

San Jose State University
SJSU ScholarWorks

Master's Theses

Master's Theses and Graduate Research

Fall 2013

The Impact of Trajectory Prediction Uncertainty on Reliance Strategy and Trust Attitude in an Automated Air Traffic Management Environment.

Sarah Marie Hunt
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Hunt, Sarah Marie, "The Impact of Trajectory Prediction Uncertainty on Reliance Strategy and Trust Attitude in an Automated Air Traffic Management Environment." (2013). *Master's Theses*. 4391.

DOI: <https://doi.org/10.31979/etd.a9t8-bwjc>

https://scholarworks.sjsu.edu/etd_theses/4391

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

THE IMPACT OF TRAJECTORY PREDICTION UNCERTAINTY ON
RELIANCE STRATEGY AND TRUST ATTITUDE IN AN AUTOMATED AIR
TRAFFIC MANAGEMENT ENVIRONMENT.

A Thesis

Presented to

The Faculty of the Department of Human Factors and Ergonomics

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Sarah M. Hunt

December 2013

© 2013

Sarah M. Hunt

ALL RIGHTS RESERVED

The Designated Thesis Committee Approves the Thesis Titled

THE IMPACT OF TRAJECTORY PREDICTION UNCERTAINTY ON
RELIANCE STRATEGY AND TRUST ATTITUDE IN AN AUTOMATED AIR
TRAFFIC MANAGEMENT ENVIRONMENT.

by

Sarah M. Hunt

APPROVED FOR THE DEPARTMENT OF HUMAN FACTORS AND
ERGONOMICS

SAN JOSÉ STATE UNIVERSITY

December 2013

Dr. Kevin Jordan Department of Psychology

Dr. Sean Laraway Department of Psychology

Dr. Lynne Martin San José State Research Foundation

ABSTRACT

THE IMPACT OF TRAJECTORY PREDICTION UNCERTAINTY ON RELIANCE STRATEGY AND TRUST ATTITUDE IN AN AUTOMATED AIR TRAFFIC MANAGEMENT ENVIRONMENT.

by Sarah M. Hunt

Future air traffic environments have the potential to exceed human operator capabilities. In response, air traffic control systems are being modernized to provide automated tools to overcome current-day workload limits. Highly accurate aircraft trajectory predictions are a critical element of the automated tools envisioned as part of the evolution of today's air traffic management system in the United States, known as NextGen. However, automation accuracy is limited due to the effects of external variables: "errors" such as wind forecast uncertainties. The focus of the Trajectory Prediction Uncertainty simulation at NASA Ames Research center were the effects of varied levels of accuracy on operator's tool use during a time based metering task. The simulation's environment also provided a means to examine the relationship between an operator's reliance strategy and underlying trust attitude. Operators were found to exhibit an underlying trust attitude distinct from their reliance strategies, supporting the strategic use of the Human-Automation trust scale in an air traffic control environment..

ACKNOWLEDGEMENTS

A master's thesis is never completed in a vacuum, yet I have to give special thanks to all the people who contributed above and beyond to the formulation and development of my research and myself. I would like to acknowledge the support of the San José State University Research Foundation and NASA Ames Research Center's Human Integration Division in making it possible for a student to experience and work in such a rich and supportive environment. Special thanks are for Joey Mercer, M.S. for his support and unwavering guidance- by which I mean telling me to make sure I can explain it drunk at a bar- through all stages of my process. My committee went above and beyond, reviewing my work more often than strictly necessary and certainly discussing the complexities of the problem with me and each other beyond the requirements of the appointment. And gratitude beyond gratitude goes to Dr. Kevin Jordan who took a chance on a student he saw promise in by offering me a job resume unseen. I will forever be in your debt.

Thank you to everyone who has paved this path with wisdom, endurance and humor. In many ways this is your accomplishment as much as mine.

Table of Contents

Introduction.....	1
Dynamic Predictive Systems in Air Traffic Control.....	1
Theoretical Foundations of Calibration, Temporal Specificity and Trust	3
Simulating a Dynamic Error Environment to Examine Trust and Reliance.....	13
Method	17
Design	17
Test Environment.....	17
Operational Concept.	18
Management Tools.....	18
Error Conditions.....	19
Subjective Data	22
Variables	23
Materials	24
Participants.....	25
Procedures.....	26
Recruitment.....	26
Training and Testing.....	26
Results.....	27
Design 1: Performance Analysis for Reliance Strategy.....	28
Design 2: Measuring for Underlying Trust Subjectively.....	31
Design 2.1.	31
Design 2.2.	35
Discussion.....	36
Limitations and Suggestions for Further Research.....	41
References.....	43

List of Tables

<u>Number</u>	<u>Table Name</u>	<u>Page</u>
1	Wind forecast errors, relative to the environment winds, per level.	29
2	Wind and aircraft performance error levels combined for six tested error conditions.	30
3	Adaption processes for the Jian et al. Human-automation trust scale.	31
4	Run schedule, only runs 1-12 were part of the primary experimental design included in this analysis.	33
5	Spearman's rho for the correlation between DIA and DTT grouped by error condition.	37
6	Descriptive statistics for repeated measures ANOVAS.	41
7	Estimates for the human-automation trust scale ANOVAs.	42
8	Cronbach's alpha for the eight items used in the human-automation trust scale.	45

List of Figures

<u>Number</u>	<u>Figure Name</u>	<u>Page</u>
1	Lee and See's model of trust in automation evolution based on their 2004 review of literature.	16
2	The Theory of Planned Behavior Model, modified from Ajzen 1991.	19
3	Simulated Atlanta airspace with two test sectors and scheduling meter fix.	26
4	Schedule based tools used in the Trajectory Prediction Uncertainty experiment prototyped in a mid-term NextGen environment.	28
5	Error distributions for aircraft performance, as applied to all traffic per run.	30
6	Jian et al.'s (2000) Human-Automation trust scale as it was presented live.	32
7	Simulation lab layout, separating Worlds 1 and 2.	34
8	Scatterplots of all Spearman's Rho Correlations. Only Combined, RL and RM were significant ($p < .05$) in the relationship between DTT and DIA, (X, Y), respectively.	39
9	Trust Statement Type (Positive, Negative) means over Error Condition.	44

List of Acronyms

FAA	Federal Aviation Administration
A	Traffic Scenario A
AOL	Human Systems and Integration Division's Airspace Operations Laboratory
ATC	Air Traffic Control
ATM	Air Traffic Management
B	Traffic Scenario B
DIA	Difference in Advised and Issued Speeds
DTT	Delay Tag Time
L	Large
M	Moderate
N	Nominal
NextGen	Next Generation Air Transportation System
R	Realistic
SJSURF	San José State University Research Foundation
TPU	Trajectory Prediction Uncertainty Experiment
TS	Trial Plan Speed

Introduction

Dynamic Predictive Systems in Air Traffic Control

The Federal Aviation Administration (FAA) is currently developing the next stage in air traffic-control, known as NextGen (next generation air transportation system). The goal is to advance today's very safe but antiquated system in anticipation of the rising number of flights expected as 2025 approaches (Joint Planning and Development Office, 2012). As traffic increases, the ability of air traffic controllers (ATC) to maintain personal oversight and manually separate every aircraft in their sectors will potentially come under error-causing strain (Prevot, Homola, Martin, Mercer, & Cabrall, 2012). To avert this strain, automated systems with decision-making and predictive skills are being developed to work in conjunction with the operators and assist in their duties. Both Europe and the USA are working to develop systems that support this human-automation collaboration in air traffic management (Martin, et al., 2011). As this automation is developed to work cooperatively with human operators, issues relating to trust and reliance strategies become barriers to both experimentation with and adaptation of these new tools. This thesis focuses on the building blocks of assessing and understanding the relationship between an operator's underlying trust attitude in his / her system and the actual behavioral-based reliance strategies resulting from the interaction of the human and the automation.

While there is a body of work in the field focusing on the relationship of an operator to automated systems which function in a hit / miss / false alarm manner, such as

conflict detection, there is another branch of NextGen technology to which this format does not apply. Wickens et al, 2009 identified false alarms, or “crying wolf,” as one of the most detrimental errors an automated system can make when working with a human operator. However, a portion of the NextGen function’s advisories do not lend themselves to a correct / incorrect classification. These systems are predictive by design and as such are dependent on the quality of the data with which they make their advisories. These predictions are estimates of the future states of aircraft, such as “how soon will an aircraft arrive at a certain point?” or “when the aircraft arrives at waypoint X, what will the speed and altitude be, based on current flight plans and conditions?” But the algorithms producing these predictions are only as accurate as the data fed into the system. Of specific concern to the calculation of speed and arrival times are aircraft performance assumptions and weather data. Weather variables, such as wind, can significantly affect the speed of an aircraft. Not only is weather difficult to forecast accurately, the weather environment is not static. Instead, it is a dynamic environment that may change suddenly and with little warning. Thus it is highly unlikely these systems will ever be 100% accurate. In this situation, a false alarm or miss could be read as anything not 100% correct. However, the dynamic nature of the system provides information to the operator that is degrading in accuracy over a continuum. This suggests predictions that are only mostly accurate may contribute to the operator’s decision making process. Because of the expected air traffic increase in the next 20 years, any system that decreases the cognitive workload and expands the capacity of the human controller should not be discarded unnecessarily.

Trust in automation research has identified several traits considered to be necessary for an operator to work in an environment where the accuracy of the system is not completely stable. An ideal operator would approach the automated system with an understanding of both the current capabilities and the limitations of the system. This operator would have proper *calibration* (Lee & Moray, 1994). However, because these systems can fluctuate in accuracy due to changes in data such as weather, operators must not only properly calibrate their use and expectations from the automation at the start of their relationship with the system, but also recalibrate at the beginning of every shift. This recalibration is termed *temporal specificity* and is necessary for an operator to continue to use a dynamic predictive system both safely and to the full extent of its capabilities (Lee & See, 2004). The goal for operators working with predictive NextGen systems is that through proper calibration and temporal specificity they both relinquish and regain the use of their tools based on environmental conditions such as errors in weather data.

Theoretical Foundations of Calibration, Temporal Specificity and Trust

Both calibration and temporal specificity are measured using the behavior of the human operator, typically operator reliance strategies when interacting with the automated system. These experiments are conducted using a human-in-the-loop design, building complex simulations of real traffic management environments with real air traffic controllers. However, while this is an appropriate method to test the overall performance of a system which includes automation, environmental conditions, and human operators, it can be difficult to identify the causal relationships between air traffic

controller's trust in automation, their reliance strategies, and their ability to recalibrate in a temporally specific way.

Trust has been a factor in understanding a human's relationship to automation since Parsuraman and Riley's (1997) foundational paper. They identified three potential pitfalls in human-automation relationships which are all types of calibration failures: misuse, disuse, and abuse. Misuse is the operator relying on the automation beyond the capabilities of the system, whereas if operators are displaying disuse they are underusing a system and potentially compromising the safety of their task. Additionally, the abuse of automation concerns the development and implementation of automation without consideration of the human and integration of the human-automation relationship into the system as a whole. While trust has been found to affect these calibration issues, it is difficult to separate trust from other contributing elements in the behavioral expression of reliance strategy.

Two models developed in the last 20 years help explicate the relationships between the variables encountered in a human-in-the-loop experiment. Both Lee and See (2004) and Ajzen and Fishbein (2010) identified the reliance action or behavior (respectively) as the objective expression of the relationship between the human and the automation. Lee and See's work greatly advanced the field by presenting a model which synthesized research from multiple disciplines from a human factors perspective, drawing from social psychology concerning interpersonal trust, trust in organizations and systems, as well as physiological and neurological research. Their model, as seen in Figure 1, proposed reliance action as a conclusion based on a progressive distinction between trust,

intent, and reliance. Lee and See's model focused on not using trust as an umbrella term, but instead recognizing that trust is actually an attitude that is related to, but not directly predictive of actual reliant behavior. Lee and See's model suggested that trust is first developed based on what the operator knows about the system before experiencing it first-hand. As shown in Figure 1, the beliefs an operator uses to construct this first initial trust rely heavily on second-hand knowledge such as gossip and system reputation and not from first-hand experience of the system. When an operator is exposed to a system and obtains first-hand knowledge on which to formulate trust, while these other information sources may still matter, of greatest importance is the operator's personal experience. This then moves into a discussion of history-based trust, which has different contributing features and implications (Merritt & Ilgen, 2008). However, the focus of this thesis is initial trust and training implications, and second-hand beliefs are assumed to underlie trust formation.

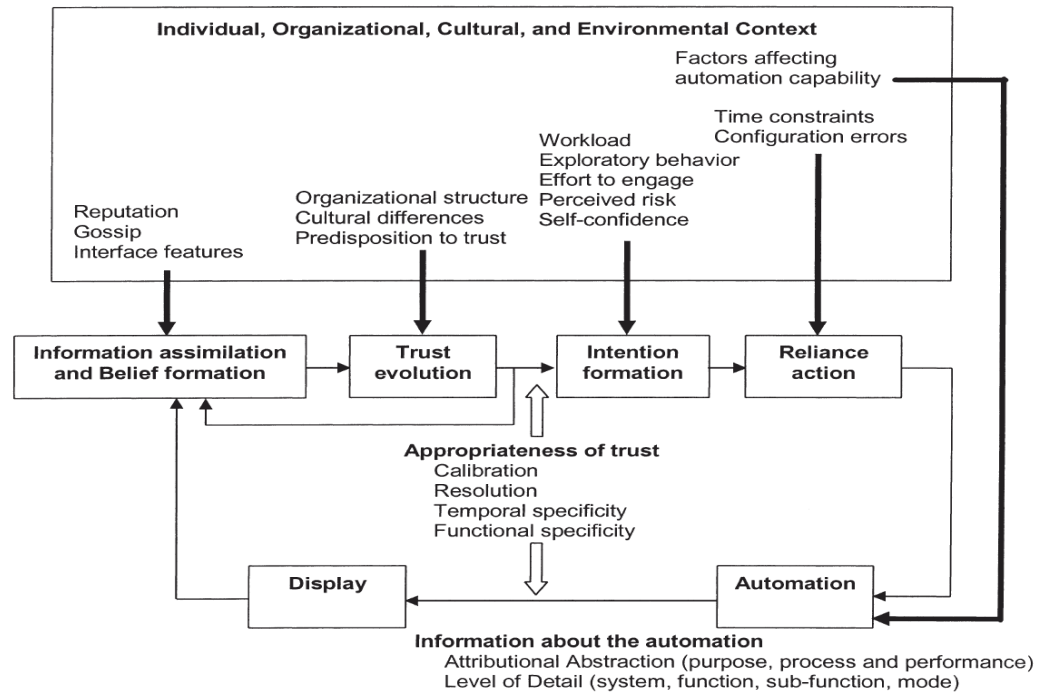


Figure 1: Lee and See's model of trust in automation evolution based on their 2004 review of literature. Copyright the Human Factors and Ergonomics Society 2004; reprinted with permission.

According to Lee and See's model, *trust evolution*, *intention formation* and *reliance action* are progressive and interrelated constructs that are nevertheless distinct. This model promoted the possibility of intervening variables (i.e. workload, system error, or time constraints), moderating a reliance action despite the true nature of the individual's trust. In this case, *reliance* is a specific term defined most commonly as usage of a target automated system. The conclusion that reliance can be modified through intervening variables such as workload or environmental considerations such as system errors is founded in work done by Kirlik (1993). He developed a modeling technique to assess and account for unexpected strategic behavior in a human-automation

study. The unexpected results involved human use of autopilot automation in ways that the designers of the system did not intend. It was the case of significant individual differences in strategy that had Kirlik examining the study through the context of the interaction between aid design and task context. The conclusions drawn from his paper are specifically relevant to an air traffic management environment, in that individual operators use a strategy specifically tailored to their preferences even though all received the same fundamental training. Kirlik presented his conclusion on the importance of task context as a moderating effect on individual operator strategy for automation use to achieve optimal performance. Reliability and accuracy alone are not sufficient for automation to benefit strategy. It is the operator's strategic management through the context of the environment that realizes the benefits of the automation (Kirlik, 1993). This means that an operator's choice not to use the automation because of an inappropriate environmental context is just as important as him/her using it in the environments where he/she finds it beneficial. While this may seem like a foregone conclusion, it is fundamental to the logic underlying Lee and See's instance that environmental context can moderate reliance despite intent to use the automation.

Unfortunately, while Lee and See's (2004) model is extremely valuable and popular in the field, it glossed over a few important distinctions made by its precursor, the Theory of Planned Behavior by Ajzen and Fishbein (1991). Of concern is the fact that Lee and See generalized the relationship between trust, intent, and reliance (behavior) and summarized Ajzen and Fishbein's work in this manner: "behaviors result from intentions and that intentions are a function of attitudes. Attitudes are in turn based

on beliefs.” (Lee & See, 2004, p. 53). In Figure 1, not only do reputation, gossip, and interface features contribute to trust evolution, but organizational structure, cultural differences and predisposition to trust are also considered to be directly contributing to trust evolution. However, according to Ajzen and Fishbein, the organizational structure / cultural differences Lee and See cite as directly contributing to trust actually form a separate variable which directly influences intent separate from its effect on trust.

This distinction is considered relevant to air traffic management environments because it is necessary to affect initial trust and understand potential pitfalls during the training and implementation of NextGen systems. Figure 2 below is a re-drawn version of Ajzen’s Theory of Planned Behavior model. This model separates trust from two other variables thought to contribute to intent. In this model, what Lee and See identified as organizational structure / cultural differences is classified as the *subjective norm*. The subjective norm is understood as the operator’s perception of other relevant person’s beliefs about the unfavorable or favorableness of performing the target behavior (Ajzen, 1991). This identifies social pressure on an individual to form intent irrespective of personal attitude. Having the subjective norm function as a separate variable from trust attitude is considered relevant in this thesis due to the confined and team nature of an air traffic management facility. The social state of the air traffic management facility may have just as much influence on intent as attitude, independent of the actual trust attitude of the operator. This issue is considered important here because (as previously mentioned) Lee and See stated regarding Ajzen and Fishbein’s model, that “behaviors result from intentions and that intentions are a function of attitudes.” (2004, p. 53). This

is an interpretation of the Theory of Planned Behavior model which glosses over the weighting system which suggests either the subjective norm or the *perceived behavioral control* (a factor identified by Ajzen and Fishbein but not addressed in this analysis) could overpower the influence of attitude on intentions (1980, p. 58). Therefore, this paper maintains Ajzen and Fishbein's distinction among attitude, subjective norm and perceived behavioral control due to both the potential for bias in the social structure of the air traffic management facility and potential for personal control expectancies air traffic controllers might hold. Here trust attitude is not considered to be a direct precursor of Intent due to the additional variables not controlled in this experiment.

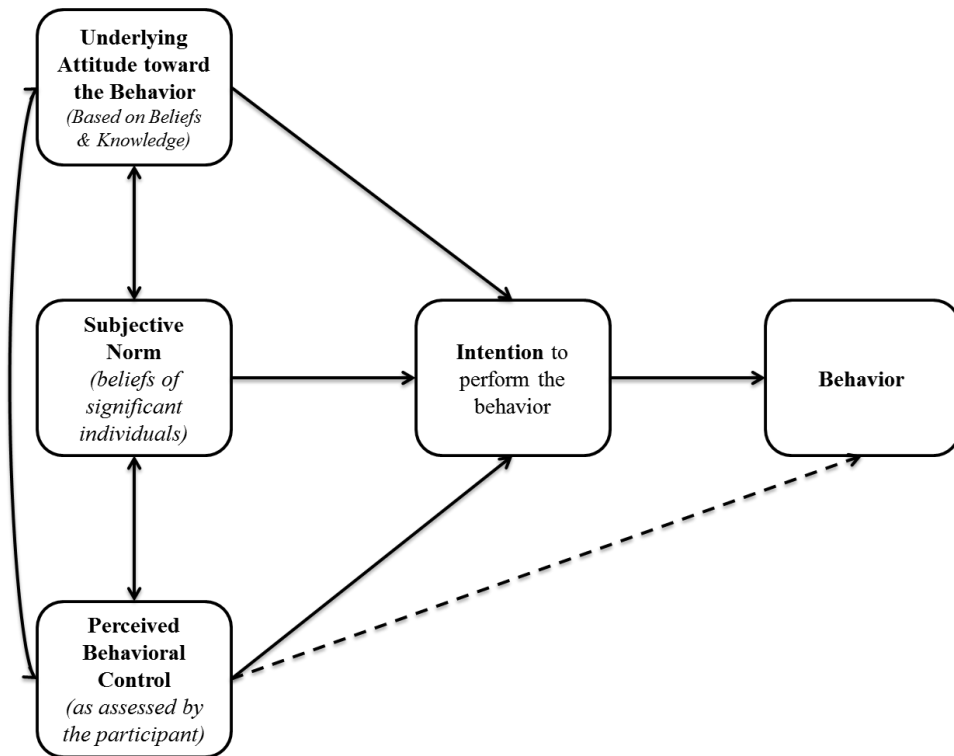


Figure 2: The Theory of Planned Behavior model, modified from Ajzen 1991.

Although for this analysis the distinction between attitude, subjective norm and personal behavioral control is considered important to maintain, it is significant that in the literature there is some dissension over the extent of the effect of what Ajzen and Fishbein term “external variables” or variables which moderate attitude, subjective norm or perceived behavioral control. Current human factors papers on human-automation trust discuss the influence of affective states such as “liking” that may contribute to the more analytical approach of the Theory of Planned Behavior model (Merritt S. , 2011). Ajzen and Fishbein maintain that these external variables exert their effects on behavior through the attitude, subjective norm and perceived behavioral control, and as such are explained in their model (Fishbein & Ajzen, 2010, p. 400). However, Lee and See gave

affective factors more weight in their analysis. This issue is referenced as a caveat to constrain interpretation of the results presented here. This thesis solely focuses on *trust attitude* and its suggested construct distinction from *reliance action*, not only through intent, but through the mediating factors of subjective norm and perceived behavioral control. Affective factors may or may not influence intent directly or as moderated by attitude, but are not measured explicitly in this experiment and no conclusions are made about affective factors. When examined in conjunction, both Lee and See's and Ajzen and Fishbein's work suggests trust is a construct related to reliance, but distinct from it due to the number of intervening factors. Examining this distinction is the focus of this analysis because of the contradictions in the field over the exact relationship of trust and reliance.

Several previous human-in-the-loop studies examined air traffic management environments which contained errors. Masalonis and Parasuraman (2003) examined reliability and trust during an air traffic management conflict detection task. Discussed was the need for an examination of automation information usage, as opposed to a simple acceptance or rejection seen in signal detection theory based models. They specifically examined the difference in individuals who were trained to expect some error in the system and the potential sources of that error and in individuals who were not trained. However, their first hypothesis contrasts with the model Lee and See published a year later. Malasonis and Parasuraman hypothesized that trust would fluctuate when automation reliability fluctuated for the trained participants. This does not reflect the distinction between reliance and trust suggested by Lee and See's model. Lee and See's

model expected that the opposite of Masalonis and Parasuraman's hypothesis was true, that trust would remain stable in fluctuating reliability conditions if participants were trained on why and where unreliability might occur. While Masalonis and Parasuraman's hypothesis was supported (i.e., they found that trained participants had high trust in high reliability situations and low trust in low reliability situations), there may be a difference in measurement and operationalized definitions driving the conflict between these results and Lee and See's model. In Masalonis and Parasuraman's experiment, individuals were rating their trust in system accuracy during a specific scenario. The important factor here is that the individuals were rating a system they knew to have a fluctuating accuracy at a specific point in time. By adjusting their trust in the accuracy of the automation between conditions with differing error levels, these participants were actually evidencing calibration and temporal specificity. What Lee and See classified as *trust* is an underlying attitude based on a participant's beliefs in the overall quality of the system regardless of a temporary state change, not trust in accuracy as driven by a dynamic external factor such as current error state. It is this fine line between measuring user's assessment of current accuracy and underlying trust in the system that is important when examining the relationship between trust and reliance.

As laid out in Lee and See's Model, initial trust in the system along with subjective norm and perceived behavioral control is a precursor to an operator's first calibration to an automated system. Incorrect calibration can lead to a number of safety issues, especially in situations such as aviation collision warning systems (Lee & Moray, 1994). Once calibrated, the temporal specificity of the operator to recalibrate in real time

based on environmental changes should adjust their reliance strategy. While Lee and See used the term ‘reliance action’, the experiment examined here addresses changes in reliance over a defined block of time containing multiple actions and therefore uses the term ‘reliance strategy.’ If the distinction laid out by both Lee and See and Ajzen and Fishbein is true, the underlying trust in the automation should remain more stable across varying error environments than an operator’s reliance strategy. Ideally, an operator should evidence two traits: (a) the ability to both relinquish and regain the use of some or all tools based on changes in error environment, and (b) a fairly stable trust in the overall automated system. An operator who is not temporally specific is likely to relinquish a tool and not regain it or, based on initial calibration, never use a tool.

Simulating a Dynamic Error Environment to Examine Trust and Reliance

The Trajectory Prediction Uncertainty Simulation (TPU) was run at the NASA Ames Research Center’s Airspace Operations Laboratory in January 2013. Unlike Rovira and Parasuraman’s (2010) study concerning imperfect automation in an air traffic management conflict detection program, the automation in TPU was not based on a miss / false alarm model. Instead these prototype features of the NextGen system were designed to simulate a continuum of errors. This simulation varied levels of wind forecast error and predicted aircraft performance (system) error. Data was collected on the performance capabilities of the operators to achieve metering times within +/- 25 seconds and maintain a safe environment. The air traffic controllers in this simulation had access to a number of predictive tools which used the weather and aircraft performance data. One of the preferred tools was the delay tag time, or the time

displaying how early or late the aircraft was in arriving to the meter fix. This time directly affected meeting the simulation goals regarding metering times. When the automation was working with perfect or close to perfect data, these times would match up to the controller's own perception of where an aircraft was in their sector and how fast it was going. However, when the system was fed incorrect weather and performance data, the times stopped creating a logical representation of the behavior of the aircraft in the sector. This level of system transparency (Gao & Lee, 2006), while not actually displaying the calculations the automation was undertaking, allowed the controllers to make a rough assessment of the automation accuracy based on their own expertise and knowledge of the airspace. Action sequences taken by the controllers based on this delay tag time were the focus of this analysis, specifically sequences in which an operator requested an advisory on a speed and then subsequently issued a speed to the aircraft. The speed issued by the controller may be identical to the automation's advisory or may be a modification of the advisory. Of most interest was the extent of modification and whether the indicator of the state of the automation, the delay tag time, could be used to measure the relationship between the automation accuracy and the reliance strategy of the controllers. In order to meet the requirements of identifying ideal operator traits, the first goal of this thesis was to assess whether this was (a) a viable means for identifying the relationship between automation state and reliance strategy and (b) whether these controllers evidence temporally specific reliance strategies. Temporal specificity in this instance was measured over changes between fifty five minute blocks, and not minute by minute.

In addition to measuring reliance strategies, the operator's trust in automation was also assessed at multiple points during the simulation. This measure was obtained using the Human-Automation Trust Scale developed by Jian, Bisantz, and Brury (2000). This scale was chosen because Jian, et al. empirically generated a list of trust statements for human-automation research based on a three-phase experiment (Jian, Bisantz, & Drury, 2000). This scale was then used in research on automated combat identification systems (Wang, Jamiseon, & Hollands, 2009) and in understanding trust in medical technology (Montague, Kleiner, & Winchester 3rd, 2009). Use of this scale in the TPU study included seven of the original twelve keywords which were identified as relevant to human-automation relationships by Jian, et al., plus a single addition of an eighth keyword that was felt more appropriate for the simulation environment. The original seven items were considered the most relevant to the system capabilities and goals in this simulation. *Intent* was not addressed directly because participants in this simulation were instructed to use the automation to the best of their ability until there came a point when they were unable to do so. Because of this, all participants had intent to use the automation driven by the subjective norm of this design. As mentioned previously, Ajzen and Fishbein stated (1980, p. 58) that each primary factor (attitude, subjective norm, and perceived behavioral control) contributes to the formation of intent in a weighted manner. The sum of those weights equals intention. In the case of this simulation, the subjective norm of the need to use the automation is assumed to outweigh attitude and perceived behavioral control, driving intention. However, this does not necessarily mean that in real air traffic management facilities the subjective norm will

carry the same weight as in this experiment. The goal in using the Human-Automation Trust Scale was to assess whether (a) it evidenced construct validity in this environment, and (b) whether the scores suggested a fairly stable underlying trust or a fluctuating trust in the system.

In summary, this paper will discuss the following lines of questioning: (a) Can a temporally specific operator be identified through an objective relationship between the delay tag time presented by the automation and the subsequent actions taken by the controllers in regard to speeds? (b) Does Jian, Bisantz and Dury's (2000) Human-Automation Trust Scale display construct validity in this air traffic management environment to measure underlying trust? And (c) does the contrast of these two measures suggest trust attitude and reliance strategy are related but distinct constructs?

Method

Design

Test Environment. Shown in Figure 3, to conduct this human-in-the-loop simulation, two en route sectors (one-high altitude and one low-altitude) were created based on historical traffic of the Atlanta airspace feeding into the northwest meter-fix of Atlanta's Terminal Radar Approach Control. The aircraft simulated in this environment were equipped with Flight Management Systems and Automatic Dependent Surveillance-Broadcast -out capabilities. All instructions from the controllers to the pilots were issued via voice communications.

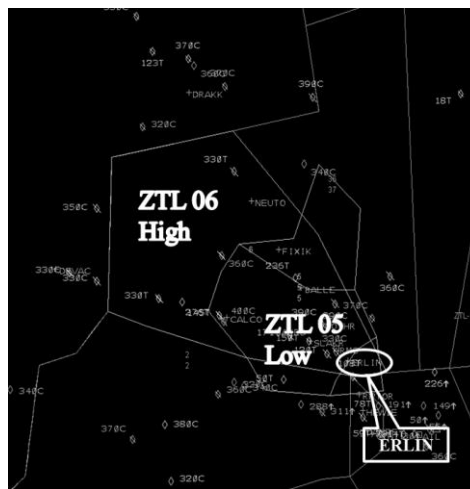


Figure 3: Simulated Atlanta Airspace with two test sectors and scheduling meter fix.

In parallel, two different teams were simultaneously tested on this airspace, with each team consisting of two test sector controllers and four confederate controllers. Confederate controllers staffed the surrounding airspace (high and low ‘ghost’ sectors). These confederates enabled the simulation to staff the adjacent airspace, while limiting the number of participants necessary. The two independent teams (or ‘worlds’) were

presented with the exact same traffic scenarios. The simulation employed two different traffic scenarios, A and B, which had similar but different flow problems to provide a broader range of testing and assist in avoiding learning effects.

Operational Concept. The traffic scenarios simulated current-day metering operations during an arrival push. The high sector controllers were expected to condition traffic in such a way that the traffic feed to the low sector controller supported their goal of delivering traffic to the meter-fix ERLIN within +/- 20 seconds of the scheduled arrival time displayed on their meter list and in their aircraft data blocks (see Figure 4). The complexity of the problem was compounded by simultaneous over-flights and departures in the test sectors which must still be managed for safety.

Management Tools. Controllers had access to prototype metering tools envisioned as part of a NextGen 'mid-term' time frame, which assumes the presence of more accurate surveillance data and advanced decision support tools. Available to the controllers were tools such as a meter list, conflict list, and a trial plan tool. Trial planning was available for speed, trajectory, or a combination of both. This analysis used data provided by the trial plan speed (TS) function. This function allowed the controller to request a speed advisory from the automation. The subsequent advisory (shown in cyan on the controller's scope) included both a speed and an updated delay tag time. As seen in Figure 4, this delay tag time reflects the automation's best prediction of how early or late the aircraft will be to the meter fix ERLIN if the advised speed is issued.

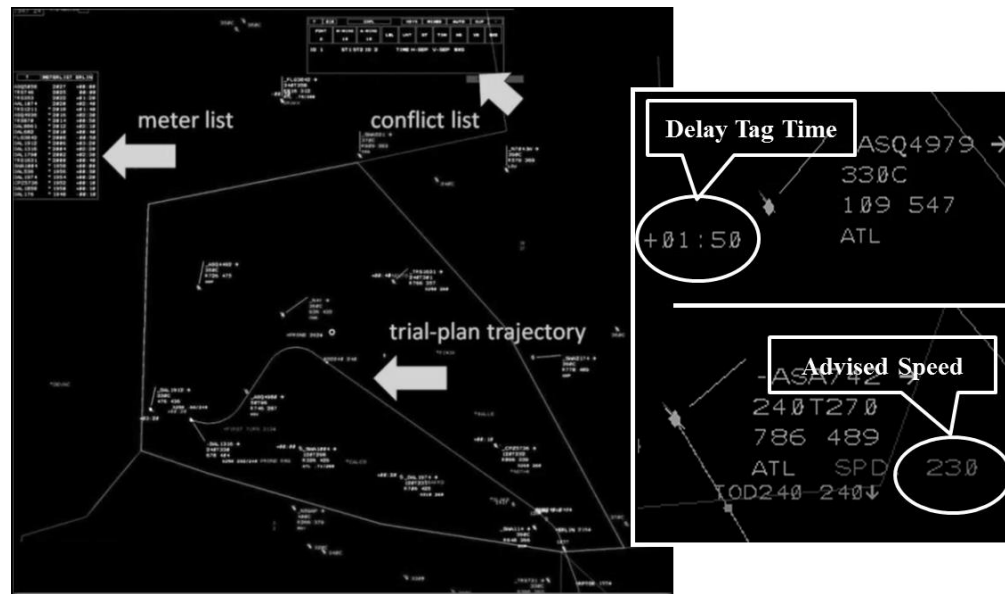


Figure 4: Schedule based tools used in the Trajectory Prediction Uncertainty experiment prototyped in a mid-term NextGen environment.

Error Conditions. In order to simulate error degradation in the automated system, two sources of error were targeted, wind forecast and aircraft performance. Wind Forecast Error assumes a difference between the actual wind environment and the forecast values the automation uses to predict estimated arrival times used in the meter list and all trial planning tools. As seen in Table 1, four levels of wind forecast error were created (Nominal [N], Realistic [R], Moderate, [M], and Large [L]). An extension of this experiment not addressed in this thesis did include additional levels, but the primary design is the sole focus of this thesis. More information about TPU can be found in Mercer, et al. (in press).

Table 1
Wind Forecast Errors, Relative To The Environment Winds, Per Level

Condition Level	Wind Conditions: forecast-actual	
	<u>A</u>	<u>B</u>
--		
Nominal	0	0
Realistic	+ 10 knots	-10 knots
Moderate	+ 20 knots	-20 knots
Large	+ 30 knots	-30 knots

Note. Scenario A over-predicts and scenario B under-predicts. This was done to create a balanced design and eliminate forecast bias.

Aircraft performance error was based on a discrepancy in the automation's calculations of the ideal distance by which an aircraft should begin its descent to ERLIN and the actual top of descent flown by each aircraft. The difference between the assumption and actual was applied to all aircraft in the scenario according to a scaling factor. Three levels of performance error were applied, according to standard deviation, where the Realistic (R) condition had $\leq 5\%$ at 1 SD, $\leq 10\%$ at 2 SD, and $\leq 15\%$ for the remainder (5% two tailed). The Large (L) condition had $\leq 12.5\%$ at 1SD, $\leq 25\%$ at 2 SD and $\leq 31.25\%$ for the remainder (5% two tailed). The Nominal (N) condition did not contain any aircraft performance errors. Figure 5 is a graphic display of these distributions.

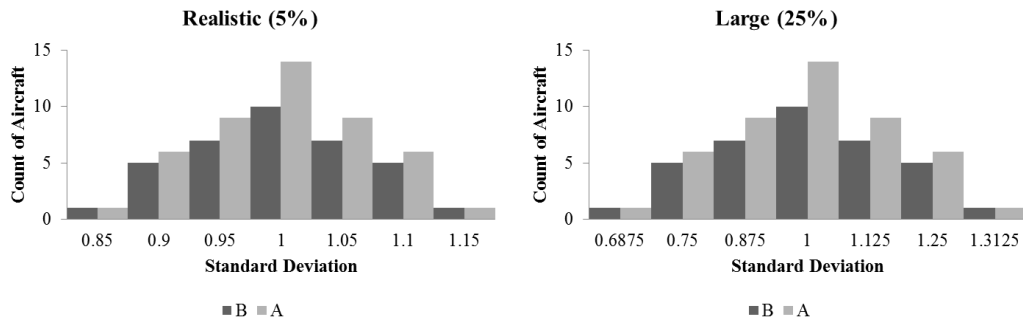


Figure 5: Error distributions for Aircraft Performance, as applied to all traffic per run.

An exhaustive set of wind and aircraft performance error level combinations was not tested. Only those combinations thought to be important by the research team were selected for this simulation. Table 2 contains the error conditions that were ultimately used. The first letter always refers to aircraft performance and the second always to wind forecast error. Each error condition was tested twice, once for each traffic scenario (A and B). These iterations were not randomized, and A always preceded B.

Table 2

Wind And Aircraft Performance Error Levels Combined for Six Tested Error Conditions.

<u>Aircraft Performance</u>	<u>Wind Error</u>	<u>Nominal</u>	<u>Realistic</u>	<u>Moderate</u>	<u>Large</u>
Nominal		NN			
Realistic			RR	RM	RL
Large			LR		LL

Subjective Data. Subjective data were also collected at the end of each error condition in the form of a Human-Automation Trust Scale (Jian, Bisantz, & Drury, 2000). This scale was adapted for the TPU simulation, retaining modified versions of seven of the original twelve items, with the addition of one statement added by the experimenters. The modified and original versions of this scale can be seen in Table 3.

Table 3

Adaption Processes for The Jian Et Al. Human-automation Trust Scale (2000).

<u>Action</u>	<u>Keyword</u>	<u>Keyword Type</u>	<u>Original Scale</u>	<u>Adapted Scale</u>
Adapted	Suspicious	Negative	I am suspicious of the system's intent, action or outputs.	I am suspicious of the automation's solutions.
Adapted	Wary	Negative	I am wary of the system.	I am wary of the automation.
Adapted	Harmful	Negative	The system's actions will have a harmful or injurious outcome.	The automation's suggestions will lead to a harmful or dangerous outcome.
Adapted	Security	Positive	The system provides security.	The automation provides safe solutions.
Adapted	Confident	Positive	I am confident in the system.	I am confident in the automation.
Adapted	Reliable	Positive	The system is reliable.	The automation is reliable.
Adapted	Trust	Positive	I can trust the system.	The automation is trustworthy.
Removed	Deceptive	Negative	The system is deceptive.	--
Removed	Underhanded	Negative	The system behaves in an underhanded manner.	--
Removed	Integrity	Positive	The system has integrity	--
Removed	Dependable	Positive	The system is dependable.	--
Removed	Familiar	Positive	I am familiar with the system.	--
Added	Comfortable	Positive	--	I am comfortable with the automation.

Note. 7 point Likert Scoring: Not at all – Moderately – Very. Presented in a stable, randomized array. Only items deemed most relevant to the ATM environment were kept due to space restraints.

The Human –Automation Trust Scale (Jian, et al. 2000) was presented using the online survey tool Lime Survey Software as the randomized array seen in Figure 6. Also seen in Figure 6, the anchor points used were *Not at All – Moderately –Very*. Additional subjective data not referenced in this thesis were also collected in conjunction with the human-automation scale.

* Please indicate your *current thoughts* about the automation for the following:

	1 Not at All	2	3	4 Moderately	5	6	7 Very
The automation provides safe solutions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The automation is trustworthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am suspicious of the automation's solutions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am comfortable with the automation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The automation is reliable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The automation's suggestions will lead to a harmful or dangerous outcome.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am wary of the automation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am confident in the automation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Resume later Next >>

Exit and clear survey

Figure 6: Jian et al.'s Human-Automation trust scale as it was presented live. Items were randomized within the array for every presentation.

Variables. The independent variable (IV) for this simulation was Error Condition, consisting of the six levels described in error conditions (NN, RR, RM, RL, LR and LL, page 29). Three dependent variables (DV) were collected for this analysis. The objective dependent variables concerned the TS function described in management tools (pg. 28). The first objective DV was the difference in the advised speed and the speed ultimately issued by the controller on the basis of that advisory. This is referred to as the Difference in Issued vs. Advised (DIA) in knots. This speed was operationalized by collecting all instances where a controller used the TS advisory request then issued a speed to the same

aircraft within two minutes of the advisory as long as no additional trial planning action was taken on the same aircraft within that time. This variable was considered to be representative of an operator's Reliance Strategy because the operator has the option to either accept the advisory as-is or modify it to some degree. Some level of modification was generally expected as operators tend to round speeds to whole numbers whereas the advisories are not rounded. The second objective DV was the delay tag time (DTT) displayed next to the aircraft's data block (see Figure 4) at the time of the TS advisory request. The subjective DV data were the Likert scores for the Human-Automation trust scale (Jian, et al, 2000).

Materials

The TPU simulation was conducted using MACS (Multi-Aircraft Control Simulator) developed at NASA Ames Airspace Operations Laboratory (AOL) for the purpose of testing ATM environments (Prevot, 2002). All apparatus used for this experiment was Aerospace Operations Laboratory equipment, and in order to mimic a real-world environment, replicated FAA-fielded workstation hardware as closely as possible. Each ATC workstation contained: Dell Precision PC, model T7400, 64-bit Vista Ultimate (SP1), Intel® Xeon® CPU – X5482 @ 3.20GHz (2), 8GB RAM; Cortron, Inc. keyboard, model 109-50008C; Measurement Systems, Inc. trackball, PN XCL250-1; Dell mouse, model MOA8BO; Barco ISIS (MDP-471) display; Toshiba Portege M700 Tablet PC; Plantronics USB headset; Delcom Products foot pedal, model 803653. All workstations were identically configured. Further information on human-in-the-loop simulations of ATC operations at the Airspace Operations Laboratory is available in

separate publications (Prevot, et al., 2010). Three separate rooms (Figure 7) were prepared for participant ATC positions and the confederate pilots.

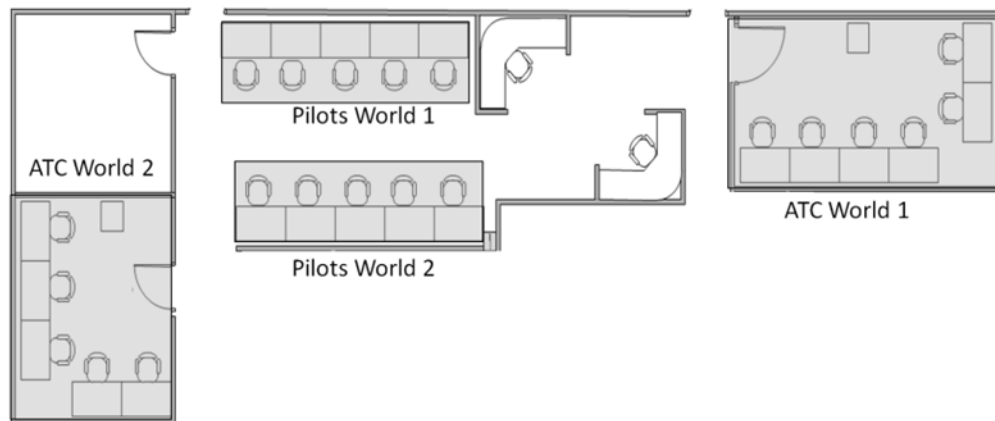


Figure 7: Simulation lab layout, separating Worlds 1 and 2.

Participants

The positions staffed to conduct this experiment can be grouped into four categories: ATCs for the test sectors, radar associate positions for each of the test sectors, ghost ATC sectors, and pseudo-pilots. Between both worlds, four ATCs handled the test airspace positions, and four ATCs handled the ghost airspace surrounding the test sectors. All ATC positions were staffed by retired air traffic controllers (eleven males and one female, all 40+ years old with normal or corrected-to-normal vision). Participation eligibility required retirement from a career in air traffic control in the USA. Twelve students from the aviation program at San José State University participated as pseudo-pilots. No data were collected on the pseudo-pilots, but they were all male and ranged in age from 18-30 years old. Each pilot was assigned to handle the planes in the sector of one ATC participant.

Procedures

Recruitment. The participants were recruited by the San José State University Research Foundation (SJSURF) through a previously constructed database of ATC participants. They were recruited based on experience obtained throughout their ATC career.

Training and Testing. The simulation consisted of two training days and five days of data collection, with a weekend in-between training and data collection. All simulations were conducted by members of the Human Systems Integration Division's AOL team. Informed consent was obtained by all participants on the first day of training. Two days of training from 8:30am-4:40pm, with an hour lunch break, were conducted. Day 1 began with a full briefing on the experimental design, MACS tools and functions, and the purpose of this study. Five 55-minute training runs were completed during Day 1 and again during Day 2. These included all developed scenarios, as well as baseline runs where the participants did not have access to any NextGen tools, such as the conflict list and trial planner. Breaks were administered after every run. Days three through seven comprised the data collection phase. All days ran from 8:30 am – 4:30 pm Monday- Friday, allowing for between four and five 55-minute runs each day. Short surveys, including the Human-Automation Trust scale were administered after each run, followed by a 15-minute break. Also, an hour was taken each day for lunch at noon. The experiment concluded with a debriefing session. Table 4 below shows the run schedule for data collection. This thesis is only concerned with runs 1-12, the primary

experimental design. Runs 13-23 were extension runs for exploratory designs used for other analyses by the AOL.

Table 4

Simulation Run Schedule, Only Runs 1-12 were Part of The Primary Experimental Design Included in this Analysis.

Time	M	T	W	R	F	Time
8:30	briefing	briefing	briefing	briefing	briefing	8:30
8:40	training	A-RL 5	A-RM 10	A-LXL 15	RRS4 20	8:40
10:00	A-NN 1	B-RL 6	A-LL 11	N-RR 16	N-LLS4 21	10:00
11:20	B-NN 2	B-LR 7	B-LL 12	B-LXL 17	RRDC 22	11:20
12:20	Lunch	Lunch	Lunch Post Primary Sim. Questionnaire	Lunch	Lunch	12:20
13:20	A-RR 3	A-LR 8	A-LL 13 Modified Schedule	RRBL 18	N-LLDC 23	13:20
14:40	B-RR 4	B-RM 9	N-LL 14	N-LLBL 19	Debrief	14:40
15:40	debrief	debrief	debrief	debrief		15:40

Results

Controllers in this simulation chose which tools they used during each run.

Because of this freedom, the controller working World 1 Low did not use the TS advisory past the second run. As the TS was the primary data source for this analysis, the data from World 1 low was removed from all analyses. A two tailed alpha of .05 was used for all tests.

Design 1: Performance Analysis for Reliance Strategy

Design one assesses the relationship between the delay tag time (DTT) displayed by the automation and the controller's actions and the difference in the advised speed and the issued speed (DIA). The goal was to attempt to identify any temporal specificity that may have occurred as the accuracy of the automation's assumptions fluctuated through the error conditions. To accomplish this, this analysis examined the predictive relationship between automation state (DTT and controller reliance strategy (DIA) across Error Condition. Error Condition used the same six levels outlined in the methods section (NN, RR, RL, LR, RM, LL), with an additional Combined level. This level was included to examine whether overall DTT predicts DIA.

Table 5

Spearman's Rho for the Correlation Between DIA and DTT Grouped by Error Condition

<u>Condition</u>	r_s	p	n	df
Combined	.139	.001	548	546
NN	.180	.076	98	96
RR	.003	.979	91	89
RL	.330	.001	90	88
LR	.071	.504	91	89
RM	.283	.005	97	95
LL	-.105	.351	81	79

Note. alpha=.05, two tailed.

Because of the violations of normality in the data (as assessed by a Shapiro-Wilk test $p < .05$), nonparametric Spearman's Rho was used to complete this correlation. Before running this test, three outliers were removed from this data. Outliers were

identified as instances when the operator issued speed was greater than 15% different from the advised speed, and are considered aberrations in the data. Two additional instances were removed for lack of a recorded Delay Tag Time. As seen in Table 5, Combined, RL and RM significantly correlated DIA and DTT. The scatter plots in Figure 8 graph the results, with the dotted line providing a visual reference for the zero mark. The results of Design 1 suggest that in two conditions (RL, RM), the error required a shift in operator reliance strategy to more closely track the information provided by the automation, while the other conditions were compatible with reliance strategies that did not require close (predictive) analysis of the automation's output. The shift from predictive (RL) to not predictive (LR) back to predictive (RM) evidence temporal specificity.

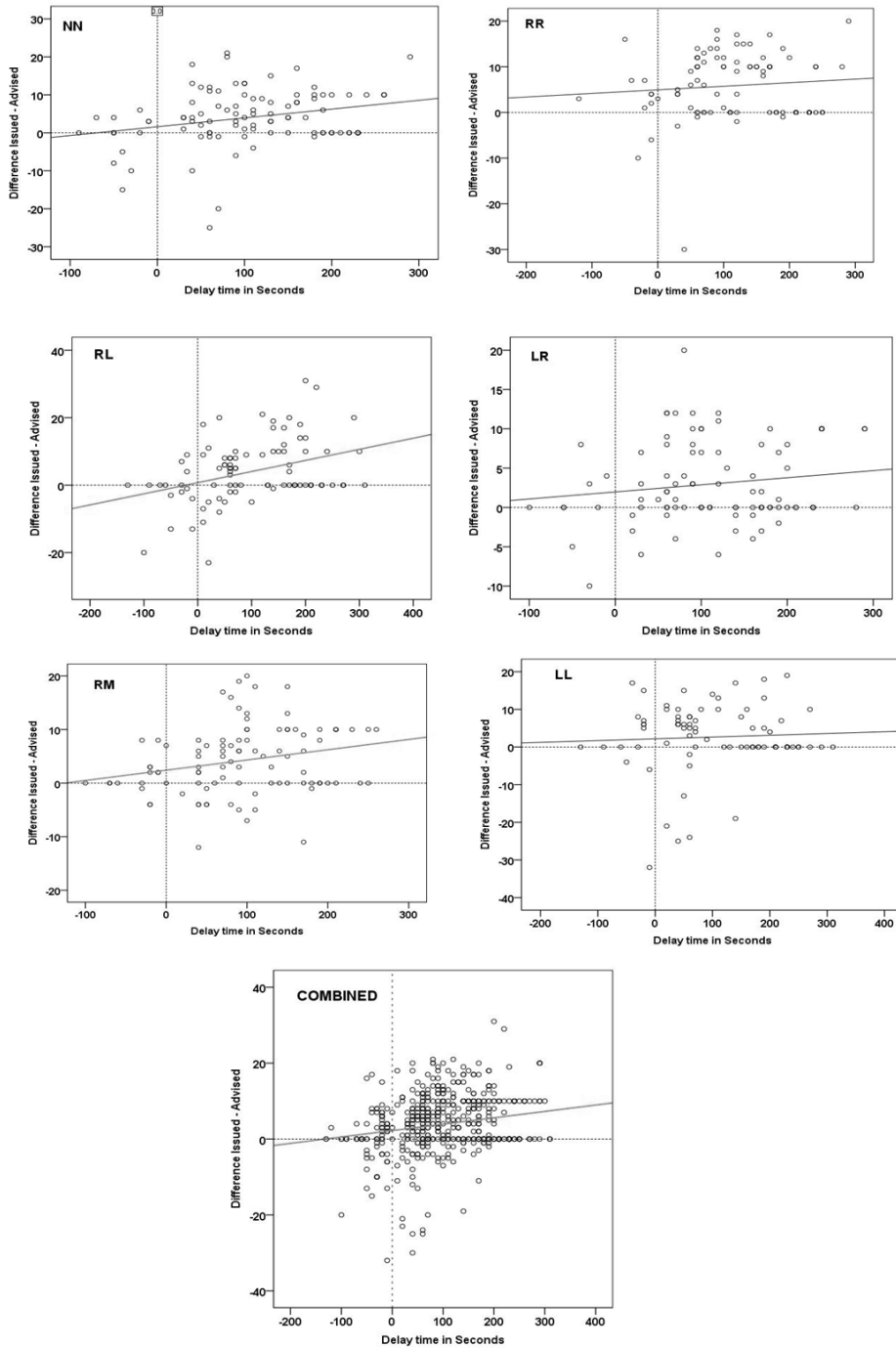


Figure 8: Scatterplots of all Spearman's Rho Correlations. Only Combined, RL and RM were significant ($p < .05$) in the relationship between DTT and DIA, (X, Y) respectively.

Design 2: Measuring for Underlying Trust Subjectively

Design 2 focuses on the Human –Automation Trust Scale (Jian, et al., 2000) and its construct validity in an air traffic management environment. Design 2.1 specifically groups positive and negative statements to assess whether they evidence significant separation by trust statement type and if they respond to changes in error condition. If there is an overall positive trust in the automation, positive statements were expected to group toward the seven in the Likert scale and negative statements towards the one. Design 2.2 examines the internal consistency of the scale using Cronbach’s alpha. As discussed in the introduction and methods, the Human-Automation Trust Scale was adapted to fit this environment using only 7 of the original 12 keywords. An eighth was added, Comfortable, by the researchers. Because all statements were randomly presented within a stable array (see Figure 6, page 33), recoding of reverse scaled data was necessary for Design 2.2.

Design 2.1. For analysis 2.1, Positive and Negative Statements were separated and compared across Error Condition using a 2X6 repeated measures ANOVA. The independent variables were: (a) Trust Statement Type (Positive ($n=5$) and Negative ($n=3$)) and (b) Error Condition (NN, RR, RL, LR, RM, and LL) and the dependent variable was mean Likert score. Although the data violated Shapiro-Wilk’s test for normality ($p<.05$), the ANOVA was considered robust enough to continue. Mauchly’s Test of Sphericity was also violated ($p<.05$), and a Greenhouse-Geisser correction was applied. There was a statistically significant interaction between Trust Statement Type and Error Condition on mean Likert Score, $F(5, 40) = 11.902, p = .002, \text{partial } \eta^2 = .598$. Descriptive statistics can

be found in Table 6. There was no significant main effect for Error Condition, $F(5, 40) = .409, p = .839$, partial $\eta^2 = .049$. However, a significant main effect was observed for Trust Statement Type, $F(1, 8) = .729, p < .001$, partial $\eta^2 = .989$. Pairwise comparisons for Trust Statement Type were significant, $m=3.907, SE = .145, p < .001, 95\% CI [3.574, 4.241]$.

Table 6

Descriptive Statistics for Repeated Measures ANOVAS

<u>Condition</u>	<i>Mean</i>	<i>SD</i>	<i>n</i>
<u>Positive Statements, missing pairwise comparisons removed</u>			
NN	6	0	9
RR	5.67	0.5	9
RL	4.78	0.833	9
LR	6.11	0.333	9
RM	6	0.707	9
LL	6.22	0.441	9
<u>Negative Statements</u>			
NN	1.78	0.441	9
RR	1.89	0.782	9
RL	2.78	0.833	9
LR	1.67	0.5	9
RM	1.56	0.527	9
LL	1.67	0.5	9
<u>Positive Statements, no pairwise removals.</u>			
NN	6.13	0.352	15
RR	5.73	0.458	15
RL	4.8	0.775	15
LR	6.13	0.352	15
RM	6	0.655	15
LL	6.33	0.488	15

Notes. The 2x6 repeated measures ANOVA removed missing cases pairwise, whereas the simple main effects for positive statements did not.

Two post hoc repeated measures ANOVAs were performed on Positive and Negative statement types, respectively. Negative statements were found to have statistically significant within subject effects, $F(5, 40) = 5.805, p = .022$, partial $\eta^2 = .421$, but no significant pairwise comparisons when a Bonferroni correction was applied. Means and standard deviations can be found in Table 6. Positive statements also had

statistically significant within subject effects, $F(5, 70) = 18.036, p < .001$, partial $\eta^2 = .563$. Pairwise comparisons revealed RL to be significantly different from every other condition except RR: NN ($p = .001$), RR ($p = .053$), LR ($p < .001$), RM ($p < .001$) and LL ($p < .001$). Means and standard errors for these comparisons can be found in Table 7. No other pairwise comparisons were statistically significant, singling out the RL condition.

Table 7

Estimates for the Human-Automation Trust Scale ANOVAS

Test	Mean	SE	95% CI	
			Lower Bound	Upper Bound
<u>2X6 Repeated Measures</u>				
<u>Trust Statement Type</u>				
Positive	5.796	0.077	5.618	5.975
Negative	1.889	0.115	1.625	2.153
<u>Error Condition</u>				
NN	3.889	0.073	3.719	4.058
RR	3.778	0.088	3.575	3.98
RL	3.778	0.188	3.343	4.212
LR	3.889	0.111	3.633	4.145
RM	3.778	0.147	3.439	4.117
<u>1X6 Simple Main Effect, Positive</u>				
NN	6.133	0.091	5.938	6.328
RR	5.733	0.118	5.48	5.987
RL	4.8	0.2	4.371	5.229
LR	6.133	0.091	5.938	6.328
RM	6	0.169	5.637	6.363
<u>1X6 Simple Main Effect, Negative</u>				
NN	1.778	0.147	1.439	2.117
RR	1.889	0.261	1.288	2.49
RL	2.778	0.278	2.137	3.418
LR	1.667	0.167	1.282	2.051
RM	1.556	0.176	1.15	1.961
LL	1.667	0.167	1.282	2.051

The results of these analyses suggest that the controller's faith in the automation remained strong throughout the fluctuating error conditions, as evidenced by the

maintenance of a significant distance between positive and negative trust statement types. This maintenance of distance, as seen in Figure 9, suggests that the scale is measuring an underlying trust construct and not a temporally specific reaction to automation state. Additionally, the statistically significant change in the RL condition for positive statements only indicates a reaction to what is likely the same factor which caused a shift from a non-predictive to predictive state during RL in the DIA vs. DTT data (design 1). This coordination with the reliance strategy data supports the conclusion that these measures are examining constructs related to the same automated system, and not, for example, trust in automation in general. The lack of a statistically significant main effect of error condition on trust scores is a reflection not of an issue with the scale, but the expertise of the controllers. Each controller completed their career with a 100% safety rating. While this experiment was designed to challenge them with detrimental error conditions, they maintained performance integrity throughout. The ceiling effect caused by their expertise accounts for the lack of the main effect of error condition and suggests the scale is remaining true to the actual experience of the controllers used in this study.

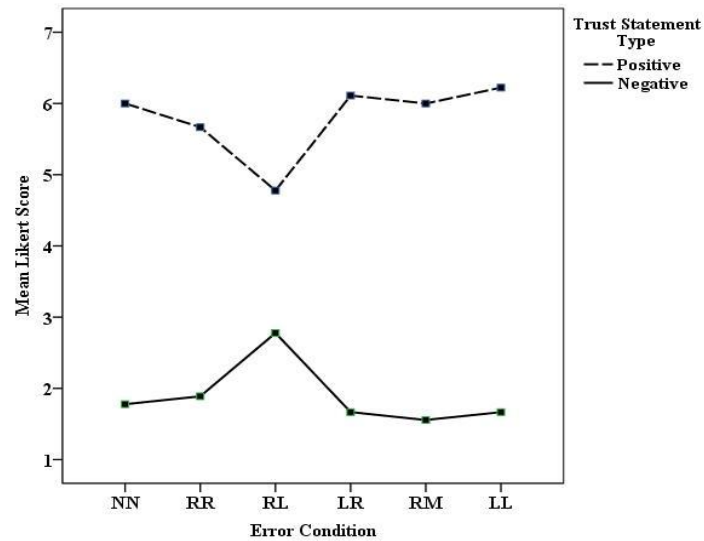


Figure 9: Trust Statement Type (Positive, Negative) means over Error Condition.

Design 2.2. Cronbach's alpha was used to test the internal consistency of the adapted Human-Automation Trust Scale (Jian, Bisantz, & Drury, 2000). Eight items based on keywords were tested: *Safe*, *Confident*, *Reliable*, *Trustworthy*, *Suspicious*, *Wary* and *Harmful*. *Suspicious*, *Wary* and *Harmful* were transformed before the analysis due to reverse scaling. The scale had a high level of internal consistency, with Cronbach's alpha equaling .923. As seen in Table 8, only the removal of the experimenter added keyword *Comfortable* would have improved the internal consistency of the scale. This supports the conclusion that the Human-Automation Trust Scale (Jian, Bisantz, & Drury, 2000) does measure the underlying trust construct.

Table 8

Cronbach's Alpha for the Eight Items Used in the Human-Automation Trust Scale

Keyword	Item Mean	Item Standard Deviation	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Safe	5.78	0.65	0.61	0.67	0.92
Confident	5.89	0.76	0.82	0.90	0.91
Reliable	5.72	0.75	0.87	0.87	0.90
Trustworthy	5.72	0.83	0.90	0.91	0.90
Suspicious	6.06	0.80	0.77	0.95	0.91
Wary	6.00	0.77	0.76	0.96	0.91
Harmful	6.28	0.57	0.66	0.66	0.92
Comfortable	6.17	0.62	0.55	0.84	0.93

Note. Suspicious, Wary and Harmful's data were transformed due to the reverse coding before analysis.

Discussion

The goal in conjointly examining both objective data pertaining to reliance strategy and subjective data on a controller's trust in automation was to provide measurement methods capable of identifying ideal controller traits for the NextGen environment. As discussed in the introduction, three exploratory questions were laid out at the beginning of this research to assist in that goal: (a) Can a temporally specific operator be identified though an objective relationship between the delay tag time presented by the automation and the subsequent actions taken by the controllers in regard to speeds? (b) Does Jian et al.'s Human-Automation Trust Scale (2000) display construct validity in this air traffic management environment to measure underlying trust? And (c) does the contrast of these two measures suggest trust attitude and reliance strategy are

related but distinct constructs? These questions drove the conclusions discussed in this section.

Temporal specificity is rooted in calibration, but represents a controller's ability to recalibrate their reliance strategies for the automation in real time. Of the three controllers who participated in using the TS advisory function, temporal specificity was found when comparing the delay tag time (the visible indicator of automation state, see page 24), and the difference in their issued speeds to the automation's advisories. It is to be expected that each controller would come into the simulation or a real air traffic environment with a personal reliance strategy based on their expertise, experience and preferences. However, the fact that they became more aware of the output of the automation as their environment fluctuated in accuracy during the RL and RM conditions shows a recalibration moving away from their personal strategies to include a higher awareness of automation state. It is the shift from not only non-predictive to predictive, but back again to non-predictive and once more to a predictive state, returning finally to a non-predictive reliance strategy which shows they are re-evaluating the state of the automation in each simulated error condition. Proper calibration is required to complete this temporally specific reaction, as a controller must be aware of the underlying capabilities of the system, especially in a nominal environment, and be able to compare current state against this mental model of the overall efficacy of their system. This calibrated base allows them to expand and contract their use of the system based on external environmental factors which may be affecting it. It is important to note that if an issue occurs which effects the actual underlying construction of the automation

permanently, controllers will eventually recalibrate to a less efficacious system and change their expectations. However, that situation was not the focus of this simulation, where error was introduced from sources outside of the automation, such as the weather forecast.

That the World 2 low controller chose not to use the TS advisory has some implications worth noting. This tool is known among researchers in the NextGen environment to be a desirable attribute of the automated system. While this controller may have not preferred the tool, when questioned after the simulation was completed, they stated that it was “jumping all over the place and wasn’t working correctly.” When queried if they found this to be true even in the nominal conditions where no errors were introduced, they replied an affirmative. While this controller may be properly calibrated and temporally specific in their use of other aspects of the automation, they began the simulation with an improperly calibrated view of this tool. This is an example of disuse, when the automation is not used by the human operator, and can lead to significant safety issues in the case of tools supporting critical tasks. If this simulation had been an actual training session, disuse of this kind would need to be addressed.

This Human-Automation Trust scale (Jian, Bisantz, & Drury, 2000) was chosen specifically because of its empirical background and use in other types of automation research. However, to the best of the author’s knowledge, before January 2013 it had not been previously applied to decision making aids in air traffic management. The ability of this scale to maintain stable scoring across fluctuating error conditions while discriminating between positive and negative statements, as well as the strong

Cronbach's alpha score promotes confidence in its applicability to air traffic management research. As mentioned in the results for design 2.1, a ceiling effect for performance was created due to the expertise of the participant controllers. They maintained their performance despite the effects of the fluctuating error conditions. As such, the Human-Automation Trust scale not recording a significant main effect for error condition was not surprising, and does not suggest an issue with the scale itself. Use of this scale with controllers at various stages of training should produce more variation. The strong inter-item reliability score attained by the Cronbach's Alpha shores up the conclusion that, while this scale was developed for a more general Human-Automation usage, it can be tailored to apply directly to a specific environment such as air traffic management. The result that the one keyword added by the experimenters which was not contained in the original scale was the only item which, if removed, would have improved the scale's score strongly supports the original research in creating this scale.

However, while the Human-Automation Trust scale (Jian, Bisantz, & Drury, 2000) does seem to measure the underlying trust attitude of a controller, this conclusion is put forth with some reservations and caveats. Due to the adaptation necessary to amend the scale to this simulation environment, this scale was not properly counterbalanced. If used in the future, counterbalancing is recommended. Also recommended is a factor analysis beyond Cronbach's alpha. As the Theory of Planned Behavior explicates (Ajzen, 1991), there are multiple factors influencing intention, and it is not impossible either the subjective norm or perceived behavioral control intruded on

the controller's answering of this scale. Finally, care is recommended when adapting this scale to a new environment.

Still, the performance of the Human-Automation scale in conjunction with the objective data examining reliance strategy is sufficient to support the conclusion that reliance strategy and underlying trust are related but distinct constructs. It should not be assumed that trust is a direct precursor to intent and intent to reliance. As shown in Design 1's objective data, controllers responded significantly to a shift in automation state both in RL, then again in RM. However, the fluctuations in the trust scores only occurred in RL, not RM. This suggests that whatever caused enough of a disturbance to impact both reliance strategy and trust was accounted for by the controllers and, once processed, was incorporated into a temporally specific reliance strategy that protected it from impacting trust a second time. But it remains important to measure both constructs, as trust is one component in creating the initial and continuing calibration on which temporal specificity is based.

In summary, to reference to the exploratory questions posed in the introduction and guiding this discussion: (a) An action sequence which requires the controller to interact with automation state such as delay times and speed advisories can objectively identify both calibration and temporal specificity over time. (b) To the extent that it is generalizable from this experiment, the Human-Automation Trust scale does seem to be measuring an underlying trust in the controllers which maintains construct validity. Finally (c), that both measures responded to the same error conditions but to different

degrees suggest that they do indeed measure distinct constructs that are nevertheless related.

Limitations and Suggestions for Further Research

Due to simulation constraints, this experiment was unable to additionally measure the subjective norms and perceived behavioral control of these participants. These factors are considered relevant due to the fact that either (Ajzen, The theory of planned behavior, 1991) could control an operator's intent to use an automated system if strong enough. In this simulation, it was a simulation requirement that all controllers use the automation to until they were unable to do so. Additionally, all of these controllers were extremely capable and confident in their abilities. As such, the subjective norm was expected to have overridden any general disinclination not to interact with the automation. This may not be true in a real air traffic management facility. Especially with the introduction of a new system, the social or political perceptions within a group in addition to controllers potentially losing confidence in being able to use the new tools makes these critical measures. Any further research is suggested to include measures for these factors and to verify them for an air traffic management environment.

Of concern are the potential impacts of failing to measure trust, subjective norm and perceived behavioral control when these predictive, dynamic systems are introduced. According to Lee and See's (2004) model discussed in the introduction, as well as the body of Ajzen and Fisbein's work, initial attitudes such as trust towards automation are formulated by knowledge often based on secondhand knowledge, gossip, the system's reputation and visual cues such as interface features. If not regulated, a situation such as

a single dissatisfied operator could heavily bias the attitude of a new operator towards a negative view of a new system. Then, if operators are found to have a biased underlying trust attitude, they can be provided with both experience and accurate knowledge to assist in formulating a new trust attitude (Merritt & Ilgen, 2008). Further research into the changing of attitudes is suggested before training procedures are finalized.

One final caveat must be made on the ability to generalize from this sample. Because of the retraining necessary for a controller to use NextGen tools, use of active controllers as participants is discouraged. And while these controllers were subject matter experts, a younger active controller may differ slightly. As such, in addition to the factor analysis suggested for the Human-Automation Trust scale, the examination of all of these measures in a more current environment with a larger sample is suggested.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211.
- Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall Inc.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior*. New York: Taylor & Francis Group.
- Gao, J., & Lee, J. D. (2006). Extending the decision field theory to model operator's reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 36(5), 943-959.
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Joint Planning and Development Office. (2012). *Concept of operations: For the next generation air transportation system*. Federal Aviation Administration, Department of Transportation, Department of Defense, Department of Commerce, Department of Homeland Security, NASA, Executive Office.
- Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35, 221-242.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operator's adaptation to automation. *International Journal of Human-Computer Studies*, 40, 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, 50-80. doi:10.1518/hfes.46.1.50_30392
- Martin, L., Kupfer, M., Palmer, E., Mercer, J., Callantine, T., & Prevot, T. (2011). Acceptability and effects of tools to assist with controller managed spacing in the terminal area. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics* (pp. 432-441). Orlando: Springer-Verlag Berlin Heidelberg.
- Masalonis, A. J., & Parasuram, R. (2003). Effects of situation-specific reliability on trust and use of automated air traffic control decision aids. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 533-537). Denver: SAGE.
- Mercer, J., Mienert, N., Gomez, A., Hunt, S., Kraut, J., Martin, L., . . . Wu, G. M. (in press). The impact of trajectory prediction uncertainty on air traffic controller performance and acceptability. *Aviation*. Los Angeles: American Institute of Aeronautics and Astronautics.
- Merritt, S. (2011). Affective processes in human-automation interactions. *Human Factors: The Journal of the the Human Factors and Ergonomics Society*, 53, 356-370.

- Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*(50), 194-210.
- Montague, E., Kleiner, B., & Winchester 3rd, W. W. (2009). Emperically understanding trust in medical technology. *International Journal of Industrial Ergonomics*, 39, 628-634.
- 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.
- Prevot, T. (2002). Exploring the many perspectives of distributed air traffic management: The multi aircraft control system MACS. *HCI-AERO* (pp. 149-154). Edmonton: American Association for Artificial Intelligence.
- Prevot, T., Callantine, T., Mercer, J., Homola, J., Smith, N., & Palmer, E. (2010). Human-in-the-loop evaluation of nextgen concepts in the airspace operations laboratory. *AIAA Modeling and Simulation Technologies Conference*. Toronto, Canada.
- Prevot, T., Homola, J., Martin, L., Mercer, J., & Cabrall, C. (2012). Toward automated air traffic control- investigating a fundamental paradigm shift in human/systems interaction. *International Journal of Human-Computer Interaction*, 28(2), 77-98.
- Rovira, E., & Parasuraman, R. (2010). Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52, 411-425.
- Wang, L., Jamiseon, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51, 281-291.
- Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51, 446-462.