable manner
Suspicious	Mistrust, Suspicion, Distrust	I am often suspicious of the system's intent, action, or outputs
Unsure	Beware	I am sometimes unsure of the system
Harmful	Cruel, Harm	The system's action can have a harmful or injurious outcome
Confident	Assurance, Confidence	I am confident in the system
Security	Security	The system can provide security
Integrity	Honor, Integrity	The system has integrity
Dependable	Fidelity, Loyalty	The system is dependable
Consistent	Honesty, Promise, Reliability, Trustworthy, Friendship, Love	The system is consistent
Trust	Entrust	I can trust the system
Familiar	Familiarity	I am familiar with the system

Note. Factors used to create scale items for M-HAT Scale. Scale items were rated on 7-point Likert scale: 1- *not at all* to 7- *extremely* (Kunii, 2006[1]; Jian et al., 2000[2]).

The Complacency-Potential Rating Scale (CPRS) was developed for assessing general trust in automation to everyday systems such as automated teller machines and automatic cruise control (I. L. Singh, Molloy, & Parasuraman, 1993). The scale was developed in three phases. The first phase involved generating questions to probe positive and negative attitudes about automation. Factor analysis was used to analyze the inter-item correlation matrix to assess the statistical significance of the items and get reliability measures. Next, an initial validation study was performed to assess the interconnections of the complacency scale with variables such as age and gender. The final phase was a second administration of the scale to the sample in order to determine the test-retest reliability. The CPRS (I. L. Singh et al., 1993) employs a 5-point Likert rating and shown to have a high internal consistency (*r* = .87) and test reliability (*r* = .90). High propensity to trust is depicted by higher scores and low propensity to trust by lower scores. For this study, I used a version of the original CPRS with 12 items on a 5-point Likert scale to measure the general inclination to trust systems, not specific to any particular type of system (Merritt & Ilgen, 2008). The scale is based on four factors: trust, safety, confidence, and reliance and has been previously considered a scale designated to measure the level of propensity to trust automated systems in general (Merritt & Ilgen, 2008). These four factors can also be found in Jian et al. (2000) 12-factor HAT scale (trust, security, confident, and reliability.)

A version of the CPRS (I. L. Singh et al., 1993) was also used in the current study (Verma, Kozon, Ballinger, Lozito, & Subramanian, 2011). The scale was adapted in order to make it specific to trust in automated ATM tools (Verma et al., 2011). Questions were modified to be specific to NextGen tools such as conflict alerting, conflict probes, and Data Comm. Scale reliability was checked through question rating scale reversals. The scale has a total of 36 items, using a 5-point Likert scale that ranges from strongly disagree to strongly agree. High scores show higher levels of trust in NextGen tools while lower scores show lower levels of trust. For the remainder of the paper, Complacency-Potential Rating Scale (I. L. Singh et al., 1993) will be referred to as CPRS, Complacency Potential Rating Scale (Verma et al., 2011) as ATM-CPRS, and Modified Human-Automation Trust Scale (Kunii, 2006) as M-HAT (see Table 4 for a summary).

TABLE 4. Three Subjective Trust Scales

Trust Scale	Abbreviation	Type	Measures	Reference
Complacency-Potential Factor Rating Scale	CPRS	General automation trust	Internal consistency (*r* = .87), test reliability (*r* = .90).	I. L. Singh et al. (1993)
Complacency Potential Factor Rating Scale	ATM-CPRS	Specific to NextGen tools	Not reported	Verma et al. (2011)
Modified Human-Automation Trust Scale	M-HAT	ATM system	Not reported	Kunii (2006)

The three subjective trust scales (CPRS, ATM-CPRS, M-HAT) used in this study were related to various forms of trust. The CPRS was a general scale used to look at one's propensity to trust automation without relating to one particular system (I. L. Singh et al., 1993). The ATM-CPRS took a more specific stance and assessed trust related to a specific system and the automation used such as NextGen automated tools (Verma et al., 2011). The M-HAT was only specific to one system, ATM, but did not probe participants on specific automation within the system (Kunii, 2006). The three scales provide breadth of subjective trust measures to examine if one scale may be more sensitive to changes in trust over time. The current study attempted to assess trust both subjectively with trust scales and objectively with behavioral measures. However, there have been very limited attempts at developing behavioral measures of trust in automation. Only one study attempted to create a behavioral measure of trust in automation for NextGen (Higham, 2013). Higham (2013) used the number of near-miss aircraft to measure ATCo trust in the NextGen automated conflict detection tool. Near misses were characterized as two equipped aircraft within 6-10 nautical miles (nm) but did not lose separation (less than 5 nm laterally and 1,000 ft vertically). Therefore, by moving these aircraft, the participant shows mistrust in the automated conflict detection tool. Results showed that ATCos who were trained to trust automation moved less near-miss aircraft by the end of their internship. However, this experiment was limited because only one near-miss aircraft pair was provided in 50% mixed-equipage test scenarios to objectively measure trust.

The current study also used near-miss aircraft as an objective behavioral measure of trust. Through these subjective and behavioral measures, the study will be able to further establish if a relationship exists between the subjective trust ratings and what the controllers actually portray as their trust levels by moving or not moving near-miss aircraft. The present study will also further investigate findings of Higham (2013) by providing more opportunities for ATCos to show trust in automation with NextGen tools by increasing the number of opportunities to move near-miss aircraft to three per mixed-equipage scenario.

Factors Influencing Trust in Automation

The development of trust in automation is a complex and dynamic process, which involves several factors and sources of variability. Various aspects guide the decisions to trust automation, but Lee and See (2004) state emotions fall at the crux of establishing trusting behavior in automation. Some researchers agree the reliability of the system as well as the operator's predisposition to trust automated technologies guide this process (Lee & Moray, 1992; Muir, 1988). Other evidence suggests factors such as operator workload and risk levels of the situation play a part in the decision to use automation (Riley, 1989). In addition to factors that have an influence on the development of trust in automation there is also variance for developing trust in automation.

Based on an empirical review of 127 studies related to human trust in automation, Hoff and Bashir (2015) identified three broad sources of variability in trust. These are based on the human operator, the environment, and the automated system. The human operator may be a source of variability in trust due to factors such as gender, age, culture, personality, and individual's tendency to trust automation. The environment can play a role in the variability of the formation of trust based on external factors such as task difficulty, perceived risk, and organization setting. The automated system itself may alter trust formation due to the operator's previous experience (initial learned trust) or current interactions (dynamic learned trust) with the automated system.

Along with multiple sources of variability involved in the formation of trust, some trust metrics recognize four components of trust responsible for the individual's development of trust in automation. These four factors are represented in the CPRS and ATM-CPRS (I. L. Singh et al., 1993; Verma et al., 2011) as *trust*, *safety*, *confidence*, and *reliance*. The M-HAT also recognizes these four factors among the 12-item scale but uses different terms for *safety* (*security)* and *reliance* (*consistent)*.   *Trust* is related to an individual's trust in automation. For example, an individual with high levels of trust would rather place an order online than over the phone with a sales representative because they believe the order will be more accurate with a computer. *Safety* is the individual's awareness of the perceived safety supplied from automation. For example, an individual with high levels of trust would feel more comfortable using an automated teller machine than a human teller at the bank. The *confidence* factor is related to the potential for complacency or having overconfidence in automation. For example, believing robotic surgery is more reliable and safe than manual surgery. Finally, the *reliance* factor corresponds to ones reliance based on their understanding of the reliability of the automation. For example, by using an automated teller machine, an individual's bank account will be less likely to be subject to fraudulent use. Together these factors describe the proposed main components involved with automation-induced complacency.

Training Trust in Automation

The operator's beliefs and knowledge of the purpose and process of the automation can influence the formation of trust (Hoff & Bashir, 2015). Without the appropriate information on the purpose of an automated system or the role of the automation, operators lack the ability to fully and accurately trust automation as intended by the system designers. By training operators to hold the appropriate level of trust in automation, the occurrence of misuse and disuse should diminish (Lee & See, 2004). Training operators on the actual reliability of automation can lead to changes in operator trust and reliance (Hoff & Bashir, 2015). According to Parasuraman and Riley (1997), training for automation use should focus on the way the automation works as well as the principles behind the design of the automation technology. Training design should incorporate strategic and rational decision-making on when to use or not use automation (Parasuraman & Riley, 1997). Operators should be specially trained to appropriately deal with the conflicting demands automation presents such as passive monitoring versus active control (A. L. Singh, Tiwari, & Singh, 2009). Through evidence and understanding of the implications of trust in automation, future designers and supervisors will be better equipped to provide valuable automation training methods to operators.

Currently, very little research has focused on training trust in automation; most training research has focused on how to use an automated tool. Specifically, training effectiveness has been looked at with varying training lengths and methods on automated monitoring tasks. One study further examined the effects of training on automation-induced complacency (failure to detect automation malfunctions) by employing a flight simulation task (A. L. Singh et al., 2009). The flight simulator included engine system monitoring, fuel resource management, and tracking tasks. For the engine system monitoring task, participants were provided with an automated tool which aided them in detecting failures of automation, with varied reliability. In addition to the ability to detect automation failures, workload was assessed using the NASA Task Load Index (NASA-TLX). The NASA-TLX is a multi-dimensional rating procedure used as a subjective measure of the participant's workload (Hart & Staveland, 1988). Training time (30 minutes vs. 60 minutes) and Automation Reliability (Low-25%, Moderate-50%, High-87.5% detection rates) were manipulated to investigate their impact on operator complacency and workload. The researchers expected more training would lead to a decrease in automation-induced complacency, measured through the ability to detect automation failures, while high reliability would lead to an increase in automation-induced complacency and decrease in mental workload. The results indicated that the length of automation training did not significantly impact operator performance on the flight simulation tasks, but did lead to a reduction of workload for three NASA-TLX workload dimensions (temporal demand, frustration, and mental effort). Results also showed that low automation reliability reduced operator complacency (more failures were detected) while high automation reliability led to increased complacency. These results suggest that training an operator for a longer period of time on an highly reliable automated task can facilitate increased levels of trust, which may increase automation-induced complacency while reducing mental workload.

Previous research examined training trust in automation for NextGen (Higham, 2013). A group of 15 student ATCos participating in a 16-week internship were divided into two training classes, Trust Training or No-Trust Training. The Trust Training group was provided with verbal feedback from the instructor if they moved an equipped aircraft that came close to losing separation but did not. The feedback was provided in order to further develop student knowledge of the dependability of the NextGen conflict detection tool. The No-Trust Training group received no feedback. Higham (2013) showed that trust training may be feasible because the Trust Training group was less likely to move the near-miss aircraft by the final exam compared with the midterm exam, especially for 50% mixed-equipage scenarios. Additionally, the Trust Training group had lower workload on 100% automated scenarios containing NextGen-equipped aircraft.

Individual Differences in Trust and Training

Individual differences in automation use are prevalent due to the complex nature and various contributing factors to trust in automation. Merritt and Ilgen (2008) proposed that just as people have a general inclination to trust people, they might also have individual tendencies to trust or distrust machines. Participants were instructed to use an Automatic Weapons Detector (AWD) machine during a medium fidelity X-ray Screening Task to make decisions while screening luggage for weapons (Merritt & Ilgen, 2008). Initial and post-task trust was rated using a 6-item scale developed for the study and propensity to trust machines was measured using 12 items from the CPRS (I. L. Singh et al., 1993). Results showed that the propensity to trust interacts with the characteristics of the automation (Merritt & Ilgen, 2008). When a high trusting individual uses faulty or unreliable automation, subsequent trust will be negatively impacted based on the perceptions of this system. Ultimately, different operators will perceive the same system differently. A person who has a higher propensity to trust the system will expect the system to be reliable. With this schema in mind, he/she will have a stronger reaction to any system errors and be more likely to remember these errors in the future. Therefore, training design for future use of NextGen tools by ATCos should consider the interaction of individual differences and machine characteristics.

Situation Awareness and Trust

Trust in automation may also affect the operator's SA. SA is the comprehension required to control a complex system in a dynamically changing environment (Durso & Gronlund, 1999). In order to maintain SA, one must continuously comprehend what is currently happening and the relationship between information, events, and actions with current and future goals and plans. Air traffic control, a high information flow environment, is one domain that requires exceptional SA to safely and efficiently carry out the task. ATCos need to maintain an awareness of the aircraft's altitude, heading, speed, and flight path within their sector. ATCos' SA is typically referred to as maintaining "the picture." Previous studies have shown that most ATCo errors are due to failures of SA or errors in SA, namely a breakdown of perception of critical information, comprehension, and projection of the role of this information in the future (Jones & Endsley, 1996; Rodgers, Mogford, & Strauch, 2000).

Ensuring ATCos maintain appropriate levels of SA and trust in automation will be essential for successful implementation of NextGen automated tools. With NextGen, many responsibilities of the controller will be shifted to the pilots or automated agents while the ATCo serves more of a supervisory role over the sector. One major difference between current-day ATM and new ATM systems is ATCos responsibility for active control versus passive monitoring (Metzger & Parasuraman, 2001). Currently, ATCos are fully responsible for ensuring that aircrafts maintain separation by issuing verbal commands. NextGen may create a more passive monitoring role for controllers, which allows them to issue commands automatically and be responsible for overriding the automated system in only urgent situations.

Monitoring automation may lead to automation-induced complacency and overtrust. With automation, operators are highly likely to because more complacent, or over-reliant on the automated properties of the system. Automation complacency has been related to putting a high level of trust in an automated system (Parasuraman et al., 1993). Operator complacency has also been connected with the demands of having multiple tasks within a complex and dynamic task environment such as ATC (Parasuraman et al., 1993). Shifts between various tasks can be taxing to your attentional resources and lead to complacent behavior as a coping mechanism to supplement limited attention (Endsley, 1996). As a result of automation-induced complacency, SA for the system can be reduced through misuse (Endsley, 1996). Being over-reliant on automation can have a detrimental impact to SA.

Kunii (2006) attempted to understand the relationship between trust and SA in novice pilots. The study used the M-HAT scale and a modified version of the Situation Awareness Global Assessment Technique (SAGAT) to measure trust and SA. The trust scale was administered before a flight simulation training session. Flight instructors probed 30 participants with six SAGAT questions during the simulation. The two different scenarios used in the simulation included an electrical and engine failure. It was predicted that pilots with low SA would have high levels of trust while pilots with high SA would have low levels of trust. The reverse effect, however, was seen where pilots with high trust also had the highest SA. Even though a high level of trust is placed in the system, the pilots were still able to maintain a good level of SA and did not become over-reliant on the technology. They were able to understand their flight instruments to a great extent, which allowed them to have high trust in the system while maintaining awareness. This study provided a good starting point for measuring the degree to which trust in automation may impact SA.

Present Study

The present study was a secondary analysis of a larger scale simulation study. This study examines several questions to better understand the role of trust in automated

NextGen tools. As previously discussed, understanding the factors related to trust in automation as well as measuring increased trust levels through training with the automation will allow designers, supervisors, and trainers to be better equipped for future NextGen use. Through the implementation of NextGen, many benefits, such as safety and efficiency in the NAS, can be gained with the appropriate use of this new complex system. To properly use an automated system, one must appropriately trust the automation.

The first question investigated whether the trust scales (CPRS, ATM-CPRS, M-HAT) differ in their sensitivity to changes in trust over time. The CPRS is a general trust scale that measures an individual's propensity to trust systems in general. The ATM-CPRS was a more specific trust scale measuring ones trust exclusively for NextGen tools. The M-HAT measures trust for the ATM system as a whole. Sensitivity is the ability for a measurement to detect the degree to which the variable of interest changes over time (Löwe, Kroenke, Herzog, & Gräfe, 2004). Measuring sensitivity of various trust scales can help to determine which scale may be the better choice to measure trust in automation. The study may be able to discover if administering trust questionnaires about the ATC system as a whole (M-HAT) can be more or less sensitive compared to asking about trust in specific NextGen tools (ATM-CPRS) or automated systems in general (CPRS).

The second question looked at the relationship between the behavioral trust measure of the number of near-miss aircraft moved and the subjective trust scale ratings. If the trust scales are related to the behavioral trust measures, then this provides evidence to support the notion that the scale(s) appropriately measures trust in automation. The expected relationship would be as the individual scores higher on the trust scales then they will be less likely to move near-miss aircraft since they trust the automated conflict detection tool to do its job. On the other hand, a less trusting individual will be more likely to move the near misses instead of depending on the automated tool.

The third question examined if a controllers' trust levels effect ATCo SA. It is important to determine if putting a high level of trust in an automated system can have a negative impact on SA, such as automation-induced complacency. If a relationship between SA and trust is established, the findings can suggest new methods for determining the role of automation on SA and the appropriate level of trust needed to maintain satisfactory levels of SA.

CHAPTER 2

METHOD

Participants

Twelve students (one female and 11 male) from Mount San Antonio College Federal Aviation Administration (FAA) Collegiate Training Initiative (CTI) program participated in the study as part of a 16-week radar simulation internship at the Center for Human Factors in Advanced Aeronautics Technologies (CHAAT). The average age of the students was 23.5 years and they were compensated $10.00 per hour for their participation in the study ($120 for full participation).

Design

A 3 [*Trust Scale:* CPRS (I. L. Singh et al., 1993), ATM-CPRS (Verma et al., 2011), M-HAT scale (Kunii, 2006)] x 3 (*Internship Week:* 1, 9, 16) within-subjects analysis of variance (ANOVA) was used to investigate which trust scale(s) was most sensitive to changes over time. The internship week intervals were based on the points at which the trust scales were administered to the participants (first day, midterm exam [Week 9], and final exam [Week 16]). A median split was conducted based on the trust scale that was most sensitive to changes in trust behavior over the course of the internship. High and low-trust groups were created. A 2 (*Trust:* High, Low) x 2 (*Traffic Density:* High, Low) x 2 (*Test Session:* Midterm, Final) mixed-design ANOVA was also used to examine the effect of low or high subjective trust levels on trust behavior and SA.

Traffic density was manipulated to include two high-density scenarios and two low-density scenarios in the larger scale study. Dependent variables included average near-miss aircraft moved, SPAM ready latency, SPAM probe latency, and SPAM probe accuracy. Pearson correlation analyses were also conducted to look at the relationship between trust scales, trust behavior, and SA.

Materials

Testing and training scenarios were developed using Multi Aircraft Control System (MACS). MACS is a medium fidelity software built in JAVA that provides human-in-the-loop air traffic simulation. The scenarios portrayed Indianapolis Center (ZID-91) traffic flows, departures, and arrival streams from Louisville International Airport (Prevot, 2002). During lab sessions, students switched between pseudo-pilots and ATCos. For the experiment, confederate researchers served as pseudo-pilots.

Simulation Set-Up

Each participant managed traffic for one “world” and three to four worlds were run at one time, depending on the number of participants scheduled. The world included eight computers working together to simulate the ATM scenarios. See Table 5 for a list of the computers used and their role in the simulation.

TABLE 5. Simulation Computer Stations and their Role in the Simulation

	Computer Station	Task(s)
1	Aeronautical Datalink and Radar Simulator (ADRS)	Provides communication between computers in each world
2	ATCo Radar Screen with Screen Recording Software	Radar scope to manage all traffic in ZID-91
3	Ghost Sector	Confederate pilot station to manage all traffic in ZID-91 and surrounding ZID-90
4	Pseudo-Pilot Control Screen	Confederate pilot station to gain control and execute actions to all aircraft in ZID-91 and ZID-90
5	Simulation Manager Station	Allows manager to select, start, and stop all scenarios
6	Server Station	Single computer to provide server for each of the worlds voice communication
7	Voice Recording Software	Records communications between ATCo and pseudo-pilot in each world
8	Touchscreen Probe Station	Provides SA probe queries using visual basic program and recorded responses
9	Screen Recording Software	Camtasia screen capture software at the ATCo stations to record ATCo scenarios

NextGen Tools

For equipped aircraft, students had several NextGen tools for managing traffic including Data Comm, conflict detection, and conflict probing. Data Comm enabled digital handoffs, frequency changes, and clearances. The conflict detection tool alerted the ATCo of any loss of separation (LOS) between equipped aircraft pairs that would occur in the next eight minutes by flashing aircraft in red and showing the number of minutes to LOS next to the call sign. Note that this tool was perfectly reliable for equipped aircraft. The conflict probing tool allowed the ATCo to plan a new conflict-free path for an aircraft by shading a conflict area in blue. Throughout the internship, the instructor reminded the students of the reliability and consistency of the conflict detection tool in detecting potential conflicts between equipped aircraft. Therefore, moving any non-alerting NextGen equipped aircraft for potential traffic conflicts suggested mistrust in the automated tools.

Probe Question Development and Procedure

The Situation Present Awareness Method (SPAM) was used to administer SA probe questions, which queried participants about their sector at various times throughout the testing scenario (Durso & Dattel, 2004). Probes pertained to ATM of the sector and were counterbalanced among participants. Probing began 6 minutes into the scenario and continued for 3-minute intervals through minute 38 in the scenario. Each scenario included 10 probe questions. Probe questions were presented on a separate touch screen station located to the right of the simulated radarscope where participants made their responses. The participant would first hear a “ding” in their headset alerting them that a probe question was ready for response. From this ding they were instructed to only accept the ready prompt if they felt their workload was at an appropriate level to answer the question, and to otherwise not accept the question. Participants had 1-minute to respond to the ready prompt otherwise it would time out. Workload was measured by SPAM ready latency, the time to accept the ready prompt. Once the participant accepts the ready prompt, a question with multiple choice or true false answers is presented. Participants had 1-minute to answer the question. Participants were instructed during the briefing to answer the questions as quickly and accurately as possible within 1-minute.

SPAM probe latency (time to answer probe question) and probe accuracy (accuracy of probe response) were also recorded based on the response.

Measures

Subjective Trust

Through a literature review on measures of trust, three trust questionnaires were leveraged and distributed to participants on the first day of the internship (Week 1), at the beginning of the midterm exam (Week 9), and at the beginning of the final exam (Week 16). The three trust questionnaires included (see Appendix B-D for complete questionnaires): The CPRS (I. L. Singh et al., 1993), the ATM-CPRS (Verma et al., 2011), and M-HAT scale (Kunii, 2006).

Behavioral Trust

In addition to the three subjective measures, this study also collected an objective behavioral measure of controller trust during the simulation measured by the average number of near-miss aircraft moved. As a behavioral measure of trust in the study, the numbers of near misses were calculated using a Visual Basic program, which records the movements of any pre-programmed near misses per participant and scenario. A near miss is defined as an aircraft separation of 6-10 nm laterally. See Figure 1 for a visual representation of a near-miss aircraft with the nose of the aircraft outside of the 5 nm yellow ring.

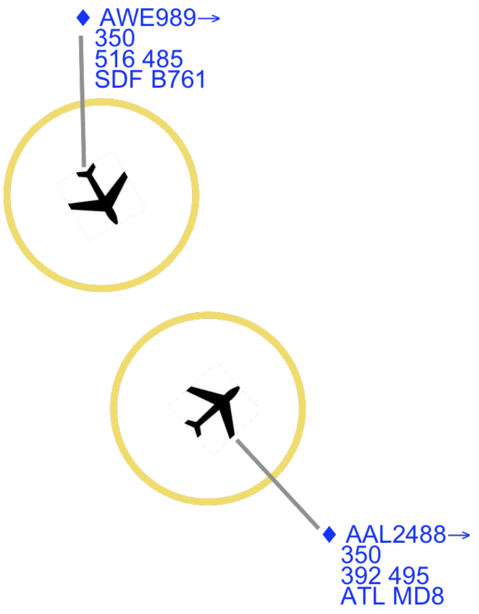

FIGURE 1. Near miss outside of 5 nm range ring.

Procedures

Training Procedure

The 16-week radar simulation internship in CHAAT took place every week on Saturday from 8:00 - 6:00 p.m. The lab portion lasted 3.25 hours while the lecture portion lasted 1.5 hours. A retired, radar-certified air traffic controller taught the internship lab and class. The lab was split into a morning and afternoon time to allow optimal time for each student with the ATCo tools in the lab. Students had the freedom to attend either morning or afternoon lab class. All students attended the classroom lecture at one time. On the first day of the internship, all participants signed informed consent forms and filled out preliminary questionnaires consisting of the three trust questionnaires, one personality questionnaire, and a questionnaire related to their demographics and previous ATCo experiences. During the first eight weeks of the internship, students were introduced to MACS and taught basic ATM techniques such as altitude, speed, vector, structure, and phraseology. Students also learned how to manage NextGen equipped and unequipped aircraft equally. Additionally, the instructor introduced the students to NextGen tools such as conflict detection, and assured them of the 100% reliability of such tools. The conflict detection tool (see Figure 2) automatically alerts controllers of aircraft on conflicting trajectories, at least eight minutes before a LOS.

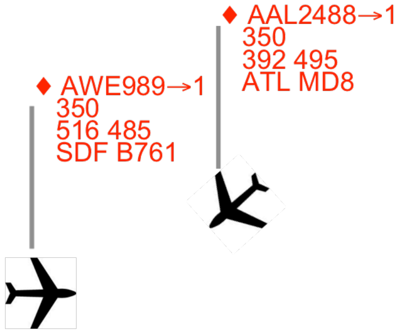

FIGURE 2. Conflict detection tool. The automated tool alerts potential conflicts by flashing the data blocks red on both aircraft involved in the conflict.

Experimental Procedure

The experimental testing portion of the internship took place at Week 9 for the midterm exam and Week 16 for the final exam. Testing procedures were the same for both exams. At the start of the testing session participants were provided with the three trust questionnaires. After, they were briefed for thirty minutes on the general purpose of the study, the MACS tools, scenarios, probe questions, and how to fill out post-scenario questionnaires. Following the briefing, students were taken into the testing cubicles and instructed to initiate a voice check with the pilots. Next, they participated in a ten-minute training scenario to warm-up their ATCo skills and practice using the probe touch screen. Following the training session, participants were given post-scenario questionnaires to fill out. Next the participants were run through four 42-minute mixed-equipage experimental scenarios. The order of the scenarios and probe questions were counterbalanced across participants for both the midterm and the final exam. Immediately after each trial, participants were given the post-scenario questionnaires and a ten-minute break. At the end of the final exam, the principal investigators of the lab debriefed participants to gather information on their experiences during the internship and exam sessions.

Scenarios

Training Scenarios

The training scenarios used throughout the lab internship were the same across participants. Depending on the participant's level of expertise with air traffic control, the instructor would select some scenarios with more (difficult) or less (easy) aircraft. Over the first eight weeks of the internship, the participants gained more experience with manual tools using 25% equipped scenarios. Based on previous studies, learning manual tools first allows the participant to have more efficient ATM skills (Kiken, Strybel, Vu & Battiste, 2012). Following the midterm exam at Week 9, the participants were trained with 50% equipped scenarios for the remainder of the internship until the final exam at Week 16.

Experimental Scenarios

The experimental scenarios were developed as 50% mixed-equipage, meaning half the aircraft (AC) were equipped with NextGen tools. Additionally, the scenarios varied in traffic density with two high-density traffic scenarios with 14-16 AC in the sector at a time and two low-density traffic scenarios with 11-13 AC in the sector. A subject matter expert ensured the equivalent difficulties between the two high-density and low-density scenarios and scenario presentation was also counterbalanced. Within each scenario, six conflicts and three near misses were designed into the traffic flows. The conflicts were between combinations of pairs of equipped aircraft, unequipped aircraft, and equipped and unequipped aircraft. For the near misses, all of the aircraft were NextGen equipped. The three near misses occurred around minute 10, minute 20, and minute 30 in each scenario.

CHAPTER 3

RESULTS

The following table (See Table 6) represents the average trust scores from the three trust scales administered on the first day of the internship, at the midterm exam (Week 9), and at the final exam (Week 16).

TABLE 6. Trust Scale Means and Standard Deviations

Internship Week	CPRS	ATM-CPRS	M-HAT
Initial (Week 1)	3.29 (.32)	3.48 (.31)	3.85 (.72)
Midterm (Week 9)	3.49 (.33)	3.79 (.44)	4.28 (.26)
Final (Week 16)	3.47 (.41)	3.66 (.42)	4.56 (.44)

Note. Standard deviations are in parentheses.

Sensitivity of Subjective Trust Measures

A 3 x 3 within-subjects analysis of variance (ANOVA) was conducted to evaluate whether there were differences in sensitivity between trust ratings measured over the course of the internship. The two independent variables in this study were Trust Scales (CPRS, ATM-CPRS, M-HAT) and Internship Week (1, 9, 16). The dependent variable was the trust score. For all analyses, we used a more liberal alpha level, .10, due to the small sample size. Furthermore, if any analyses had Mauchly's Test of Sphericity violations, Huynh-Feldt corrections were used. There was a significant main effect of the three trust scales administered, *F*(2, 22) = 28.87, *p* < .001. There was also a significant main effect of internship week, on the level of trust, *F*(2, 22) = 6.86, *p* = .005. This suggests that with training, the level of student ATCo trust increased. These main effects were modified by a significant interaction between Trust Scale and Internship Week, *F*(2.51, 27.63) = 2.65, *p* = .078 (see Figure 3).

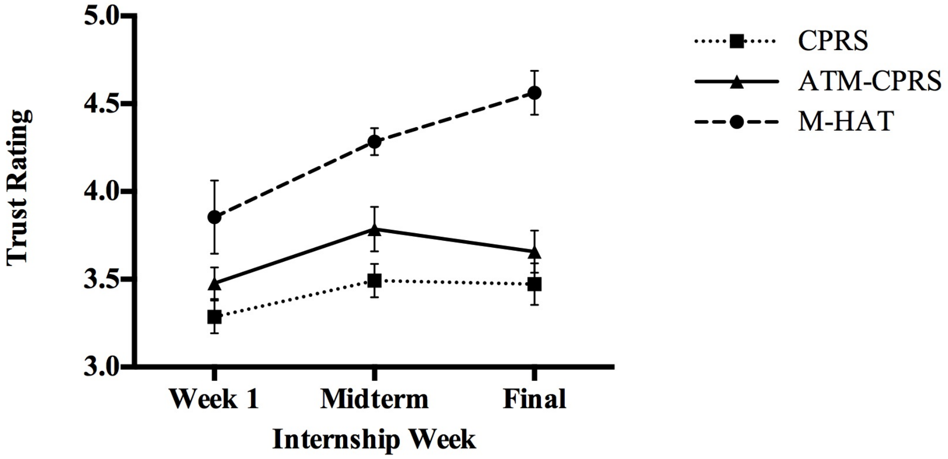

FIGURE 3. Significant interaction between Internship Week and Trust Scale rating. Error bars represent ±1 SEM. Trust scale values significantly increased over the 16-week internship.

Since the M-HAT is a 7-point Likert scale, and the CPRS and ATM-CPRS are 5-point Likert scales, the analyses were re-ran on the z-score transformations. There was only a main effect of Internship Week, *F*(2, 22) = 6.40, *p* = .006. The main effect of Trust Scale and the interaction between Trust Scale and Internship Week were not significant, *p*s > .20. However, the interaction pattern observed of the trust scores over the internship weeks was the same for all three trust scales. The M-HAT trust scores increased over the internship while the CPRS and ATM-CPRS only increased from Week 1 to midterm and decreased from midterm to final.

Trust Scale per Week in Internship

Simple effects tests were conducted to further break down the interaction of Trust Scale and Internship Week. The CPRS and ATM-CPRS trust scales did not significantly change over the internship, *p*s > .10. However, the M-HAT ratings significantly increased over the course of the internship, *F*(1.30, 14.29) = 6.16, *p* = .020. Bonferroni tests indicated participants' trust ratings on the M-HAT on the first day of the internship (*M* = 3.85, *SEM* = .21) were significantly lower than at the final exam (*M* = 4.56, *SEM* = .13), *p* = .063. Participants trust ratings on the M-HAT at the midterm exam (*M* = 4.28, *SEM* = .08) were also significantly lower than at the final exam (*M* = 4.56, *SEM* = .13), *p* = .089. From the first day of the internship to the midterm exam, there was no significant differences between trust ratings on the M-HAT, *p* > .10.

Pearson correlation analyses were conducted to determine the relationship between trust scale ratings over the course of the internship (see Table 7). A significant positive correlation was seen between CPRS at Week 1 (*M* = 3.29, *SD* = .32) and the midterm exam (*M* = 3.49, *SD* = .33), *r* = .516, *p* = .086. There was also a significant positive correlation between the CPRS at the midterm exam (*M* = 3.49, *SD* = .33) and the final exam (*M* = 3.47, *SD* = .41), *r* = .578, *p* = .049, indicating that the CPRS can reliably measure a general propensity to trust. For ATM-CPRS, there was a significant positive correlation between trust scores at the midterm exam (*M* = 3.79, *SD* = .43) and the final exam (*M* = 3.66, *SD* = .41), *r* = .386, *p* = .063. A similar trend was seen with the M-HAT scale. There was a significant positive correlation between the midterm (*M* = 4.28, *SD* = .26) trust scores and the final trust scores (*M* = 4.56, *SD* = .43), *r* = .484, *p* = .017.

TABLE 7. Trust Scales and Internship Week Significant Correlation Matrix

Measure	CPRS (*Midterm)*	CPRS (*Final)*	ATM-CPRS (*Final)*	M-HAT (*Final)*
CPRS (*Week 1*)	.516 *			
CPRS (*Midterm*)		.578 **		
ATM-CPRS (*Midterm*)			.386 *	
M-HAT (*Midterm)*				.484 **

Note. *** *p* < .001, ** *p* < .05, * *p* < .10.

Sensitivity of a Behavioral Measure of Trust and its Relationship to Subjective Trust

To determine if trust in automation affected the number of near-miss aircraft moved, participants were divided into high and low-trust groups based on the M-HAT scores, since it was the most sensitive to changes in trust over the course of the internship (see Table 8). The ratings averaged across the internship were used to divide the participants into a high (*M* = 4.54, *SEM* = .09) and low-trust (*M* = 4.02, *SEM* = .05) group using a median split (*Mdn* = 4.14) for further analyses.

A 2 (*Trust Group:* High, Low) x 2 (*Test Session:* Midterm, Final) x 2 (*Traffic Density:* High, Low Traffic) mixed-design ANOVA was conducted to assess whether the high and low-trust groups differed in the average number of near-miss aircraft moved at the midterm and the final with high and low-density traffic.

TABLE 8. Mean M-HAT Trust Scores for Participants in the High-Trust and Low-Trust Groups based on a Median Split

Trust Group	
High Trust	Low Trust
4.39 (.10)	3.81 (.75)
4.39 (.47)	3.86 (.29)
4.42 (.30)	4.00 (1.62)
4.69 (.25)	4.08 (.08)
4.81 (.25)	4.08 (.30)
	4.14 (.17)
	4.14 (.63)

Note. Median split based on average M-HAT trust scores over 16-week internship (*Mdn* = 4.14, *N* = 12). Standard deviations are in parentheses.

The dependent variable in this analysis was the average number of near-miss aircraft moved (maximum of three near misses per scenario). All main effects and interactions on the average number of near-miss aircraft moved were non-significant (*p*s > .20). At the midterm, students moved on average 1.00 (*SD* = .77) near-miss aircraft in high-density scenarios, and 1.29 (*SD* = 1.08) near-miss aircraft in low-density scenarios. At the final, students moved on average .96 (*SD* = .86) near-miss aircraft in high-density scenarios, and 1.04 (*SD* = 1.05) near-miss aircraft in low-density scenarios.

Pearson correlation analyses were also conducted to determine if there was a relationship between Trust Scale ratings and the average number of near-miss aircraft moved at the midterm and final Test Sessions. At the midterm exam, there were no significant relationships between Trust Scales and average number of near-miss aircraft moved, as shown in Table 9. At the final exam, there were significant negative correlations between the ATM-CPRS (*M* = 3.66, *SD* = .41) and the M-HAT (*M* = 4.56, *SD* = .43) on the average number of near-miss aircraft moved, *r* = -0.444, *p* = .030; *r* = -

0.344, *p* = .100, respectively (see Table 9). Additionally, by the final exam a significant positive correlation was found between the ATM-CPRS and M-HAT, *r* = .458, *p* = .025, (see Table 9).

TABLE 9. Trust Scales and Near Miss Correlation Matrix

Midterm Exam (Week 9)			
Measure	1. CPRS	2. ATM-CPRS	3. M-HAT
1. CPRS			
2. ATM-CPRS	0.367		
3. M-HAT	0.263	0.060	
Midterm Near Miss Average	0.206	0.054	0.094
Final Exam (Week 16)			
Measure	1. CPRS	2. ATM-CPRS	3. M-HAT
1. CPRS			
2. ATM-CPRS	0.577 ***		
3. M-HAT	0.081	0.458 **	
Final Near Miss Average	- 0.090	- 0.444 **	- 0.344 *

Note. *** *p* < .001, ** *p* < .05, * *p* < .10.

Trust in Automation and Situation Awareness

A 2 (*Trust Group:* High, Low) x 2 (*Test Session:* Midterm, Final) x 2 (*Traffic Density:* High, Low Traffic) mixed-design ANOVA was also conducted to look at the high and low trusting participants' SA levels at the midterm and final, with high and low-density traffic. Again, Trust Group was used as a between-subjects factor, and Test Session and Traffic Density were within-subjects factors. The dependent variables were SPAM probe accuracy and probe latency. All effects of the factors on SPAM probe latency were non-significant (*p*s > .30). For probe accuracy, a significant main effect of Test Session was obtained, *F*(1, 9) = 16.69, *p* = .003. SPAM probe accuracy at the final exam (*M* = .77, *SEM* = .17) was higher than at the midterm exam (*M* = .64, *SEM* = .03). There was also a significant interaction between the Traffic Density and Trust Group, *F*(1, 9) = 7.25, *p* = .025, and a significant three-way interaction between Test Session, Traffic Density, and Trust Group, *F*(1, 9) = 3.79, *p* = .083 (see Figure 4).

Simple effects analyses revealed no significant interaction between high and low-density scenarios for each Trust Group at the midterm exam. For the final exam, the interaction was significant, *F*(2, 22) = 12.33, *p* = .006. The low-trust participants had higher probe accuracy for the high-density scenarios (*M* = .86, *SEM* = .02) compared with the low-density scenarios (*M* = .71, *SEM* = .04). A reverse effect was seen for the high-trust participants who had higher probe accuracy for low-density scenarios (*M* = .81, *SEM* = .05) than the high-density scenarios (*M* = .73, *SEM* = .02). Simple effect analyses were also conducted on the simple effect of trust group at each level of traffic density. For high-density traffic, there was a significant effect of trust, *F*(1, 10) = 18.17, *p* = .001. The high-trust participants had significantly lower probe accuracy during high-density scenarios compared with the low-trust group. For low-density traffic, there was no significant effect of trust group, *p* > .10.

Another mixed-design ANOVA was carried out to examine if there was an effect of the number of SPAM probes that were not answered or “timed out” at the final exam. A 2 (*Trust Group:* High, Low) x 2 (*Traffic Density:* High, Low Traffic) mixed-design ANOVA was conducted. Trust Group was used as a between-subjects factor and Traffic Density was the within-subjects factors. The dependent variable was the number of unanswered probe questions. There was a main effect of Traffic Density, *F*(1, 10) = 11.43, *p* = .007.

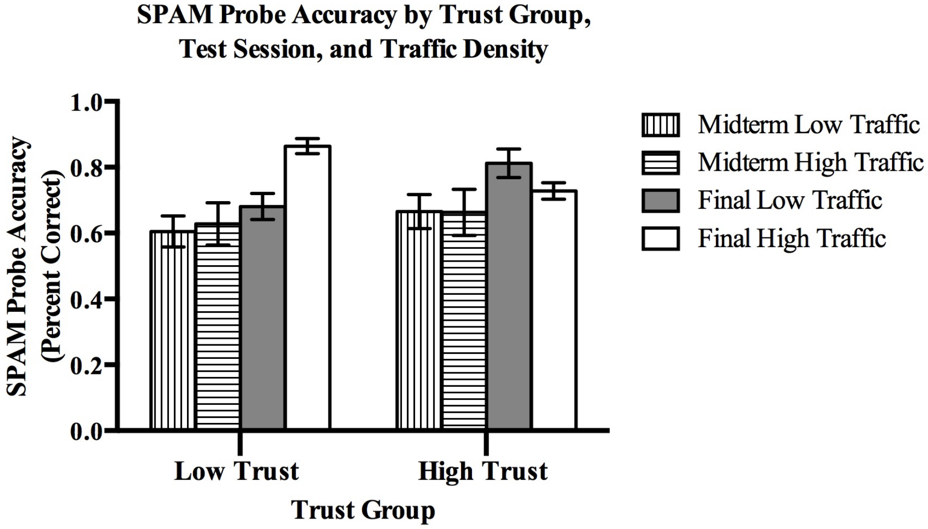

FIGURE 4. Significant three-way interaction between Test Session, Traffic Density, and Trust Group. Error bars represent ±1 SEM. The high-trust group had higher SA in low-density traffic and lower SA in high-density traffic at the final exam.

Significantly more probes were unanswered during high-density traffic scenarios (*M* = 3.71, *SEM* = 1.01) compared with low-density traffic scenarios (*M* = 1.31, *SEM* = .53). The interaction between Trust Group and Traffic Density was non-significant, *p* > .10.

Pearson correlation analyses were conducted to determine the relationship between Trust Scale ratings and SA measured through SPAM probe accuracy. At the midterm exam, there was a significant positive correlation only between ATM-CPRS (*M* = 3.79, *SD* = .43) and SPAM probe accuracy (*M* = .63, *SD* = .13), *r* = .446, *p* = .033. There were no significant correlations between Trust Scales and SPAM probe accuracy at the final exam. Thus, the relationship between trust in automation scales and SA was at best minimal.

Effect of Trust in Automation Levels on Workload

To determine whether workload may have played a role in the results between Trust Scales and SPAM probe accuracy, analyses of two workload measures was conducted. A 2 (*Trust Group:* High, Low) x 2 (*Test Session:* Midterm, Final) x 2 (*Traffic Density:* High, Low Traffic) mixed-design ANOVA was used to assess whether the high and low-trust groups differed in workload levels measured by SPAM ready latency and NASA-TLX. All effects on SPAM ready latency were non-significant. For NASA-TLX workload, a significant main effect of Traffic Density was found, *F*(1, 10) = 61.84, *p* < .001. As expected, participants reported lower workload (*M* = 45.23, *SEM* = 2.21) for low-density scenarios compared with high-density scenarios (*M* = 64.77, *SEM* = 2.76). There was also a significant main effect of Trust Group on reported workload for the NASA-TLX, *F*(1, 10) = 4.12, *p* = .070. These main effects were qualified by a significant interaction between Traffic Density and Trust Group, *F*(1, 10) = 5.17, *p* = .046. Overall, the low-trust group reported significantly lower workload during low-density traffic scenarios (*M* = 38.00, *SEM* = 2.85) compared to the high-trust group in low-density traffic scenarios (*M* = 52.46, *SEM* = 3.38) *F*(1, 10) = 10.70, *p* = .008 (see Figure 5).

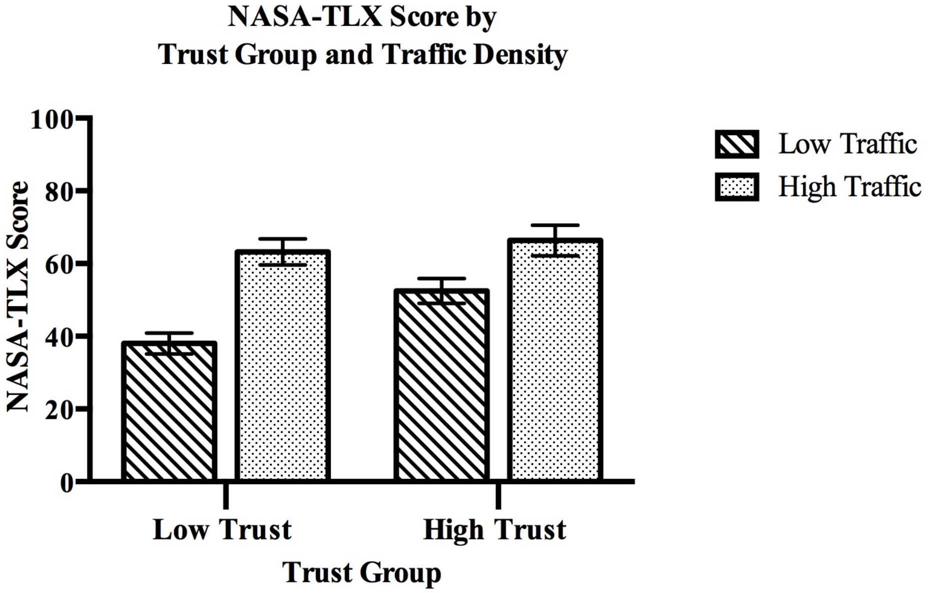

FIGURE 5. Significant interaction between Traffic Density and Trust Group. Error bars represent ±1 SEM. The low-trust group had lower workload compared to the high-trust group in low-density traffic.

CHAPTER 4

DISCUSSION

The purpose of the present study was to investigate participant's trust levels and the relationship between subjective trust, trust behavior, and SA over the course of a 16-week internship. The following three questions were explored in this study: (1) Are subjective trust scales sensitive to changes in trust brought about by training during a 16-week internship? (2) Are subjective trust measures related to a behavioral measure of trust based on the number of near-miss aircraft moved? And (3) what is the effect of ATCo trust levels on SA? Additionally, workload levels were observed to further investigate differences between SA levels in high and low-density traffic. The following three subjective trust scales were administered at three points of the internship (Week 1, Midterm, Final): CPRS, ATM-CPRS, and M-HAT. Over the course of the internship, students were trained with an equal mix of NextGen-equipped aircraft and current-day unequipped aircraft. For the experiment, two test scenarios consisting of low-density traffic and two test scenarios consisting of high-density traffic were used. The average number of near-miss aircraft moved was used to measure trust behavior. Near misses were defined as two aircraft coming close to losing separation but remaining 6-10 nm apart. There were three near misses per scenario at the midterm and final exams. SA at the midterm and final exams was measured with SPAM probe latency and accuracy and workload was measured with SPAM ready latency and NASA-TLX.

Sensitivity of Subjective Trust Scales

The results indicated that with training, ATCo trust levels in automation could be increased. Through the 16-week internship, participants reported higher trust levels of the subjective trust scales at the final relative to the beginning of training. Further analyses showed that M-HAT was the only trust scale to significantly increase over the internship. Although not significant, z-score analyses showed the same pattern. The M-HAT trust scale increased through the internship while the CPRS and ATM-CPRS only increased from the first day to the midterm exam and decreased from the midterm to the final exam. Therefore, the M-HAT scale may be the most sensitive to changes in trust throughout the 16-week internship. The M-HAT is a 7-point Likert trust scale that was made specific to ATM. Theses results suggested using trust questionnaires that focused on an overall system (M-HAT) may be more sensitive to capturing changes in trust over time than using trust questionnaires focused on specific automated tools (ATM-CPRS). It is possible that by administering a more specific scale, the changes in attitudes of trust are more difficult to detect. Of course, it is also possible that trust in automation was unchanged over the course of the internship, meaning that the CPRS scales were more valid measures of trust. It is difficult to determine which outcome is more valid, which is why I attempted to obtain a behavioral measure of trust. From the first day to the final exam, the M-HAT trust ratings significantly increased and students reported more trust. Therefore, the M-HAT scale was used to divide the participants into high and low-trust groups based on a median split.

The relationships between trust scales and internship week also suggested reliability of the scales based on the consistent trust ratings. The CPRS had significant positive correlations between Week 1, the midterm exam, and the final exam. The ATM-CPRS and M-HAT had significant positive correlations between the midterm and final exam. The CPRS results are not surprising since this scale measures a general propensity to trust in automation, and a person's propensity to trust automation should change more slowly. The ATM-CPRS and the M-HAT are more specific trust scales so it makes sense there was no relationship between the scales at Week 1 since the participants had no formal training with NextGen tools.

Relationships Between Trust Scales and Behavioral Trust Measure

High-trust and low-trust groups were formed using the M-HAT scores. Participants were divided into high and low-trust groups using a median split. The average number of near-miss aircraft moved was not significantly affected by Trust Group, Test Session, or Traffic Density. Although not significant, all participants, especially those in the high-trust group, tended to move more near-miss aircraft in low-density scenarios compared with high-density scenarios. On average, participants moved about one near miss out of three per scenario. Additionally, the differences between the average trust level in the low-trust group and high-trust group was very small. Previously, Higham (2013) found significant differences (p = .08) for the number of near-miss aircraft moved. Specifically, participants who had received trust training moved fewer near-miss aircraft at the final than at the midterm exam. Participants who did not receive trust training moved more near-miss aircraft at the final than at the midterm exam during 50% equipage scenarios. The differences between the behavioral trust measure results in these two studies can be attributed to the modification in the simulation design. For example, the current study had manipulated traffic density with two low-density and two high-density scenarios, which in turn affected workload levels of the controllers. Additionally, the current study did not manipulate trust training by providing trust feedback to the student controllers. There were no observed differences for the behavioral measure of trust in the present study. Three near misses per scenario were included. This was an increase from the previous study, which only used one near miss in the 50% equipage scenarios and two near misses in the 100% equipage scenarios (Higham, 2013). Therefore, it may not be necessary to increase the number of opportunities in the scenarios for participants to demonstrate their trust in automation as Higham (2013) suggested. These differences suggest that training to trust automation is more effective than training in general even when more opportunities for measuring trust behaviors are provided.

Correlation analyses, however, showed significant negative relationships between trust scale ratings on the ATM-CPRS and M-HAT and the average number of near-miss aircraft moved at the final exam. This indicates that as the trust scale rating increased, the number of near-miss aircraft moved decreased. Since this relationship was only seen at the final exam, it is possible some training on NextGen automated tools is required before trust in automation becomes important in ATCo performance. The CPRS and ATM-CPRS trust scales were also significantly and positively correlated. As the trust scores on the CPRS increased, trust scores on the ATM-CPRS also increased. However, only the ATM- CPRS was correlated with the number of near-miss aircraft moved. This suggests that the NextGen concepts included in the ATM-CPRS trust scale, may be required to produce the negative correlation between ATM-CPRS and near-miss aircraft moved. In other words, the CPRS is too general of a trust in automation scale to be related to ATM performance.

Effect of Trust on Situation Awareness

For SA at the final exam during high-density scenarios, participants who scored higher in trust had lower SPAM probe accuracy. One may expect that high levels of trust in automation would lead to reduced SA, which previous literature has called automation-induced complacency (Parasuraman & Riley, 1997). The low-trust participants had better SA compared to the high-trust participants because they remained engaged with the traffic. On the other hand, the ATM-CPRS showed a significant positive correlation between trust score and SPAM probe accuracy suggesting that as trust increased, SA also increased. Similar to Kunii (2006), these findings showed the same relationship between trust and SA for student pilots. Higher trusting individuals were carefully trusting the automation and in turn not compromising their SA levels. To look at these results in more detail, workload was also considered.

Difference Between Trust and Situation Awareness with Workload

Workload, measured by SPAM ready latency and NASA-TLX, was also analyzed to determine the effects of trust in automation on workload. Based on NASA-TLX ratings, participants reported lower workload levels during low-density traffic compared to high-density traffic. This shows that the manipulation to the number of aircraft in the sector during low-density traffic (11-13 AC) and high-density traffic (14-16) scenarios was successful. There was a significant interaction between Traffic Density and Trust Group on NASA-TLX workload ratings. For low-density traffic, low-trust participants reported lower workload compared to the high-trust participants. This can be considered the opposite of what is typically expected. Operators tend to work harder when they have low trust in automation since they are not appropriately using the automation's capabilities (Parasuraman & Riley, 1997).

Similar to the findings from Higham (2013), overall workload decreased from the midterm to the final testing sessions, which implies that skills were developed for ATM and using the NextGen tools over the 16-weeks of the internship. Higham (2013) found no significant differences between workload levels in 50% mixed-equipage scenarios while the current study, which only used 50% mixed-equipage scenarios, found differences between workload during low-density traffic for high-trust controllers who reported significantly higher workload than low-trust controllers. For SA, similar findings were seen to Higham (2013) study where overall SA accuracy improved from midterm to the final exam.

Training and NextGen

Overall, these findings point towards an effect of training leading to significantly different behaviors at the final exam versus at the midterm exam. This is likely due to the limited experience of participants before the midterm exam at Week 9. By the midterm, ATCos were still learning to use NextGen tools and manage traffic; therefore trust was not as critical of an issue. The level of SA at the midterm exam demonstrates this, which was equal for both the low and high-trust groups in both the low and high-density traffic. By the final exam, the participants were well versed with using NextGen tools and managing traffic, and trust may have played a greater role. Regardless of Trust Group, the participants in high traffic situations had high workload (*M* = 64.77) compared with low-density traffic (*M* = 45.23). For low-density traffic, workload played a greater role, as there were significant differences between high and low-trust participants. These findings are specific to only the final exam. At the midterm these patterns were not seen.

Findings from Higham (2013) suggested there is a potential to train trust in automated NextGen tools. This was shown through the tendency for the Trust Training group to move fewer near misses in mixed-equipage scenarios compared to the No-Trust Training group. The current study did not employ any specific trust training. Alternatively, the internship consisted of general training for ATM and NextGen using mixed-equipage scenarios. During training, the instructor only reminded ATCos of the 100% reliability of the NextGen conflict detection tools; no other feedback was given. Based on the findings of the present study, it may be likely that providing more explicit training to trust automated tools can have a larger influence on trust behavior.

Study Limitations

Several factors contributed to the limitations of this study. First, there were only a small number of participants ($N = 12$) for the 16-week internship. It was difficult to have a larger sample size as the participants were recruited from a specific population of ATCo students. A second possible factor in the limitations of this study was there were minor differences between the low and high-trust group ratings from the median split. The average trust scores in the high and low-trust groups only had a difference of .52, therefore it was difficult to find differences between the groups in subsequent analyses looking at behavioral differences in trust. This could have led to the result of no difference between the low and high-trust groups and the average number of near-miss aircraft moved. The third limitation was a possible floor effect for the dependent variable used as a behavioral measure of trust, the number of near-miss aircraft moved. On average, the numbers of near misses were very close to one. These three factors resulted in low statistical power and a significant difference was difficult to obtain even when using a more liberal alpha level of 1.0.

Future Recommendations for Design and Training

In order to address the limitations of the current study, it is recommended to increase the number of participants in the study to further increase power. This can lead to greater differences between the low and high-trust groups. Although the number of near-miss aircraft moved was increased from a previous study using the same behavioral trust measure (Higham, 2013), there was still little variability between the two trust groups. It cannot be recommended to further increase the number of near-miss aircraft to more than three per scenario, which was what was currently used, because controllers tend to change their strategies or become suspicious if they notice a similar reoccurring situation during a simulation. It may irritate a controller to experience too many near misses because that is not representative of a typical sector. Additionally, the current study's increase to three near misses per scenario did not have a greater effect on the number of near-miss aircraft moved as previously expected (Higham, 2013).

It is also recommended that future training studies explore the benefits of trust training. Based on the findings, it can be suggested that trust training may have a greater impact on using automated tools appropriately. This was evidenced by the difference in the findings for near-miss aircraft moved between the present study and Higham (2013). As long as an automated tool is reliable, it is valuable the operator is accurately calibrated to trust the automation. This may possibly be achieved with trust training, but future studies should further investigate this idea thoroughly to verify its positive impact.

APPENDICES

APPENDIX A

INFORMED CONSENT FORM

CONSENT TO PARTICIPATE IN RESEARCH

Using Situation Awareness Probes to Predict Air Traffic Controller Performance

You are being asked to participate in a research study conducted by Thomas Strybel, from the Department of Psychology at California State University, Long Beach (CSULB) as part of an ongoing development of new situation awareness metrics that are better suited for use in future procedures under considerations for the National Airspace System (NAS). You were selected as a possible participant in this study because you are a U.S. Citizen, over 18 years old, have radar simulation experience, and have been an air traffic controller.

PURPOSE OF THE STUDY

This project is a simulation effort of The Center for Human Factors in Advanced Aeronautics Technologies (CHAAT) at CSULB. We are examining the effectiveness of different situation awareness tests when applied to air traffic controllers managing aircraft that are flying in the NAS. Situation awareness refers to the ability to "have the picture," and be able to anticipate possible situations in one's sector. We are measuring situation awareness so that new technologies can be evaluated for their impact on air traffic controller situation awareness.

PROCEDURES

If you agree to participate in this study, you will attend a training briefing where you will learn to use a desktop air traffic control simulator (MACS) and become familiar with sectors in the national airspace. Then, you will participate in 4 simulation scenarios over 1 day at CSULB, Long Beach, California. Each scenario will be approximately 40-minutes. After each scenario, you will be asked to fill out questionnaires about your experience in the scenario. Rest breaks will be provided between scenarios.

POTENTIAL RISKS AND DISCOMFORTS

There are no foreseeable risks involved in participating in this experiment. You can leave the test room at any time without penalty. You may also discontinue your participation in the experiment at any time and will still be compensated for the amount of time spent in the experiment.

POTENTIAL BENEFITS TO SUBJECTS AND/OR TO SOCIETY

The results of this experiment will contribute to the development of situation awareness metrics better suited to the types of air traffic management scenarios under considerations for future National Airspace Systems.

PAYMENT FOR PARTICIPATION

You will receive $10 per hour for each hour of participation in the experiment at CSULB. Should you decide to withdraw from the experiment before completion, compensation will be commensurate with your participation.

CONFIDENTIALITY

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. The results from this experiment will not be associated with you in any way. You will have to provide your social security number for payment. This information will be stored in the office of the Psychology Department Office Administrator until we have verified that your check was received.

PARTICIPATION AND WITHDRAWL

You can choose whether to be in this study or not. If you agree to be in this study, you may withdraw at any time without consequences of any kind. Participation or non-participation will not affect your status in the university. You will be paid for all sessions completed. The investigator however, may withdraw you from this research if circumstances arise which in the opinion of the researcher warrant doing so.

IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact the Principal Investigator: Thomas Strybel at CSULB (562-985-5035; tstrybel@csulb.edu).

RIGHTS OF RESEARCH SUBJECTS

You may withdraw your consent at any time and discontinue participation without penalty. You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you have questions regarding your rights as a research participant, contact the Office of University Research, CSULB, 1250 Bellflower Blvd, Long Beach, CA 90840; Telephone: (562) 985-5314 or email to research@csulb.edu.

SIGNATURE OF RESEARCH PARTICIPANT or LEGAL REPRESENTATION

I am at least 18 years old and I understand the procedures and conditions of my participation described above. My questions have been answered to my satisfaction and I agree to participate in this study. I have been given a copy of this form.

____
Name of Subject

____ ____________________
Signature of Subject Date

APPENDIX B

COMPLACENCY-POTENTIAL RATING SCALE (CPRS)

Participant______________ **Date______________**

Complacency-Potential Rating Scale (CPRS; I. L. Singh et al., 1993)

Please **circle the number** at the point which best describes your feeling or your impression.

Confidence:

1. I think that automated devices used in medicine, such as CT scans and ultrasound, provide very reliable medical diagnosis.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

2. Automated devices in medicine save time and money in the diagnosis and treatment of disease.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

3. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer than manual surgery.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

4. Automated systems used in modern aircraft, such as the automatic landing system, have made air journeys safer.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

Reliance:

1. ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

2. Automated devices used in aviation and banking have made work easier for both employees and customers.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree **Strongly Agree**

3. Even though the automatic cruise control in my car is set at a speed below the speed limit, when I pass a police radar speed trap I worry that the automatic control may not be working properly.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

Trust:

1. Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

2. I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

3. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

Safety:

1. I feel safer depositing my money at an ATM than with a human teller.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

2. I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my DVR rather than manual recording.

1-----------------------2-------------------3---------------------4--------------------5
Strongly Disagree **Strongly Agree**

APPENDIX C

COMPLACENCY POTENTIAL RATIING SCALE (ATM-CPRS)

Participant______________ **Date**______________

Complacency Potential Rating Scale (ATM-CPRS; Verma et al., 2011)

Please **circle the number** at the point which best describes your feeling or your impression.

Confidence:

1. I think the **conflict alerting** function is reliable.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

2. I think that **conflict alerting** reduced my effort and workload as a controller.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

3. I think the **conflict probe** function available with the trial planner is reliable.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

4. I think that the **conflict probe** reduced my effort and workload as a controller.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

5. I think that the **datalink** capability is reliable.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

6. I think that the use of **datalink** reduced my effort and workload as a controller.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

7. I think the automated **conflict alerting system** is more reliable than human controllers manually detecting aircraft conflicts.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

8. I think the automated **conflict probe system** is more reliable than human controllers manually detecting aircraft LOS.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

9. I think the **datalink system** results in more reliable communication with aircraft than voice over the radio frequency.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

10. I think that use of **conflict alerting** results in a safer system than the human controller monitoring for aircraft conflicts manually.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

11. I think that use of **conflict probes** results in a safer system than the human controller determining the LOS area.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

12. I think that use of **datalink** procedures results in a safer system than voice procedures.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

Reliance:

13. **Conflict alerting** provides many safeguards against errors (e.g., miscalculations of the distance between aircraft).

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

14. **Conflict probes** provide many safeguards against errors (e.g., miscalculations of the area of LOS).

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

15. **Datalink** provides many safeguards against errors (e.g., clearances being executed by the wrong flight deck.)

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

16. I think that use of **conflict alerting** made the air traffic control task easier.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

17. I think that use of the **conflict probe** made the air traffic control task easier.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

18. I think that use of **datalink** made the air traffic control task easier.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

19. I have concerns that the **conflict alerting system** may not work properly.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

20. I have concerns that the **conflict probe system** may not work properly.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

21. I have concerns that the **datalink system** may not work properly.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

Trust:

22. When monitoring traffic, I think that automation, in terms of the **conflict alerting system**, is more reliable than my own monitoring.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

23. I think using the **conflict alerting system** as a method for conflict detection is more likely to be correct (e.g., predict actual aircraft pairs in conflict) than manually monitoring.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

24. I think that the **conflict alerting system** will make air traffic control safer.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

25. When resolving conflicts, I think that automation, in terms of the **conflict probe system**, is more reliable than my own calculations.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

26. I think using the **conflict probe system** as a method for conflict resolution is more likely to be correct (e.g., predict area of LOS better) than manually calculations.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

27. I think that the **conflict probe system** will make air traffic control safer.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

28. When communicating, I think that using **datalink** is more reliable than using voice communication.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

29. I think using the **datalink system** as a method for communication is more likely to be correct (e.g., results in aircraft executing intended commands) than voice communication.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

30. I think that the **datalink system** will make air traffic control safer.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

Safety:

31. Using **conflict alerting** makes me feel safer about detecting conflicts than doing my own monitoring.

1------------------------2--------------------3----------------------4----------------------5
Strongly Disagree Strongly Agree

32. I would choose **conflict alerting** over manual monitoring to ensure conflict detection between aircraft.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

33. Using **conflict probes** makes me feel safer about determining the LOS area than doing my own calculations.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

34. I would choose to use **conflict probes** over manual calculations to ensure no LOS between aircraft.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

35. Using **datalink** makes me feel safer about aircraft executing my clearances than voice communication.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

36. I would choose **datalink** over voice to ensure accurate communication with aircraft.

1------------------------2--------------------3----------------------4---------------------5
Strongly Disagree Strongly Agree

APPENDIX D

MODIFIED HUMAN-AUTOMATION TRUST SCALE (M-HAT)

Participant_____________ **Date__________**

Modified Human-Automation Trust Scale (M-HAT; Kunii, 2006)

Please **circle the number** at the point which best described your feeling or your impression. For the purposes of this questionnaire, “system” refers to the air traffic controller systems, for example, the radar scope, trial planner, conflict probe, conflict alerting, etc., and what they tell you as an air traffic controller. PLEASE ANSWER AS CANDIDLY AS YOU CAN.

1. The system can be deceptive

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

2. The system sometimes behaves in unpredictable manner

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

3. I am often suspicious of the system’s intent, action, or outputs

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

4. I am sometimes unsure of the system

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

5. The system’s action can have a harmful or injurious outcome

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

6. I am confident in the system

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

7. The system can provide security

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

8. The system has integrity

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

9. The system is dependable

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

10. The system is consistent

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

11. I can trust the system

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

12. I am familiar with the system

1-------------2-------------3-------------4-------------5-------------6-------------7

Not at all **Extremely**

Comments:

APPENDIX E

DEMOGRAPHICS QUESTIONNAIRE

Participant________ Lab________ Date__________

Demographic Questionnaire

1) What is your age? __________

2) What is your gender? (Please circle one)
Male Female

3) How many years have you been studying to be a controller?
_____________ years

4) Have you taken the AT-SAT?
Yes No

5) Please describe any experience you have in air traffic management such as an internship, training, or supervision (e.g. locations worked, duties, years at each location).

____
____
____
____
____

6) Please rate your radar experience. (Please circle one)

1-----------------2--------------3-------------4-------------5--------------6--------------7
No Somewhat Very
Experience Experienced Experienced

7) Please rate your experience with ZID airspace. (Please circle one)

1-----------------2--------------3-------------4-------------5--------------6--------------7
No Somewhat Very
Experience Experienced Experienced

8) Please rate your experience with the MACS software. (Please circle one)

1-----------------2--------------3-------------4-------------5--------------6--------------7
No Somewhat Very
Experience Experienced Experienced

9) Are you a licensed pilot? (Please circle one)
YES NO

If yes: please indicate your FAA certifications/ratings by placing an "X" next to all that are applicable.

_____Private
_____ATP
_____CFI

______Commercial
______Instrument
______CFII

_____Other
(please describe):____

10) Please list any other qualifications you think are relevant as a participant in this study.

____
____
____
____
____

APPENDIX F

NASA TASK LOAD INDEX (NASA-TLX)

Participant ID ______________ Day _____ Time _______

Trial Number ______________ Scenario Number _________

NASA TLX Workload Scale

RATING SCALE DEFINITIONS		
Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
PERFORMANCE	*Good/Poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Participant ID ______________ Day _____ Time _______

Trial Number ______________ Scenario Number _________

NASA TLX RESPONSE FORM

MENTAL DEMAND

Low **High**

PHYSICAL DEMAND

Low **High**

TEMPORAL DEMAND

Low **High**

PERFORMANCE

Good **Poor**

EFFORT

Low **High**

FRUSTRATION

Low **High**

APPENDIX G

POST-EXPERIMENT QUESTIONNAIRE

Participant ID ___________ Date ____________ Time _____________

Post Experiment Questions

For each information element below, please rate on a 1-10 scale, how critical the information element was for successfully managing traffic in the scenarios you worked.

1 = not at all relevant; 5 = relevant; 10 = critical

Information Item	Rating
Departure AC speed	
Departure AC altitude	
Departure AC heading	
Departure AC distance to APALO/PXV	
Departure AC exit altitude	
Departure AC equipage information	
SDF Arrival AC heading	
SDF Arrival AC airspeed	
SDF Arrival AC altitude	
SDF Arrival AC distance to airport	
SDF Arrival AC exit altitude	
SDF Arrival AC relative position to overflight traffic	
SDF Arrival AC equipage information	
SDF Arrival AC distance to ZARDA/PENTO	
Overflight AC call sign	
Overflight AC airspeed	
Overflight AC heading	
Overflight AC altitude	

Information Item	Rating
Overflight AC route to exit gate	
Overflight AC original altitude when exiting	
Overflight AC equipage information	
Relative distance between arriving aircraft	
Relative distance between arriving aircraft and overflight aircraft	
Relative distance between arriving aircraft and departure aircraft	
Relative distance between departure aircraft overflight aircraft	
Relative distance between overflight aircraft	
Relative distance between unequipped aircraft	
Relative distance between equipped aircraft	
Relative distance between equipped and unequipped aircraft	
Handoff frequency	
Altitude differences	
Speed differences	
AC headings	
Heading differences	

Please circle the number that best describes *how realistic the scenarios* were:

1 Extremely Unrealistic	2	3	4	5	6	7 Extremely Realistic

Comments:

Please circle the number that best describes *how realistic the ATC radar interface was:*

1 Extremely Unrealistic	2	3	4	5	6	7 Extremely Realistic

Comments:

Please circle the number that best describes *how realistically the simulation pilots responded to your communications and requests for flight plan changes and other traffic information:*

1 Extremely Unrealistic	2	3	4	5	6	7 Extremely Realistic

Comments:

Please circle the number that best describes *how well the training prepared you for the scenarios:*

1 Extremely Inadequate	2	3	4	5	6	7 Extremely Adequate

Comments:

Please circle the number that best describes *how interfering it was to answer questions when they appeared on your probe screen:*

1 Not at all interfering	2	3	4	5	6	7 Extremely interfering

Comments:

Please circle the number that best describes *how your workload was changed by having to respond to questions when they appeared on your probe screen:*

1 Significant *decrease* in workload	2	3	4 *No change* in workload	5	6	7 Significant *increase* in workload

Comments:

To what extent did the probe questions and your responses to the probe questions *change your awareness of traffic?*

1 No change	2	3	4 Some change	5	6	7 Significant change

Comments:

To what extent did the probe questions and your responses to them *change your strategies for managing traffic?*

1 No change	2	3	4 Some change	5	6	7 Significant change

Comments:

How helpful were the *conflicts alerts for detecting and resolving conflicts?*

1	2	3	4	5	6	7
Not at all helpful			Somewhat helpful			Extremely helpful

Comments:

Rate your preference for vertical resolutions when *using the trial planning tools.*

1	2	3	4	5	6	7
No preference			Some preference			Extreme preference

Comments:

Rate your preference for lateral resolutions when ***using the trial planning tools.***

1	2	3	4	5	6	7
No preference			Some preference			Extreme preference

Comments:

Rate your preference for vertical resolutions when you made ***manual resolutions.***

1	2	3	4	5	6	7
No preference			Some preference			Extreme preference

Comments:

Rate your preference for lateral resolutions when you made ***manual resolutions.***

1	2	3	4	5	6	7
No preference			Some preference			Extreme preference

Comments:

Is there anything about the experiment that we should have asked or that you would like to comment about?

************** ***Thank you for your participation!*** ******************

APPENDIX H

DEBRIEFING QUESTIONS

Sim 8 Final Exam Debriefing Questions

Probe

- Which lab were you in?
- How often did you rely on conflict detection and the conflict probe to help you manage traffic and resolve conflicts?
- What percentage of conflicts between datalink equipped aircraft did you try to resolve before the conflict alert?

Training

- How did the training you received in the lab and class prepare you for the testing scenarios for the midterm? For the final? What changed in your traffic management strategies between the midterm and final?
- Do you feel that your performance improved from the last time you were tested? Discuss how and why.
- Discuss which issues you faced when learning to manage traffic with Datalink tools.
- What tasks were specifically difficult to perform without datalink?
- How did you combine heading, altitude, and structured techniques into managing your traffic? What strategies did you use?
- Approximately how many class periods did it take you to feel comfortable with a mixed-equipage scenario.
- Do you have any other comments about your training?

Scenarios

- Were there any scenarios that you found to be more difficult and why?

Datalink

- When you were using the datalink or the trial planner, did you have a preference for making vertical or lateral resolution? Did you prefer to resolve conflicts with voice or datalink?
- For the mixed scenarios, did your strategies change as a result of having conflict alerts only for some aircraft?
- Approximately how many aircraft do you feel comfortable managing at one time without datalink?
- Approximately what proportion of aircraft needed to be equipped with datalink in order for you to feel comfortable while managing?
- Approximately what proportion of datalink aircraft with conflict detection would you feel comfortable managing at one time?
- What additional tools or information would have made your task easier, both while you were learning to manage traffic and once you were proficient at managing traffic?

Communications

- How often did you use "expedite" and under what situations?

REFERENCES

REFERENCES

Barber, B. (1983). *The logic and limits of trust* (Vol. 96). New Brunswick, NJ: Rutgers University Press.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution, 2*(4)*,* 265-279.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(3), 474-486.

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.). *A cognitive approach to situation awareness: Theory and application* (pp. 137-154). Aldersot, United Kingdom: Ashgate.

Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. *Handbook of applied cognition* (pp. 283-314). Chichester, England: John Wiley.

Eckel, C. C., & Wilson, R. K. (2004). Is trust a risky decision? *Journal of Economic Behavior & Organization*, *55*(4), 447-465.

Endsley, M. R. (1996). Automation and situation awareness. *Automation and human performance: Theory and applications*, (pp. 163-181), Mahwah, NJ: Lawrence Erlbaum.

Ergenli, A., Saglam, G., & Metin, S. (2007). Psychological empowerment and its relationship to trust in immediate managers. *Journal of Business Research*, *60*, 41-49.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *The International Journal of Aviation Psychology*, *23*(3), 245-266.

Hart, S. G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam, The Netherlands: Elsevier.

Higham, T. L. M. (2013). *Training trust in automation within a NextGen environment* (Master's thesis). California State University, Long Beach.

Hoff, K. A., & Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3), 407-434. doi:10.1177/0018720814547570

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

Joint Planning and Development Office. (2010). *Concept of operations for the next generation air transportation system version 3.2.* Retrieved from http://jpe.jpdo.gov/ee/docs/conops/NextGen_ConOps_v3_2.pdf

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine, 67*(6), 507-512.

Kiken, A., Rorie, R. C., Bacon, L. P., Billinghurst, S., Kraut, J. M., Strybel, T. Z., & Battiste, V. (2011). Effect of ATC training with NextGen tools and online situation awareness and workload probes on operator performance. In G. Salvendy & M. J. Smith (Eds.), *Human interface, Part II, Lecture notes in computer science,* (Vol. 6772, pp.483-492). Berlin, Germany: Springer Verlag.

Kiken, A., Strybel, T. Z., Vu, K. P. L., & Battiste, V. (2012, July). Effectiveness of training on near-term NextGen air traffic management performance. In *Proceedings of the Applied Human Factors and Ergonomics Society Annual Meeting* (pp. 311-320). San Francisco, CA: USA Publishing.

Kunii, Y. (2006). *Student pilot situational awareness: The effects of trust in technology* (Unpublished doctoral dissertation). Embry-Riddle Aeronautical University, Daytona Beach, FL.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50-80.

Löwe, B., Kroenke, K., Herzog, W., & Gräfe, K. (2004). Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the Patient

Health Questionnaire (PHQ-9). *Journal of Affective Disorders, 81*(1), 61-66.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8(*4), 277-301.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human–automation interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*, 194–210.

Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43*(4), 519-528.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5), 527-539.

Muir, B. M. (1988). Trust between humans and machines, and the design of decision aids. In E. Hollnagel, G. Mancini, & D. D. Woods (Eds.), *Cognitive engineering in complex dynamic worlds* (pp. 71-83). London, England: Academic.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3*(1), 1-23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 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.

Prevot, T. (2002). Exploring the many perspectives of distributed air traffic management: The multi-aircraft control system MACS. In S. Chatty, J. Hansman, & G. Boy (Eds.), *Proceedings of the HCI-Aero 2002* (pp. 149-154). Menlo Park, CA: AAAI Press.

Pruitt, D. G., & Rubin, Z., (1986). *Social conflict: Escalation, statement, and settlement*. New York, NY: Random House.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95-112.

Riley, V. (1989, October). A general model of mixed-initiative human-machine systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 33, No. 2, pp. 124-128). Santa Monica, CA: SAGE.

Rodgers, M. D., Mogford, R. H., & Strauch, B. (2000). Post hoc assessment of situation awareness in air traffic control incidents and major aircraft accidents. In M. R. Endsley, & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 73-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651-665.

Sheridan, T. B. (1975). Considerations in modeling the human supervisory controller. In *Proceedings of the International Federation of Automatic Control, 6th World Congress* (pp. 1-6). Laxenburg, Austria: International Federation of Automatic Control.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control.* Cambridge, MA: MIT Press.

Singh, A. L., Tiwari, T., & Singh, I. L. (2009). Effects of automation reliability and training on automation-induced complacency and perceived mental workload. *Journal of the Indian Academy of Applied Psychology, 35* (Special issue), 9-22.

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*(2), 111-122.

Verma, S., Kozon, T., Ballinger, D., Lozito, S., & Subramanian, S. (2011, October). Role of the controller in an integrated pilot—Controller study for parallel approaches. In *IEEE/AIAA 30th Digital Avionics Systems Conference (DASC)* (pp. 3B1-1). Seattle, WA: IEEE.