DOT/FAA/TC-17/37

Federal Aviation Administration
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

Traffic Flow Management Tools: Guidance for Use, Integration, and Training: Part-Task Experiment 1

Brion Woroch, Engility Corporation, Inc.
Carolina M. Zingale, Ph.D., FAA Human Factors Branch, ANG-E25
Anthony J. Masalonis, Ph.D., Spectrum Software Technology, Inc

September 2017

Technical Report

This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/TC-17/37		

4. Title and Subtitle	5. Report Date
Traffic Flow Management Tools: Guidance for Use, Integration, and Training: Part Task Experiment 1	September 2017
	6. Performing Organization Code ANG-E25

7. Author(s)	8. Performing Organization Report No.
Brion Woroch, Ph,D., Engility Corporation, Inc. Carolina M. Zingale, Ph.D., FAA Human Factors Branch Anthony J. Masalonis, Ph.D., Spectrum Software Technology, Inc.	DOT/FAA/TC-TN17/37

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	
	11. Contract or Grant No.

Sent Bill an email 12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Federal Aviation Administration Human Factors Research and Engineering Group 800 Independence Avenue, S.W. Washington, DC 20591	Technical Note/Technical Report
	14. Sponsoring Agency Code ANG-C1

15. Supplementary Notes

16. Abstract

Objective: The purpose of this study is to develop a better understanding of human behavior when using the types of decision support tools (DSTs) planned for the Traffic Flow Management (TFM) domain and other applicable Air Traffic Control domains. **Background**: DSTs are typically not 100% accurate or reliable because they base decisions on probabilistic information, such as weather predictions. DSTs may provide one or more recommendations. User trust in automation and user workload can influence the extent to which users implement the suggested recommendations and how well the task is performed. **Method**: Sixteen volunteers from the FAA William J. Hughes Technical Center with no experience with TFM tools and procedures served as participants. We designed a task that could be quickly learned by novices and focused on several key aspects of the types of tasks performed by TFM personnel. We focused on four factors that might impact DST use; situation-specific training, DST reliability, the number of recommendations made by the DST, and overall task workload. **Results**: Some of the factors had direct impact on both objective measures of task performance and subjective measures. Several of the factors interacted in meaningful ways that illustrate the complex nature of DST use and provide insights and recommendations for DST development and deployment. **Conclusion**: We found that DST reliability and task workload played important roles in task performance. The interaction of the factors and training highlights a need to consider these multiple factors when developing and deploying DSTs in the operational environment. **Applications**: The complexities of training leave room for future research. We hope future research provides an opportunity to explore these topics in more depth.

17. Key Words	18. Distribution Statement
Decision Support Tools, Traffic Flow Management Air traffic Control	This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	65	

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

Figures **Page**

Tables **Page**

Acknowledgments

We thank Nelson Brown (ANG-E25), computer scientist, and Kevin Tracy, Brett Williams, and Louis Rivera of Hi-Tec Systems, who were responsible for developing the software and data-collection system used in this study. We also thank the participants from various organizations throughout the FAA William J. Hughes Technical Center for contributing to this effort.

THIS PAGE IS BLANK INTENTIONALLY.

Executive Summary

The purpose of this research is to develop a better understanding of human behavior when using the types of decision-support tools (DSTs) planned for the Traffic Flow Management (TFM) domain and other applicable air traffic control domains. These DSTs provide users with recommended solutions and methods to assess different options (i.e., "what-if" modeling capabilities) that allow users to evaluate the likely outcomes of different potential actions. This study used a part-task design to assess DST used by non-TFM personnel. We designed a task that could be quickly learned by novices, and we focused on several key aspects of the types of tasks performed by TFM personnel. To study DSTs, we focused on four experimental variables that likely impact DST use: situation-specific training, DST reliability, the number of recommendations made by the DST, and overall task workload.

The central task was inspired by the Integrated Departure Route Planning tool (DeLaura et al., 2012; Davison Reynolds & DeLaura, 2011) and involved rerouting aircraft around severe weather. The task involved assessing how multiple factors impact the aircraft and airspace. These factors included weather, aircraft congestion in the airspace under the participant's control, aircraft congestion in airspace under the neighboring center's control, and flight delay time. We assigned each factor a score to provide a numerical measurement of behavior. We weighted the factors to represent the relative importance of each parameter in the real world, as determined by the researchers (based on our knowledge of TFM) and by the TFM subject matter experts (SMEs). The participant's task was to select the highest scoring reroute. Participants made their rerouting decision based on information about each route using "what-if" modeling capabilities. In addition, we administered several questionnaires throughout the study to assess the participant's subjective attitudes and opinions about the DST.

The DST we developed for this task recommended one or more routes as a high-scoring route. We manipulated two key aspects of the DST: its reliability and the number of recommendations it made. This allowed us to assess how the experimental variables would impact DST use.

The participants also performed two secondary tasks designed to be similar to the types of tasks performed by traffic managers: an airspace monitoring task and a National Traffic Management Log communication task. These tasks allowed us to manipulate the workload of the participants so we could study the impact of high-workload environments on DST use. We also restricted the amount of time participants had to complete their tasks to create a realistic and busy situation.

Every participant underwent interactive training on how to perform the task and how to find and select optimal routes. They also completed several practice scenarios. Half of the participants received additional situation-specific training regarding the details of how the DST was developed and how it generated its recommendations. We were able to compare the participants who received the extra training with those who did not to see how such training impacted DST use.

Some of the experimental variables directly impacted task performance and survey measures. Several of the experimental variables interacted in meaningful ways that illustrate the complex nature of DST use, and provide insights and recommendations for DST development and deployment.

Both DST reliability and task workload had direct effects on performance. More reliable DSTs increased task performance. Higher workloads decreased performance and increased the participants' reliance on the DST. These two variables interacted, indicating that low-reliability had little impact on performance when task load was low. When the route recommendations provided by the automation were not optimal, there was an increase in the need to evaluate alternative routes.

When workload was low, the participants had the necessary time and cognitive resources to do that evaluation and overcome the poorer performance of the DST. When time and workload reduced the participant's ability to evaluate their options and rely on the automation, a poor performing DST had a negative impact on performance.

The extra situation-specific training we gave half the participants did not increase performance in the scenarios with no DST. However, it did change the way the participants performed the tasks in scenarios with the DST. Situation-specific training and task workload had an interactive effect on performance. When workload was low, the trained and untrained groups performed similarly. Having more information regarding how the DST arrived at decisions was not as helpful in situations in which the participants had time to consider alternative options. The training let the participants know the situations in which the DST could be trusted and was beneficial to performance in high-workload situations. We conclude that low workload allowed the participants the opportunity to evaluate their options so the extra training did not increase performance. But when workload was high, the additional training allowed the participants to make better use of the DST and outperform the untrained group.

Workload is relevant to DST use in operational environments. Situations in which DSTs would be most useful are inherently busy, such as those that require traffic reroutes. In our study, DST reliability became an important consideration, and the additional situation-specific training helped improve performance. Additional research is needed to know more precisely the levels of these experimental variables that would most impact performance in operational settings. Our study used only two levels of reliability and workload: low and high. It is unknown how reliable a DST needs to be, only that it is an important consideration. The workload of potential users should be considered during evaluation of DSTs.

In our study, simple additional training had an impact on DST use and task performance. However, the complexities of training leave room for future research. These include topics about the types of information covered by the training and issues surrounding how training is administered. We hope future research provides an opportunity to explore these topics in more depth.

We conducted this study with novices. Next, it will be important to conduct a similar study with TFM personnel with a range of experience levels to determine whether the same results are found. It is likely that experienced TFM personnel and less experienced TFM personnel differ from one another in how they make use of the DST and the type of training that is most effective. DST training for more experienced users may need to be targeted more specifically to help them determine where benefits from the automation can be gained (e.g., by providing information as to when the tool provides a faster resolution, or identifying and providing solutions to situations that these users encounter less frequently).

In conclusion, we set out to evaluate DST use in an environment that emulated many of the demands in TFM. We designed a task that could be quickly learned and performed by non-TFM personnel that was similar to actual TFM responsibilities. We found that DST reliability and task workload played important roles in task performance. The interaction of these two experimental variables, as well as the amount and type of training provided, highlights a need to consider these issues when designing, developing, deploying, and evaluating DSTs in the operational environment.

THIS PAGE IS BLANK INTENTIONALLY.

1. INTRODUCTION

The purpose of this research is to investigate how users employ new decision-support tools (DSTs) planned for the Traffic Flow Management (TFM) domain and other applicable air traffic control (ATC) domains. These DSTs recommend solutions and methods to assess different options (i.e., "what-if" modeling capabilities) that allow users to evaluate the likely outcomes of different potential actions.

The decisions made by TFM personnel—Traffic Management Coordinators (TMCs) in Air Route Traffic Control Centers (ARTCCs), Terminal Radar Approach Control (TRACON) facilities and air traffic control towers (ATCTs), and Traffic Management Specialists (TMS) at the Air Traffic Control System Command Center (ATCSCC)—are complex and involve multiple factors that affect cognitive workload. DSTs are intended to reduce the cognitive workload of users in these domains. However, various factors must be considered in implementing the tools to ensure that they bring their expected benefits. These factors include tool design and user training.

This report summarizes a part-task study examining how DSTs affect performance of novice, non-operational personnel on an aircraft departure rerouting task. We based the part-task study on information obtained from a literature review that we summarized in an annotated bibliography (Masalonis, Zingale, & Puzen, 2016) and on our review of air traffic management (ATM) tools and concepts that are expected to be implemented in the NextGen timeframe or that have been researched for possible implementation (Masalonis, Zingale, Puzen, Thomas, & Yuditsky, 2016).

DSTs are typically not 100% accurate or reliable because they base decisions on probabilistic information, such as weather predictions. Research has shown, however, that the tool does not need to be perfect to be useful. For example, "imperfect" alerts, as long as they are at least 70-75% accurate, are useful when workload levels and task demands are high (Dixon & Wickens, 2006). Sorkin, Kantowitz, and Kantowitz (1988) reported that the use of the alerts improves when users see information about the likelihood of an event (e.g., "possible signal," "likely signal," or "urgent signal") because the likelihood information helps the user better assess the situation to make the appropriate decision. Therefore, instructing users as to the reliability or accuracy of the tool recommendations is an important factor in evaluating DST usefulness.

Although qualitative DST reliability information has been shown to be useful, presenting quantitative information about the tool's reliability has not. Wiegmann (2002), in a laboratory study of variable-reliability automation, found that some participants used a strategy of agreeing with the automated recommendations at a rate approximately equal to the tool's purported reliability (e.g., 80%). This misunderstanding of probability led to poor performance; if participants attempted to agree with the automation on 80% of trials, they failed to understand that this strategy was likely to result in overall performance well below 80%. The participants could only be 80% correct if they agreed with the DST on the "right" trials. To achieve a performance level of 80%, the participants should rely on the automation all the time. Although this result is counterintuitive, presenting more precise, quantitative reliability information may lead to a reduction in DST usefulness and task performance. These findings led us to include qualitative DST reliability information rather than quantitative information in our study.

The usefulness of DST information also depends upon the user's trust in the automation and general tendency toward complacency. However, users' trust in automation does not always lead them to use it (Masalonis & Pararsuraman, 1999). A user who trusts a system's automation may opt not to use it when workload is low (to prevent boredom) or to sustain vigilance on the task. Alternatively, during demanding tasks, the user may decide to use the automation to reduce workload if the tool

meets an acceptable level of accuracy, safety, etc. To address this issue in the part-task study, we incorporated different levels of workload in the experimental task to investigate whether users were more likely to rely on automation under high-workload conditions even if the automation was less reliable in some situations.

Trust in automation has been shown to be difficult to build but easy to break down (Masalonis and Parasuraman, 1999). Once users lose trust, they may deactivate the automation and stop using it altogether. This means that the automation never has the opportunity to prove itself. For this reason, we decided not to give the participants in the part-task study the opportunity to turn off the automation. We did not want to influence trust levels in unplanned ways. Instead, our part-task study manipulated the tool's reliability by incorporating two different DST algorithms, and we trained half of the users about which algorithm was more reliable and which was less reliable under specified conditions.

Other studies have examined the degree of automation reliability required for performance improvement (see reviews by Wickens & Dixon, 2007; Rein, Masalonis, Messina, & Willems, 2013), but Trapsilawati, Qu, Wickens, & Chen (2015) conducted one of the few studies we found that explicitly compared automation reliability levels for DSTs that provide recommendations. Although Trapsilawati et al. (2015) found that reliable automation for air traffic conflict resolution advisories resulted in better performance than "unreliable" automation, even the "unreliable" automation was successful at resolving the conflict and avoiding the creation of new conflicts 80% of the time. Furthermore, both the reliable and unreliable automation led to better performance than no automation.

The number of recommendations a DST provides may vary. Sheridan and Verplank (1978) conducted seminal work on this topic, categorizing 4 levels of automation support: Level 1, in which the human operator determines the decision; Level 2, in which the computer helps determine the decision; Level 3, in which the computer helps determine the decision and suggests options to the user; and Level 4, in which the computer calculates the optimal decision that the user can choose to implement. The DSTs currently in use in today's air traffic environment or under consideration for future operational implementation are generally classified as Level 3 because they provide multiple alternative suggestions for action. DSTs that rank the suggested solutions, such as the Conflict Resolution Advisor (CRA), are also under consideration (e.g., Trapsilawati et al., 2015). This tool comes close to Level 4 automation because it includes a highest-ranked solution. However, there is still no "pure" Level 4 automation in today's air traffic environment. Therefore, it is appropriate to research the effects of presenting a single-decision choice before widespread implementation of such a capability is implemented. We investigated this type of automation, as well as Level 3, in this part-task study.

In summary, the part-task study we conducted examined different levels of DST reliability on an aircraft rerouting task under different levels of workload. We trained one half of the participants on the logic of the DST algorithms and the conditions in which the algorithms would be more or less reliable. We provided aircraft rerouting trials with no recommendations (in which participants would have to use "what if" capabilities to evaluate options), a single recommendation, or three recommendations. We expect the results of this part-task study to be useful in making recommendations for the development of future DSTs and the development of DST training materials.

2. METHOD

We designed this part-task experiment to examine the use of DSTs in TFM. We designed a primary task involving departure rerouting of aircraft around severe weather. Participants also performed two secondary distraction tasks to increase their workload. The entire experiment took approximately 2.5 hours.

2.1 Participants

Sixteen volunteers (11 males, 5 females) from the FAA William J. Hughes Technical Center (WJHTC) with no experience with TFM tools and procedures served as participants in this study. The age of the participants ranged from 25–52 years old (M = 40.75, SD = 10.57). Half of the participants were assigned to a training group and received the situation-specific training (SST). Many of the participants in this study were research personnel in the Research Development and Human Factors Laboratory (RDHFL). To avoid any expertise-based bias, equal numbers of these participants were assigned to each group (three in each). All participants read and signed an Informed Consent Statement (Appendix A) that summarized their rights and responsibilities before participating.

2.2 Materials

2.2.1 Demographics Questionnaire

The participants completed the demographics questionnaire (Appendix B) before beginning the experiment. It included questions about age, gender, and TFM or ATC experience. We also confirmed that participants had not received information from prior participants regarding the SST before we began the experiment.

2.2.2 Complacency Rating Scale

Before the participant performed the experimental tasks, we administered a survey to assess individual differences in attitudes toward automation and susceptibility to overreliance on automation. The instrument was based on the Complacency-Potential Rating Scale (CPRS) (Singh, Molloy, & Parasuraman, 1993). We made adjustments to two items because of the scale's outdated technology references, such as an item about VCRs. We made edits to questions #7 and #20 to update the items pertaining to television and medical devices. The version used in this study is Appendix C.

2.2.3 Questionnaires

We included several questionnaires during the experiment; after each reroute, after each scenario, and at the end of all scenarios. We provide screen shots of those surveys in Appendix D.

2.2.4 Primary Rerouting Task

The primary task for the participants was a departure rerouting task we designed based on the Integrated Departure Route Planning tool (DeLaura et al., 2012; Davison Reynolds & DeLaura, 2011). The task consisted of 10 experimental scenarios, each requiring the participant to reroute several aircraft around a weather event from one airport to another in fictional airspace. Although the task was based on tools a traffic manager might use, both the number of parameters used and the task itself were greatly simplified, so a participant with no TFM experience could learn and practice the task as well as complete the 10 experimental scenarios in 2–2.5 hours.

A screenshot of the rerouting task is shown in Figure 1. It contains a map of all the routes and a list of aircraft to be routed through that airspace. During each scenario, five aircraft had their originally filed route blocked by severe weather (e.g., Route E) and had to be rerouted to a different route. The airports, route names, waypoint labels, and direction of flight varied between scenarios but stayed the same for all five flights throughout a single scenario. Table 1 shows all of the parameters that varied between scenarios. Every scenario began at 1900 hours, and each flight was scheduled to depart 15 minutes after the previous flight (1900, 1915, 1930, 1945, and 2000).

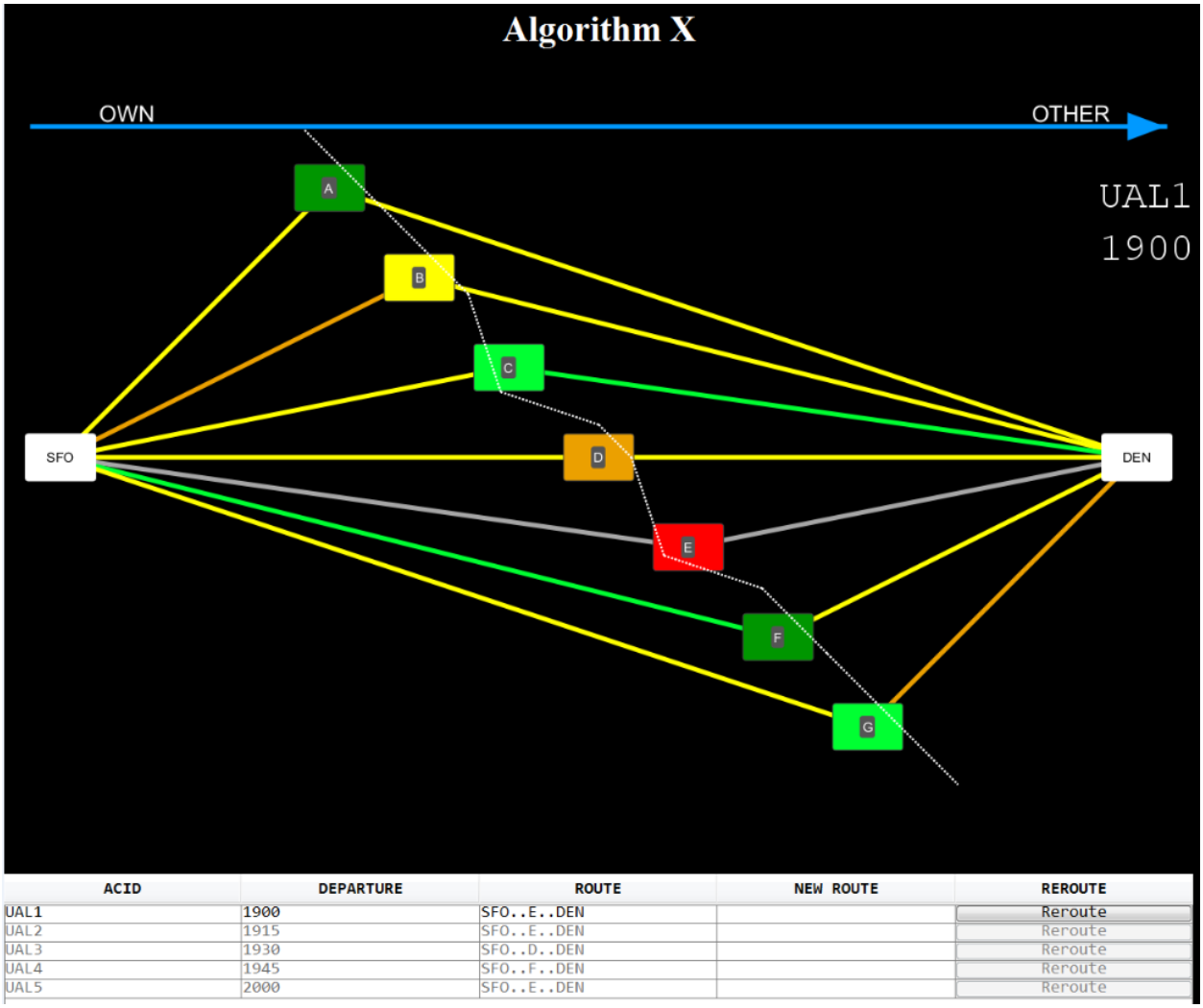

Figure 1. A screenshot of the primary aircraft rerouting task.

Table 1 Scenario variables

Airports	Traffic Direction	Aircraft ID (ACID)	Waypoint Labels	Number of Scenarios
SFO & DEN	Eastbound	UAL 1-5	A through G	5 (5 flights in each)
JFK & ORD	Westbound	JBU 1-5	H through N	5 (5 flights in each)

Each scenario featured seven routes from the departure airport to the destination airport. Each route passed through a single waypoint. At the start of each flight, the originally filed route was closed because of severe weather. For example, in Figure 1, UAL1 was originally filed as SFO..E..DEN. The participant had to choose an alternate route for the flight from among one of the six remaining routes. Each route was assigned a score. The score was based on four parameters that varied for each flight: weather, flight delay time, congestion level in the participant's center, and congestion level in the neighboring center. Each parameter was assigned a numeric value that was weighted based on their relative importance, as determined by the experimenters, and combined to yield a final score for each route. Figure 2 shows the route table that displayed the parameters for each route at the departure time of that aircraft. These numbers combined to yield a score out of 1000 possible points for each reroute. In addition, a color scale indicated the relative severity (i.e., greatest negative impact to route) of that parameter. The colors used in this task (from lowest to highest) were: light green, dark green, yellow, and orange (Table 2). Red and gray indicated a closed route. The R/G/B values used can be found in Appendix E.

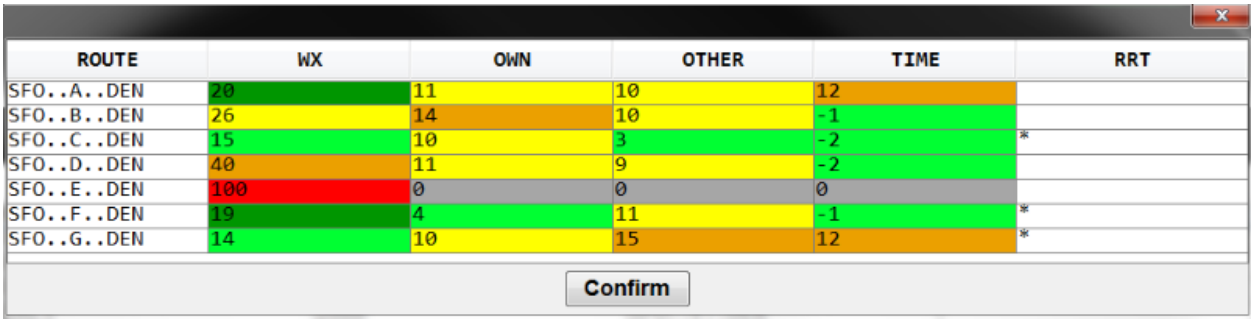

Figure 2. The route table displaying scoring parameters.

Table 2 Route parameter values

	Weather	Own Center	Other Center	Delay Time
Gray	0	0	0	0
Light Green	1-15	1-4	1-4	(-10)-1
Dark Green	16-20	5-7	5-7	1-5
Yellow	21-35	8-11	8-11	6-10
Orange	36-50	12-15	12-15	11-20
Red	>50	N/A	N/A	N/A

Weather:

The originally filed route for each aircraft was blocked because of severe weather. This was represented visually on the route map (Figure 1) with a red box around the waypoint label. This was represented in the route table (Figure 2) with a red cell and "100" parameter value. The weather parameter for the other available routes varied from 1–50, with a higher number indicating more severe weather. The weather condition predicted for each route was indicated by the color of the waypoint on the map and the color of the cells in the route table. We instructed participants that all

colors except red represented weather that was safe to fly in and that the lower the weather value the higher the score for that route.

Own & Other Center Congestion:

The airspace on the route map (Figure 1) was separated into two halves: the airspace under the control of the participants' OWN center and the airspace under the control of the OTHER center. In scenarios in which traffic was eastbound from SFO to DEN, the airspace containing SFO and the routes to the left of the waypoint were OWN, and the rest of the airspace was OTHER. In scenarios in which traffic was westbound from JFK to ORD, the airspace containing JFK and the routes to the right of the waypoints were OWN, and the rest of the airspace was OTHER. The OWN and OTHER labels were shown at the top of the screen throughout the scenario.

The aircraft congestion level along the routes was indicated by the color of the route line and the color of the corresponding cells in the route table. We told the participants that the number in the route table indicated the number of aircraft forecasted along the route. Rerouting onto a lower number (less congested) route yielded more points.

Delay Time:

The delay time parameter was the number of minutes a flight would be delayed by choosing that route. This value varied from -10 to 20. A negative number meant a route was faster than the originally filed route and was a saving time. This parameter was not represented visually on the route map; it was only available in the route table. A faster flight time (less delay) yielded a higher score.

Score calculation:

We developed the following formula to combine the individual parameters into a score for that route:

((70-Weather)*8.1)+((20-Own Center)*14)+((20-Other Center)*9)+((20-Time)*7.4)

The factors were weighted to represent the relative importance of each parameter, as determined by the researchers based on our knowledge of TFM and discussions with TFM SMEs to reflect the relative importance of these factors in the real world. We identified weather as the most important factor. A low value on the weather parameter contributed more to the score on that route than any other parameter. The second most important factor was the congestion level in the participant's OWN airspace. Congestion levels in the OTHER airspace and flight delay contributed about equally to the score and were weighted the lowest.

We instructed the participants that the overall goal of the task was to choose the highest-scoring route to obtain the highest possible score in each scenario. We trained the participants on the parameters and how the score was calculated. After each scenario, participants received feedback regarding their choices and score.

2.2.4.1 Task sequence and "What-iffing" capabilities

The participants had 4 minutes to reroute the five aircraft in each scenario. The route table (Figure 2) started as an empty table with no information except the indication that the originally filed route closed due to severe weather. The participants could gather information about other routes by selecting them with the mouse, a procedure we referred to as "what-iffing" because it allowed the participants to see the parameters associated with that route option. The alternate routes indicated the colors of the route lines and waypoints on the route map (Figure 1). Choosing to "what-if" a route option was followed by a 3-second delay before the information appeared in the route table. This

delay was added to mimic potential computer-processing time required by some tools to model information. The delay was also added to discourage participants from "what-iffing" all routes for all flights in every scenario. Because of the time pressure to complete the scenario within the allotted 4 minutes, the delay introduced a cost to acquiring more information, and we hoped to encourage a more selective use of this feature.

If all five aircraft were not rerouted within 4 minutes, the participants received a 500-point score penalty and an additional 10-point penalty for every 10 seconds over 4 minutes. We did this to encourage the participants to finish the scenario within 4 minutes. However, we still wanted to gather any rerouting choices for flights rerouted after the 4 minutes expired, so the scenario continued to run until the participants completed all of the reroutes.

2.2.4.2 Route Recommendation Tool

Most of the scenarios (8 of 10) had routes suggested by the Route Recommendation Tool (RRT). A suggested route had the parameter information for that route automatically displayed when the rerouting for that flight began. In addition, an asterisk was placed, in that route's row, in the RRT column of the routing table to indicate that it was a "recommended" route. Four of the 10 scenarios had one route automatically recommended for each flight; four scenarios had three routes recommended for each flight. The two remaining scenarios had no automation suggestions. Figure 2 shows a route table with three routes recommended by the RRT and all other routes "what-iffed."

The RRT varied in the quality of its recommendations; sometimes, it indicated the highest-scoring route available, and other times it did not. Therefore, the RRT varied whether it would reliably recommend a good route. We systematically assigned which route was recommended, allowing the creation of two types of scenarios: low-reliability scenarios and high-reliability scenarios. In a high-reliability scenario, the recommended routes were likely to include the highest-scoring route. In the low-reliability scenarios, the recommended route never included the highest-scoring route. Table 3 shows the rank of the recommended route's score (1 = highest-scoring route).

Table 3 Rank of the recommended route's score (out of six).

Number of Recommended Routes	Low Reliability	High Reliability
1	2nd, 2nd, 3rd, 3rd, 3rd	1st, 1st, 1st ,1st, 2nd
3	2-4th, 2-4th, 3-5th, 3-5th, 3-5th	1-3rd, 1-3rd, 1-3rd, 1-3rd, 2-4th

2.2.5 Secondary Tasks

The participants also performed two secondary tasks—a monitoring task and a communicating task—inspired by the types of tasks performed by traffic managers. There are many aspects to a traffic manager's job and a variety of tasks they need to perform. These secondary tasks were added to simulate some of the demands on the traffic manager's time and attention so we could perform our evaluation of DSTs in more operationally realistic circumstances.

We varied the frequency of the secondary tasks during the scenario to vary the workload of the participants. In low-workload scenarios, each secondary task required four responses, whereas in the high-workload scenarios, each task required 10 responses.

2.2.5.1 National Airspace Monitoring Task

This secondary task required participants to monitor and report congestion values from a simulated National Airspace System (NAS) monitor shown in Figure 3. At the start of each scenario, all of the cells were randomly set to green or yellow. Periodically throughout the 4 minutes of rerouting, a cell would change from green or yellow to red (similar to what a traffic manager would see if the Monitor Alert Parameter (MAP) value was exceeded). When a cell turned red, participants were to select the red cell, revealing the predicted aircraft count (i.e., 18–24) in that sector. Next, the participants had to report these three values: sector, time, and aircraft count by typing the values into three text boxes. In Figure 3, the goal was to select the red cell, enter 14 under "Sector," enter 2130 under "Time," and the "Sector Count" number from that cell (not shown), and press the "Submit" button. The cell remained red for the duration of that scenario. The sector count remained visible to indicate which cell had already been selected.

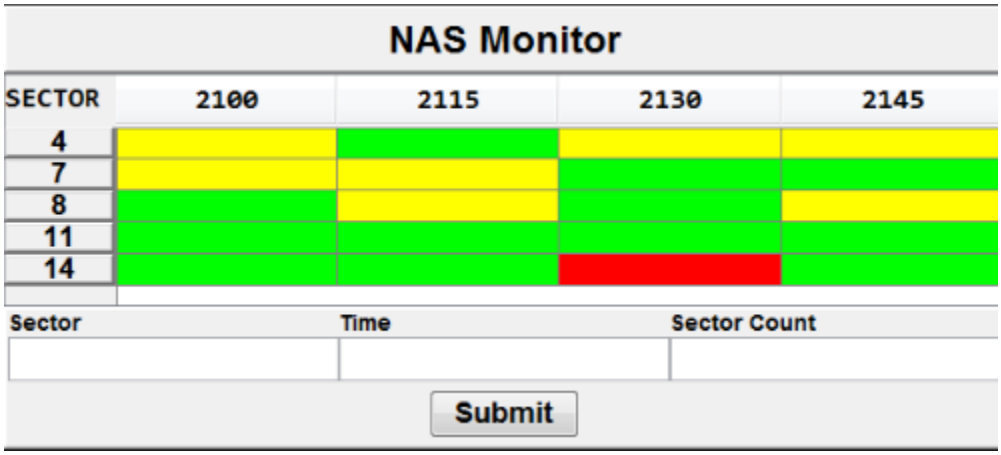

Figure 3. The National Airspace System monitoring interface.

We were concerned that participants would not pay enough attention to the secondary tasks and only focus on the primary rerouting task. To encourage vigilance on the NAS monitoring task, participants were awarded or penalized points based upon the speed and accuracy of their submissions:

- A correct response within 20 seconds of the cell turning red (+50)
- A correct response, but after 20 seconds (-50)
- An incorrect response (-100)
- No response before scenario end (-100)

Because there could be multiple cells turning red at one time, the task involved a working memory component (i.e., participants had to remember which of the red cells they were intending to respond to as they completed their submission). This led to an uncertainty regarding which cell was the intended target of an incorrect response, because a submission could contain elements belonging to two or three different red cells. Therefore, we knew whether the participants responded to any given cell correctly (all three submission fields matched a currently red cell), but we could not definitively distinguish between errors of omission (i.e., no submission) versus error of commission (i.e., a submission of incorrect information).

2.2.5.2 National Traffic Management Log Communicating Task

This task required participants to monitor the simulated National Traffic Management Log (NTML) window for new messages (Figure 4). Each message consisted of the time of the message, the message content, and the sender's initials. The message content was based on content of actual messages that may be seen by traffic managers in the NTML. Each message had to be either "Forwarded" to their supervisor or simply "Acknowledged" (ACK button in Figure 4). If the message made a reference to either airport involved in the scenario (e.g., JFK/ORD or SFO/DEN), it was to be forwarded. If it did not, it was to be acknowledged. To avoid confusion, no messages in the currently active scenario contained references to the airports used in other scenarios (e.g., in a JFK/ORD scenario, no messages contained SFO or DEN). To incentivize quick and accurate responses, the following points were awarded or deducted:

- A correct response within 10 seconds of message appearing (+50)
- A correct response, but after 10 seconds (-50)
- An incorrect response (-100)
- No response before scenario end (-200)

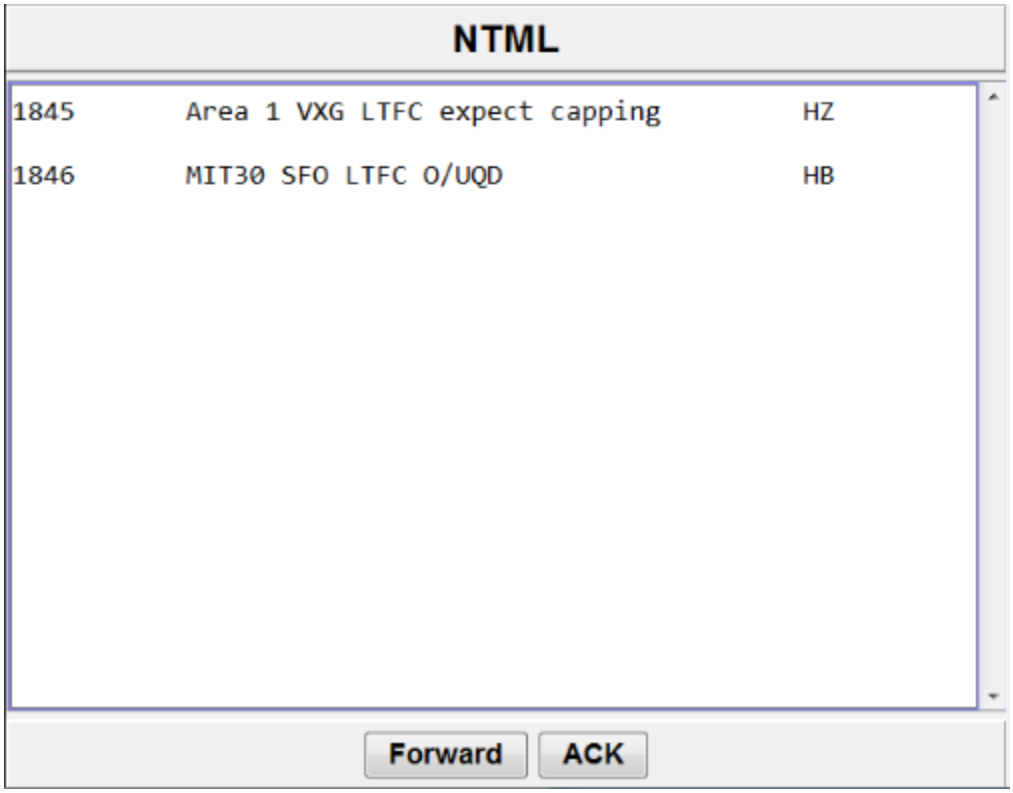

Figure 4. The National Traffic Management Log communicating interface

2.2.6 Feedback

At the conclusion of each scenario, and after completing a short questionnaire, the participants received feedback on their performance in the scenario they just completed. The goal of the feedback was to show the participants how well they did in relation to the automation. The scores for each

route for each flight were displayed on the screen. The route that the participant chose was highlighted in yellow. Any routes that would have yielded a higher score than the one chosen were also partially highlighted. All routes that were recommended by the automation were also indicated. The participants could quickly learn if their rerouting choices were optimal, how good the automation's route recommendations were, and if their own choices were better or worse than the automation's suggestions.

The participants were also shown any points accumulated or lost based on their performance on the secondary NAS monitoring and NTML communication tasks. We hoped that displaying the points earned would motivate the participants to be efficient and effective in performing these secondary tasks.

2.2.7 Training

Prior to starting the experimental scenarios, all the participants completed a training session that instructed them on how to perform the task and obtain a high score. They also completed practice scenarios to gain experience working on the tasks and to mitigate learning effects in the first few experimental scenarios. The training consisted of a slideshow and interactive demo using the experiment software (the demonstration scenarios were simply modified experimental scenarios with no time component). The slides described the goals of each task and how to perform all elements. After reading several slides, the participants would perform an action or complete an element of the task in the demonstration scenario before reading several more slides in the training slideshow. The first demonstration scenario had no automation-suggested routes and guided the participant through the primary rerouting task and both of the secondary tasks. This was followed with detailed feedback, demonstrating how they performed as well as a verbal description on how to achieve a higher score. The participants then completed one practice scenario, which was similar to an experimental scenario with high workload and no automation, followed by another detailed feedback session.

The participants then received a second slideshow about the RRT. All participants were told that the RRT automation's ability to reliably recommend a high-scoring route would vary from scenario to scenario. However, half of the participants received an additional set of three slides that revealed how to tell when an upcoming scenario would be reliable or not—the basic description of the SST that would differentiate the two participant groups. There were two factors that determined whether reliability would be high or low: direction of flight and algorithm. Each scenario featured eastbound or westbound flights. The RRT made recommendations using either Algorithm X or Algorithm Y. The algorithm names were randomly assigned labels. The SST participants were told:

> RRT Algorithm X was developed in consultation with United Airlines (UAL) dispatchers and with operational personnel at the Centers through which UAL's eastbound flights from SFO to DEN operate. It does better using the preferences of UAL and of controllers near SFO & DEN airspace to decide which combination of weather, congestion, and delay times will result in a high-scoring reroute" and "RRT Algorithm Y was developed in consultation with JetBlue Airways (JBU) dispatchers and with operational personnel at the Centers through which JBU's westbound flights from JFK to ORD operate. It does better using the preferences of JBU and of controllers near JFK & ORD airspace to decide which combination of weather, congestion, and delay times will result in a high-scoring reroute.

A summary of these reliability factors is shown in Figure 5, which was shown to the participants who received the extra training. A printout of the figure was given to them as a reference card to help remind them of the relationship. This information was critical to predicting whether the automation would be reliable or not, because the direction of traffic and algorithm to be used were shown to the

participants before they began the scenario. This information was meaningful to the trained participants, but was not meaningful to the participants who were not trained about it. Both algorithms and directions of flight were associated with both high and low reliability (e.g., algorithm X was reliable in some scenarios and unreliable in others). It was a complicated relationship and the participants not receiving the additional training could not easily deduce the relationship between algorithm, direction of flight, and RRT reliability.

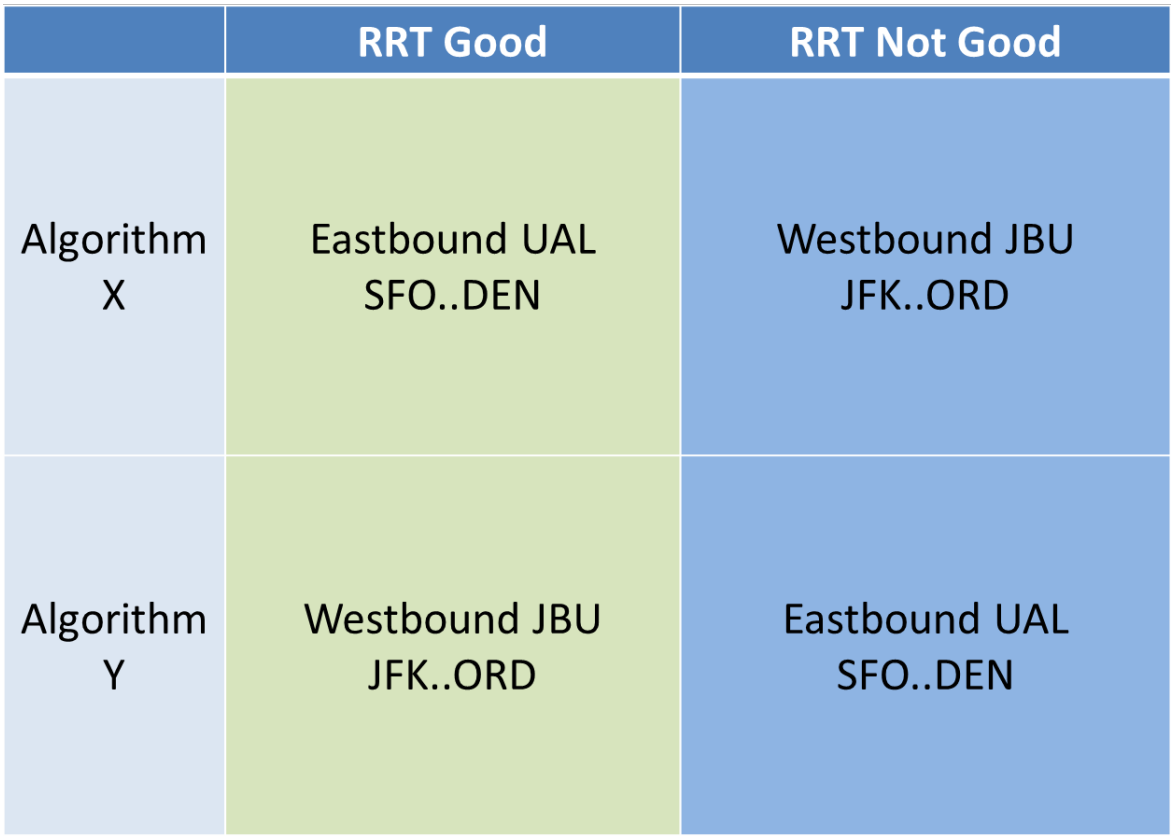

Figure 5. A summary slide of the SST for the RRT showing when the automation's recommendations would be reliable or not.

This phase of training was accompanied by a demonstration scenario and followed by two practice scenarios. The demonstration scenario had highly reliable automation recommendations. The first practice scenario had low-reliable automation and the second was high. The participants who did not receive the extra training were told we were using Algorithm P (a blend of X and Y) for the practice scenarios. We did this to make it harder for the untrained participants to figure out the pattern of reliability. The practice scenarios were similar to the experimental ones with one difference: during an experimental scenario, the RRT always made the same number of recommendations (one or three) for each of the five flights. However, during the practice scenarios, the number of recommendations alternated between one and three for demonstration purposes.

2.3 Procedure & Experimental Design

We included four experimental parameters or independent variables (IVs). We manipulated these variables in our experiment to see the effect they would have on participants' performance of the tasks, known as the dependent variables. There were ten experimental scenarios. The IVs were counterbalanced so that they co-occurred equally. Instead of randomizing condition orders or systematically varying them in a Latin square or similar design, we explicitly specified certain orders for the combinations of independent variables in a way that attempted to avoid overly biasing any participant into a certain initial attitude about the automation or the task. More details regarding the scenario development and ordering are in Appendix E.

2.3.1 Situation-Specific Training

The details of the training are outlined in section 2.2.6. Eight of our 16 participants received the supplementary SST training, whereas the other eight did not. The two groups (SST or No-SST) were matched on the basis of gender, age, and human factors research experience. All 16 of the participants performed the same scenarios, so this was a between-subjects variable.

2.3.2 Workload

The participants' workload was manipulated by the number of events in the secondary tasks during a scenario. There were two levels of workload in this study. In the low-workload scenarios, four NAS monitor cells turned red and four NTML messages appeared throughout the scenario. In the high-workload scenarios, 10 of each secondary event occurred. The two levels were informed by pilot testing with developmental personnel and other researchers. Of the 10 scenarios, five were low workload, and five were high workload.

2.3.3 The Number or Recommendations

Each scenario had zero, one, or three recommended routes for each of the five flights. We varied the number of recommendations (NREC) across scenarios. Four scenarios had one recommendation, four had three recommendations, and two scenarios had no recommendations. This variable allowed us to address the question as to what level of automation (e.g., single best, several options) is associated with higher performance.

2.3.4 Automation Reliability

There were two levels of reliability in this study. The ranks of the recommended routes (out of six) are shown previously in table 4. There were four low-reliability scenarios and four high-reliability scenarios. The two scenarios that had zero automation recommendations could not be reliable or unreliable, by definition. The combination of all variables is shown in Table 4.

Table 4 The ten experimental scenario types performed by all participants.

Scenario:	1	2	3	4	5	6	7	8	9	10
Workload	Low	Low	Low	Low	Low	High	High	High	High	High
NREC	1	1	3	3	0	0	1	1	3	3
Reliability	Low	High	Low	High	N/A	N/A	Low	High	Low	High

3. RESULTS

We analyzed the data to evaluate the effects of the independent variables (IVs) on the various dependent variables. The IVs were: Stimulus-Specific Training Group (SST versus no-SST), Workload (low versus high), Automation Reliability (low versus high), and NREC (one versus three). We conducted three principle types of data analyses and used an alpha level of 0.05 to determine statistical significance. The first analyses were T-tests between the training groups (SST versus no-SST). We tested demographic factors and CPRS scores to determine whether any pre-experiment differences were evident between the participants who were assigned to the SST group and the no-SST group. The other two analyses involved a mixed-model multiple analysis of variance (ANOVA) between group (SST versus no-SST) factor and repeated-measures factors representing the relevant independent variables. We did this to assess the impacts of all IV in the same statistical model rather than perform tests of each IV individually and inflating our alpha rate. We used the mixed model to combine a categorical variable (training group membership) with the repeated-measure factors performed by all participants.

We analyzed the behavioral effects of the IVs using a 2x2x2x2 (SST group x NREC x workload x reliability) ANOVA. The dependent variables we analyzed were the scores on the rerouting task, the response times to the secondary tasks, and responses to the survey questions. The rerouting score was the total score obtained by rerouting all five flights in a scenario. The same analyses done on the raw scores reported below were also performed using the proportion of the score obtained for a scenario out of the maximum possible score for that scenario. The results from those two sets of tests did not differ, so only the raw scores are reported.

3.1 Between-Group Demographics and Complacency Rating Scale

We attempted to match the age and experience of the SST and no-SST groups to minimize demographic differences between them. We assigned participants to a group based on age, gender, TFM experience, ATC experience, and whether or not they worked at the RDHFL. We did this so that any differences in performance we observed between the training groups could be attributed to the training manipulation and not to a pre-experimental difference. Table 5 summarizes the age and genders of the participants assigned to each group. The SST and no-SST groups did not differ in age [$t(14) = 0.70, p = 0.497$], and the groups were matched on the other factors.

Table 5 Count of gender and Mean age (standard deviation) in the SST groups.

	Gender	**Mean (*SD*) Age**
SST, N=8	F=3, M=5	38.87 *(8.84)*
No SST, N=8	F=2, M=6	42.62 *(12.39)*

We also found no difference in the Complacency Rating Scale scores between the SST ($M = 62.75$, $SD = 5.80$) and no-SST ($M = 65.13$, $SD = 6.29$) groups, [$t(14) = 0.78$, $p = 0.445$]. Therefore, we are further confident that any observed difference in performance between the SST group and no-SST group was due to the experimental training manipulation and not due to pre-experimental differences between the groups.

3.2 Primary Rerouting Task

The primary dependent variable of interest was the total score from the rerouting task. We obtained one total rerouting score per scenario by adding the scores from the five flights.

Two of ten scenarios had no automation recommendations, one of each workload condition. We used a 2x2 (group x workload) ANOVA to analyze the score for those scenarios. There was no difference in score between the SST and no-SST groups in the no-automation conditions (Figure 6), [$F(1,14) = 0.01, p = 0.942$], again indicating that the two groups did not differ from one another before any experimental variables were introduced.

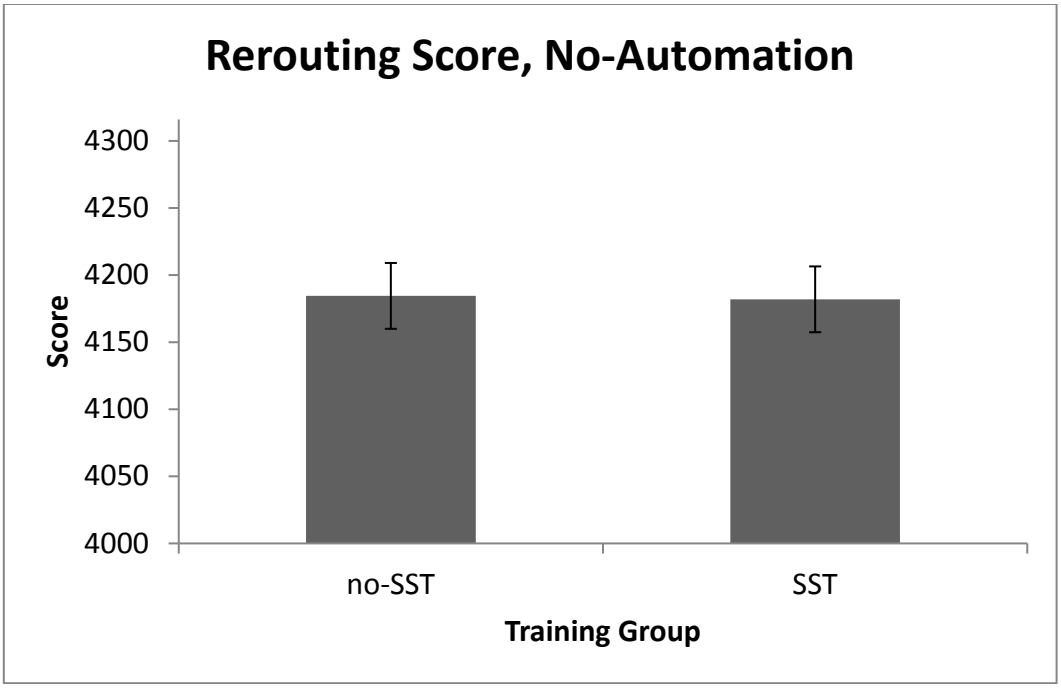

Figure 6. The average rerouting scores in the no-automation condition. The error bars represent the standard error of the mean.

Eight scenarios for each participant featured automated recommendations. The five flights in each scenario were summed and submitted to a 2x2x2x2 ANOVA to test the effects of each independent variable. We present only the results significant at $p < 0.05$. There were no three- or four-way interactions. There were several significant two-way interactions. The interactions between the independent variables offer insight into how the variables influence the use of DSTs. We conducted post-hoc T-tests on the means for the condition in the interactions to evaluate differences. Participants in the SST group did not score statistically higher than the no-SST group overall. This is due to a Training Group x Workload interaction [$F(1,14) = 13.29, p = 0.003$]. At low workload, the two groups performed at the same level [$t(14) = 0.39, p = 0.703$]. However, at high workload, the SST group outperformed the No-SST group [$t(14) = 2.99, p = 0.010$] (Figure 7).

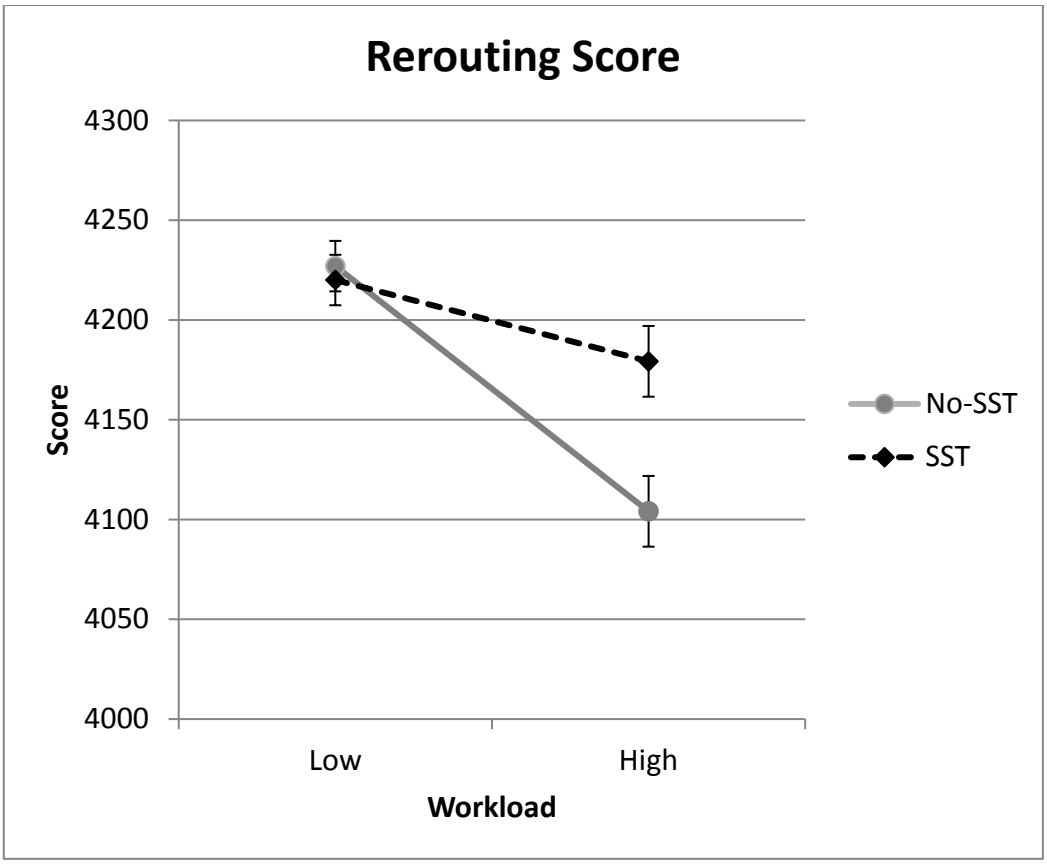

Figure 7. Interaction of Workload and SST on rerouting performance. The error bars represent the standard error of the mean.

We found an interaction between DST reliability and the number of recommendations it made [$F(1,14) = 11.07$, $p = 0.005$] (Figure 8). During scenarios when the automation was less reliable, one and three recommendation yielded similar performance [$t(15) = 1.16$, $p = 0.264$]. However, when the automation was more reliable, we found that performance was better when only one recommendation was provided, [$t(14) = 2.26$, $p = 0.039$].

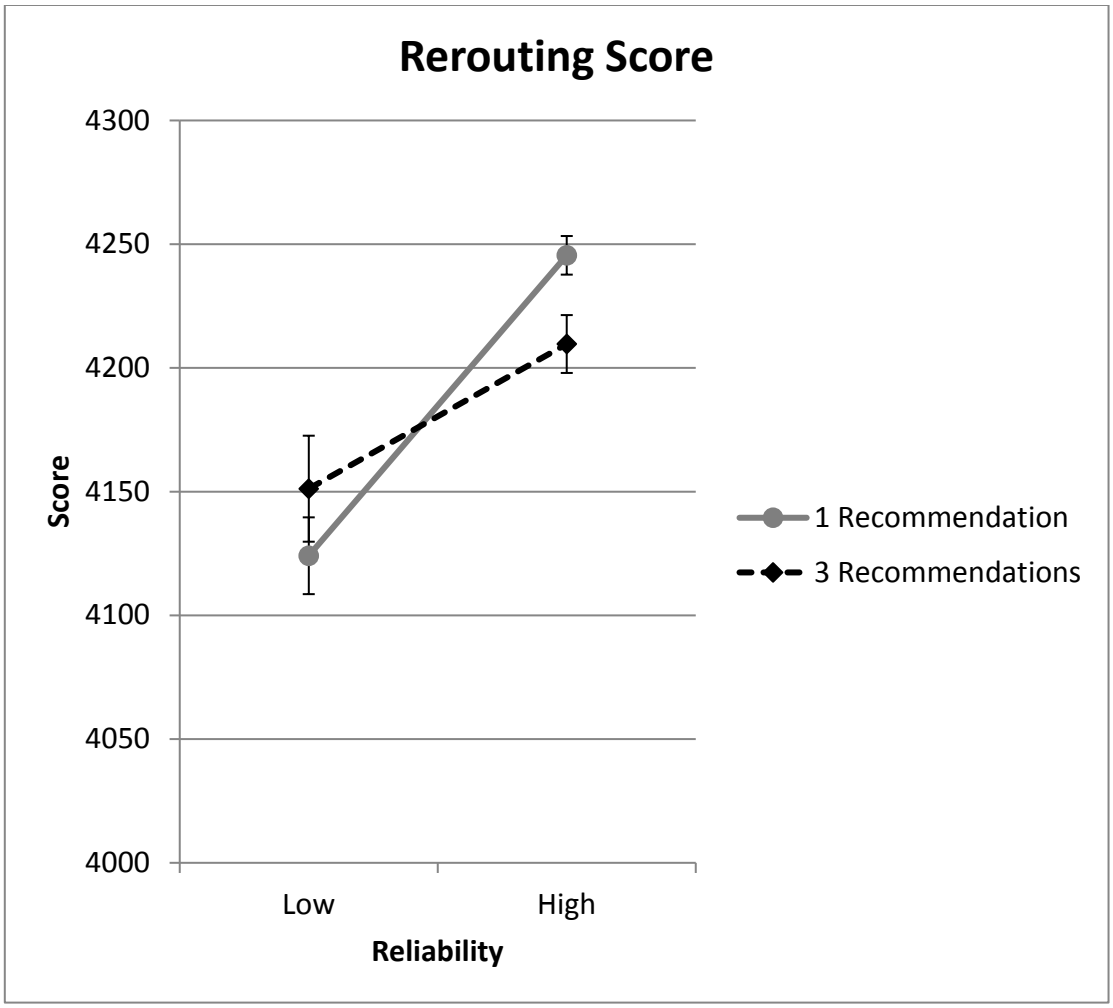

Figure 8. Interaction of automation reliability and number of automation recommendations on rerouting performance. The error bars represent the standard error of the mean.

There was an interactive effect of workload and DST reliability [$F(1,14) = 7.24$, $p = 0.018$] (Figure 9). When workload was high and automation reliability was low, performance was at its lowest When workload was low, low automation reliability had less of a negative impact on performance.

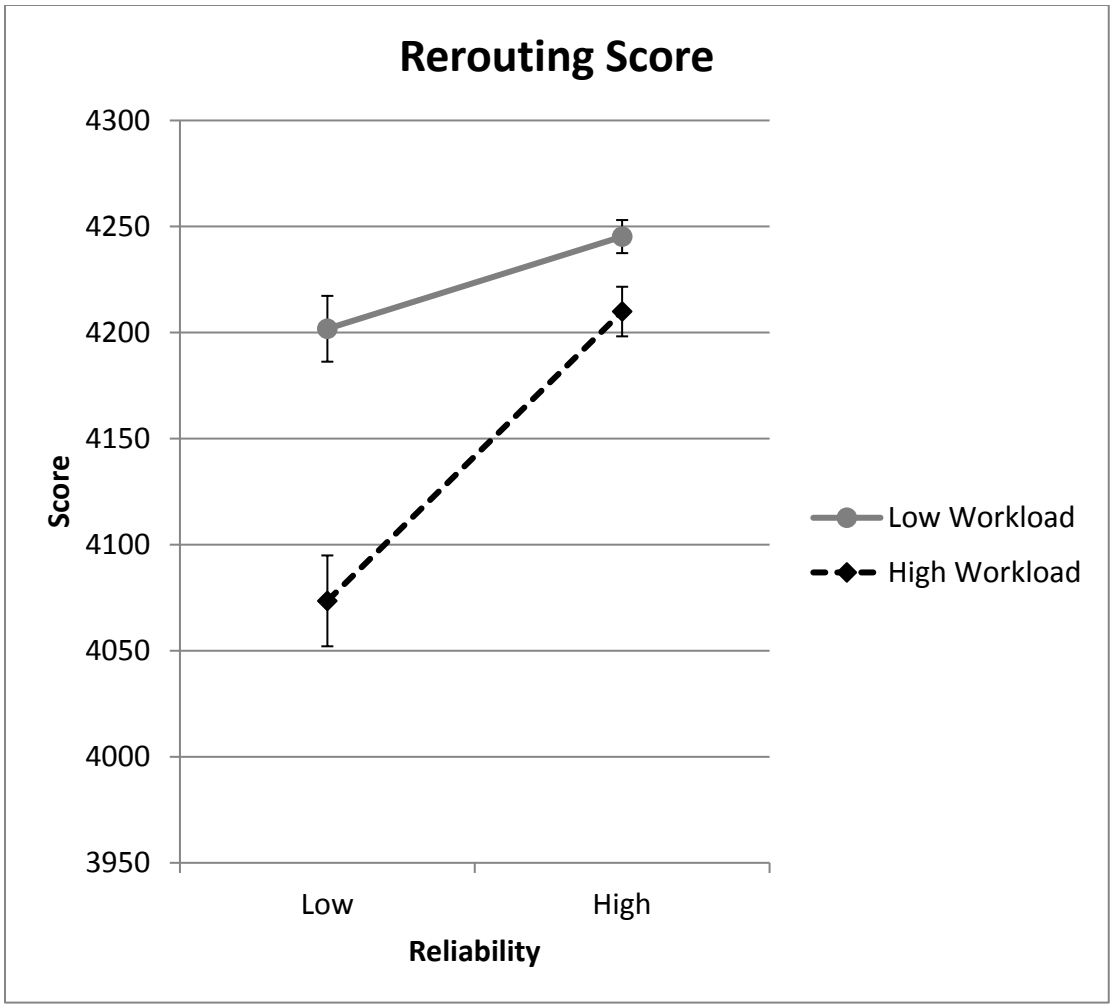

Figure 9. Interaction of workload and automation reliability on rerouting performance. The error bars represent the standard error of the mean.

There was a numerical difference in rerouting scores between the two training groups, but it was not statistically significant [$F(1,14) = 3.35$, $p = 0.089$]. This is due to an interaction between training and workload (Figure 7). We also did not find a main effect of the number of recommendations, one or three [$F(1,14) = 0.07$, $p = 0.796$]. We did find a significant difference of rerouting score as a function of workload [$F(1,14) = 52.77$, $p < 0.001$], with low workload yielding a higher score than high workload (see Figure 10). We also found that scores were higher when the automation recommendation was more reliable [$F(1,14) = 43.63$, $p < 0.001$].

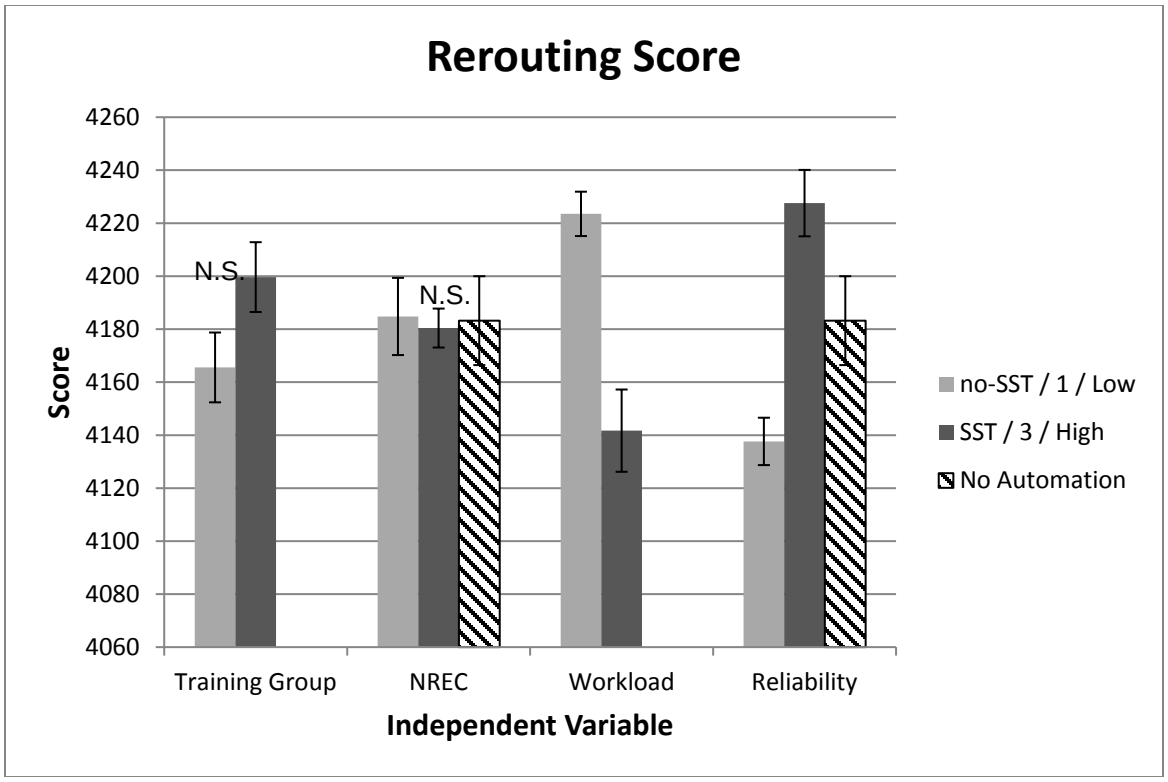

Figure 10. The average rerouting score in each condition. The error bars represent the standard error of the mean. N.S. = no statistically significant difference at $p < 0.05$.

3.3 Secondary Task Results

In addition to the primary rerouting task, participants completed two simultaneous secondary tasks. We analyzed the data from these tasks using the same ANOVA models used for the primary rerouting task, and we will only report effects that were significant at $p < 0.05$. The accuracy for both secondary tasks was extremely high (>90%), and a lack of variance in this measure meant that we could not analyze it. However, we were able to use response time to these tasks as our dependent variable.

3.3.1 NAS Monitoring and Reporting

The NAS Monitoring and Reporting task required the participants to acknowledge when a sector turned red. The participants selected the red cell and entered the number of aircraft expected in a text box. The mean accuracy for this task was 96.33% (3.42%). The ANOVA model could not be fit to the accuracy data, as some conditions had no variance (i.e., all participants scored 100%). However, direct comparisons between the means of each condition using a T-test revealed no differences in accuracy between conditions or group.

The other dependent variable of interest for this task was the response time (RT) between a cell in the NAS array turning red and the participant selecting it to reveal the flight count (Figure 3 from section 2). There was no effect on RT in the no-automation condition. There was no difference in NAS Monitoring RT between the training groups in the no-automation condition, no difference between low and high workload, and no interactions [all $p > 0.102$].

However, in the automation conditions, NAS Monitoring RT was faster when workload was lower (Figure 11) [$F(1,14) = 9.89$, $p = 0.007$]. There were no other main effects and no two-way interactions. However, there was a significant three-way interaction of NREC x Reliability x SST group [$F(1,14) = 5.14$, $p = 0.040$]. In the absence of other main effects and two-way interactions, interpretation of the functional significance of the three-way interaction is difficult to interpret. The functional significance of this effect is unclear.

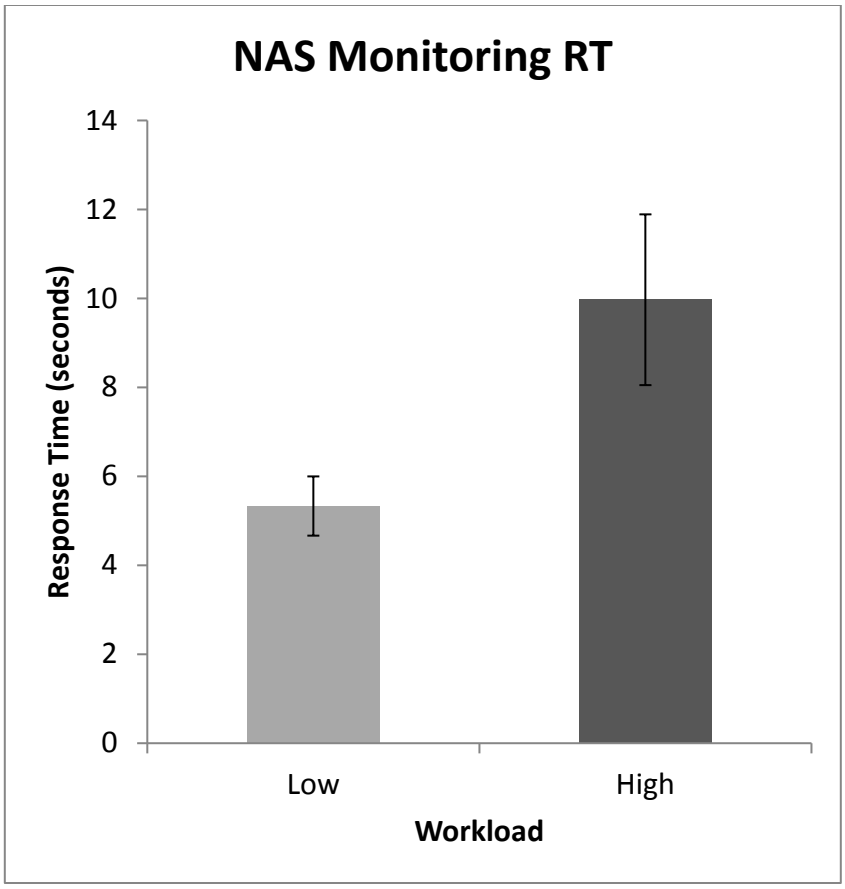

Figure 11. The average response time on the NAS monitoring task for low and high workload. The error bars represent the standard error of the mean.

3.3.2 NTML

The other secondary task pertained to NTML messages. The participants had to either acknowledge or forward each message as described in section 2.2.4.2. The dependent variable of interest for the NTML communication task was the RT between a message appearing in the NTML window and the participant forwarding or acknowledging the message.

There was no difference in NTML RT between the training groups in the no-automation condition [$F(1,14) = 1.43$, $p = 0.252$], but the NTML RT was faster when workload was lower [$F(1,14) = 4.88$, $p = 0.044$]. There was no interaction [$F(1,14) = 0.35$, $p = 0.566$].

In the automation condition, the only significant effect on NTML RT was workload (Figure 12). Response time was faster when workload was lower [$F(1,14) = 8.32$, $p = 0.012$].

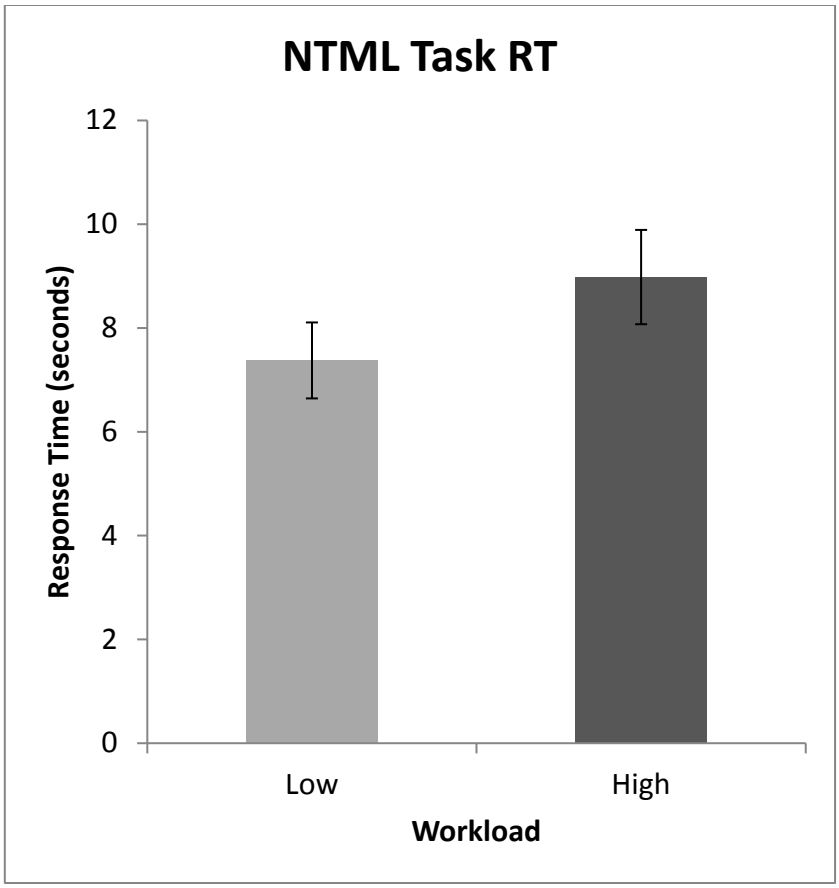

Figure 12. The average response time on the NTML task for low and high workload. The error bars represent the standard error of the mean.

3.4 Task Questionnaire Results

We presented three questionnaires to the participants during the experiment. All questionnaires used 10-point scales, with 1 indicating the lowest rating and 10 indicating the highest rating. This section lists all of the questions and any significant effects.

3.4.1 Post-Reroute Survey

After rerouting each flight, the participants answered two questions about the reroute they just performed (see Appendix D). The average rating of the five flights in each scenario was submitted to the 2x2x2x2 ANOVA for each question.

1. *How confident are you that this route is a good choice?*
   - Participants were more confident that their route was a good choice under low workload conditions [$F(1,14) = 17.03$, $p = 0.001$]. The mean ratings were 7.00 (2.32) and 6.50 (2.02) in the low- and high-workload conditions, respectively.

2. *To what extent did you rely on the* RRT *recommendation in making this rerouting decision?*

- Participants reported relying more on the automation when the workload was high [$F(1,14) = 5.38$, $p = 0.036$]. The mean ratings were 5.14 (1.80) and 5.16 (1.79) in the low- and high-workload conditions, respectively.
- Participants reported relying more on the automation in scenarios when the automation was more reliable [$F(1,14) = 48.50$, $p < 0.001$]. The mean ratings were 4.11 (2.13) and 6.48 (1.64) in the low- and high-reliability conditions, respectively.
- There was a significant interaction between automation reliability and training group. The increase in reported "reliance" on automation from low to high reliability was greater for the SST group than the no-SST group $F(1,14) = 20.79$, $p < 0.001$] (Figure 13). The no-SST group "relied" on automation to about the same extent regardless of the automation's reliability; whereas the SST group "relied" more on high-reliability automation than low-reliability automation.

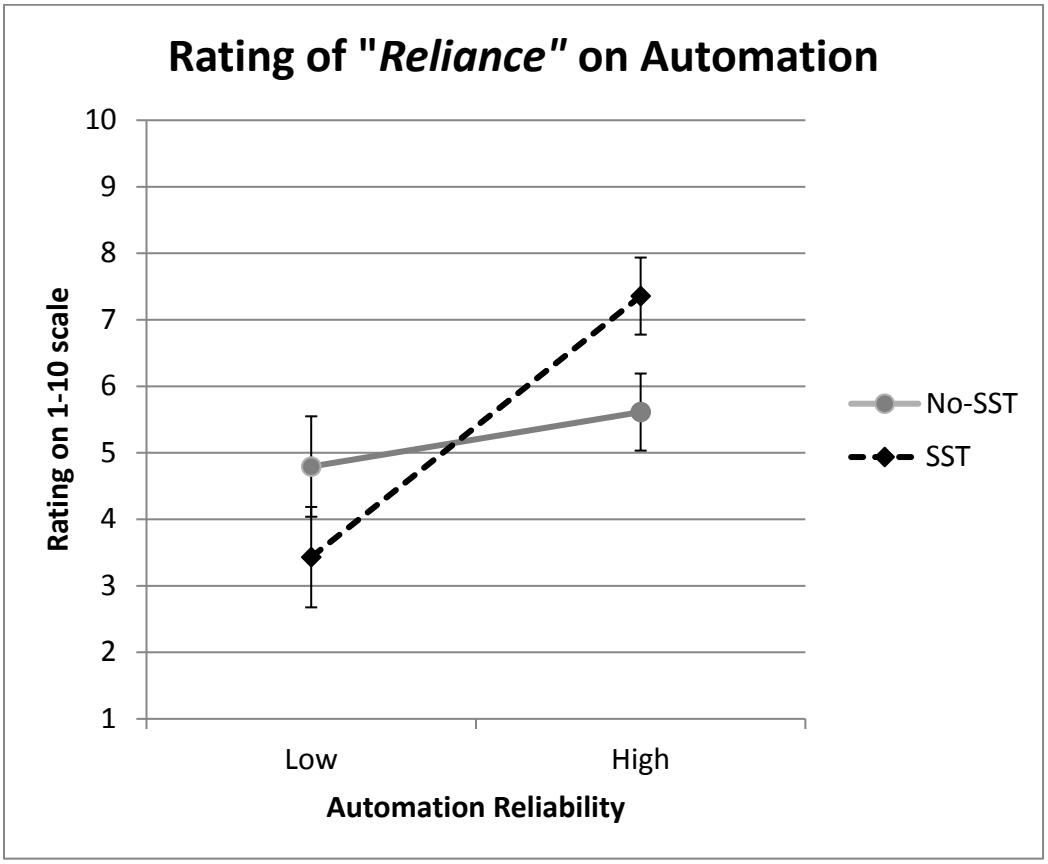

Figure 13. Interaction of SST and automation reliability on ratings of "reliance" on automation. The error bars represent the standard error of the mean.

3.4.2 Post-Scenario Survey

After each scenario, participants answered eight questions, the first six of which were from the NASA Task Load Index (NASA-TLX). The ratings were submitted to the 2x2x2x2 ANOVA for each question. The mean ratings are listed for each question and presented in Figure 14. Note that we were missing one survey after one scenario for one subject due to a data-logging error, so the

degrees of freedom for these tests are lower (*df* = 13) than the other tests in this report (*df* = 14).

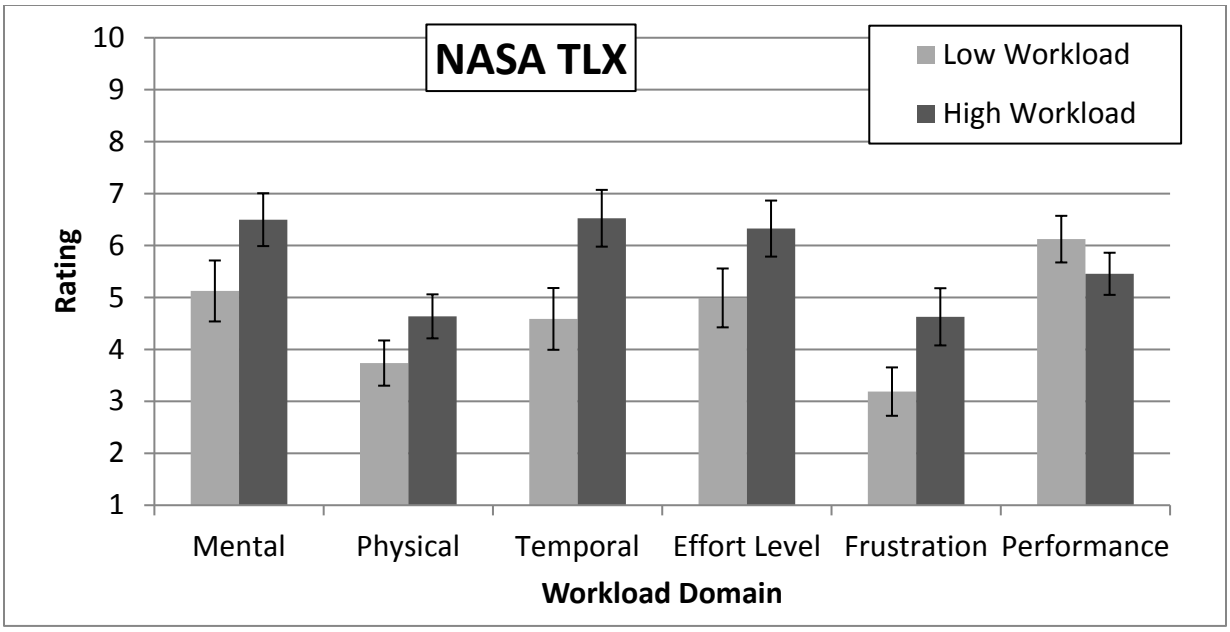

Figure 14. The average rating of six survey questions from the NASA-TLX questions. The error bars represent the standard error of the mean.

1. *Rate your mental demand during this scenario (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.).*

    - Participants rated their mental demand higher under high-workload conditions [$F(1,13) = 38.58$, $p < 0.001$]. The mean ratings were 5.13 (2.35) and 6.50 (2.03) in the low- and high-workload conditions, respectively (Figure 14).
    - There was a significant interaction between Workload and NREC. The magnitude of the increase in rating when workload increased was greater when the automation made one suggestion and then three suggestions [$F(1,13) = 5.02$, $p = 0.043$] (Figure 15).

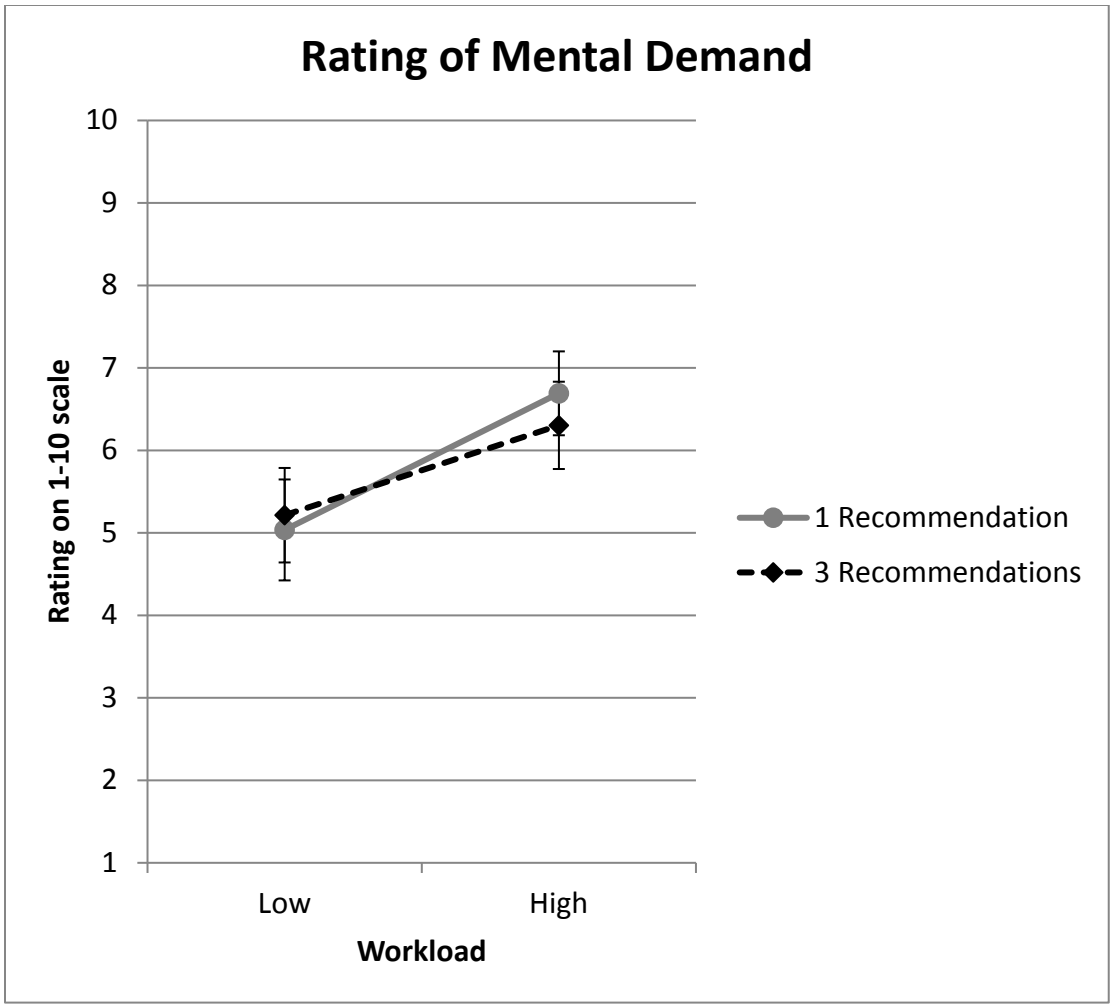

Figure 15. Interaction of workload and number of recommendations on ratings of mental demand. The error bars represent the standard error of the mean.

2. *Rate your physical demand during this scenario (e.g., communications and key presses).*

   - Participants rated their physical demand higher if they were in the No-SST group [$F(1,13) = 11.62$, $p = 0.005$]. The mean ratings were 5.61 (2.43) and 2.77 (2.28) in the No-SST and SST groups, respectively.
   - Participants rated their physical demand higher under high-workload conditions [$F(1,13) = 18.81$, $p < 0.001$]. The mean ratings were 3.74 (1.74) and 4.64 (1.69) in the low- and high-workload conditions, respectively (Figure 14).

3. *Rate your temporal demand during this scenario. (How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?)*

   - Participants rated their temporal demand higher under high-workload conditions [$F(1,13) = 59.49$, $p < 0.001$] The mean ratings were 4.59 (2.37) and 6.52 (2.19) in the low- and high-workload conditions, respectively (Figure 14).

- There was a significant interaction between workload and NREC. The magnitude of the increase in rating when workload increased was greater when the automation made one suggestion than three suggestions [$F(1,13) = 11.95$, $p < 0.001$] (Figure 16).

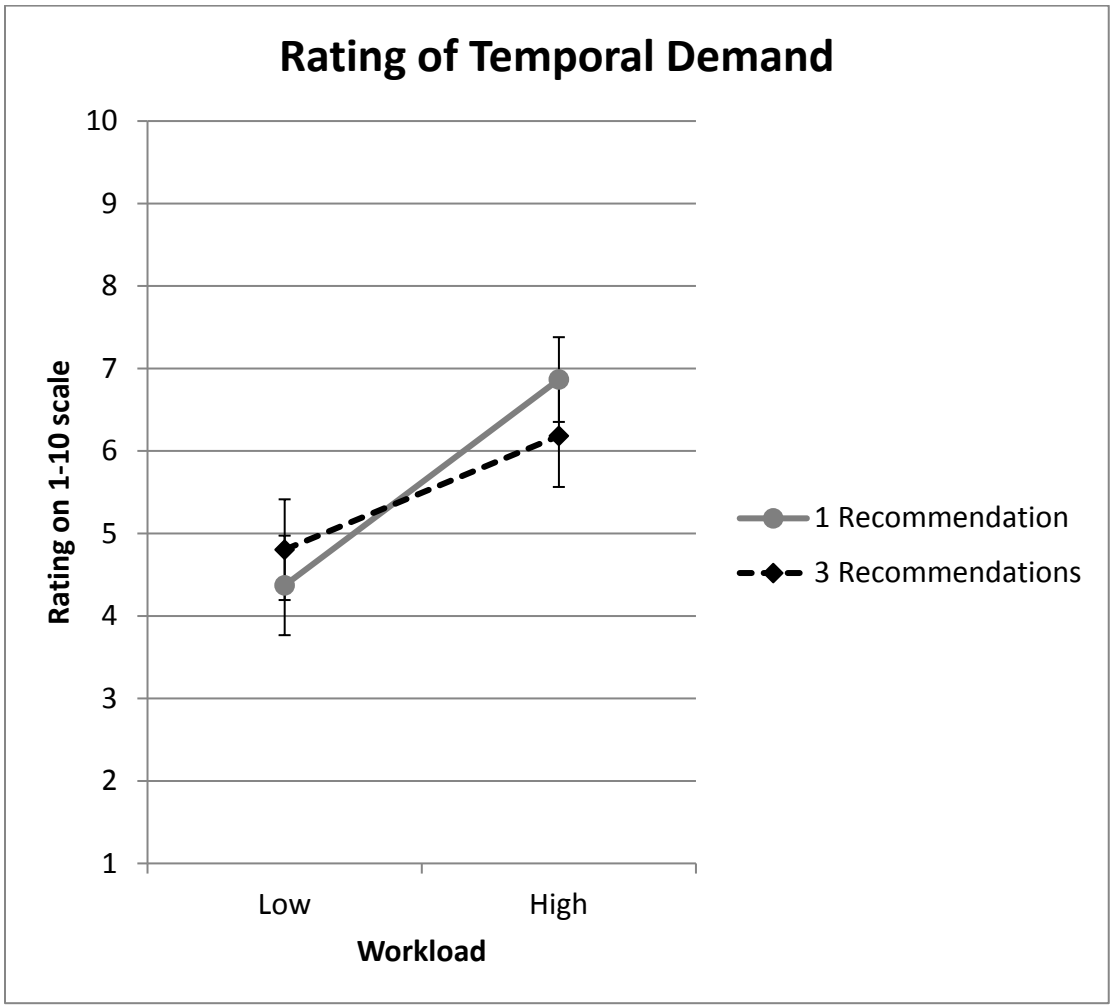

Figure 16. Interaction of workload and number of recommendations on ratings of temporal demand. The error bars represent the standard error of the mean.

4. *Rate your effort during this scenario. [How hard did you have to work (mentally and physically) to accomplish this level of performance?]*

   - Participants rated their level of effort higher under high-workload conditions [$F(1,13) = 40.33$, $p < 0.001$]. The mean ratings were 4.99 (2.67) and 6.33 (2.16) in the low- and high-workload conditions, respectively (Figure 14).
   - There was a significant interaction between workload and NREC. The magnitude of the increase in rating when workload increased was greater when the automation made one suggestion than three suggestions *[$F(1,13) = 5.47$, $p < 0.036$]* (Figure 17).

Figure 17. Interaction of workload and number of recommendations on ratings of the level of effort required to maintain performance. The error bars represent the standard error of the mean.

5. *Rate your frustration level during this scenario. (How insecure, discouraged, irritated, stressed, and annoyed did you feel during the task?)*

    - Participants rated their frustration level higher under high-workload conditions [$F(1,13) = 32.40$, $p < 0.001$]. The mean ratings were 3.19 (1.86) and 4.63 (2.20) in the low- and high-workload conditions, respectively.

6. *Rate your own performance in choosing reroute options without relying on the RRT.*

    - Participants rated their performance higher under low-workload conditions [$F(1,13) = 39.19$, $p < 0.001$]. The mean ratings were 6.12 (1.80) and 5.46 (1.62) in the low- and high-workload conditions, respectively (Figure 14).
    - There was a significant interaction between workload and SST training. The increase in rating for low workload was greater in the SST group than no-SST group [$F(1,13) = 5.02$, $p = 0.043$] (Figure 18).

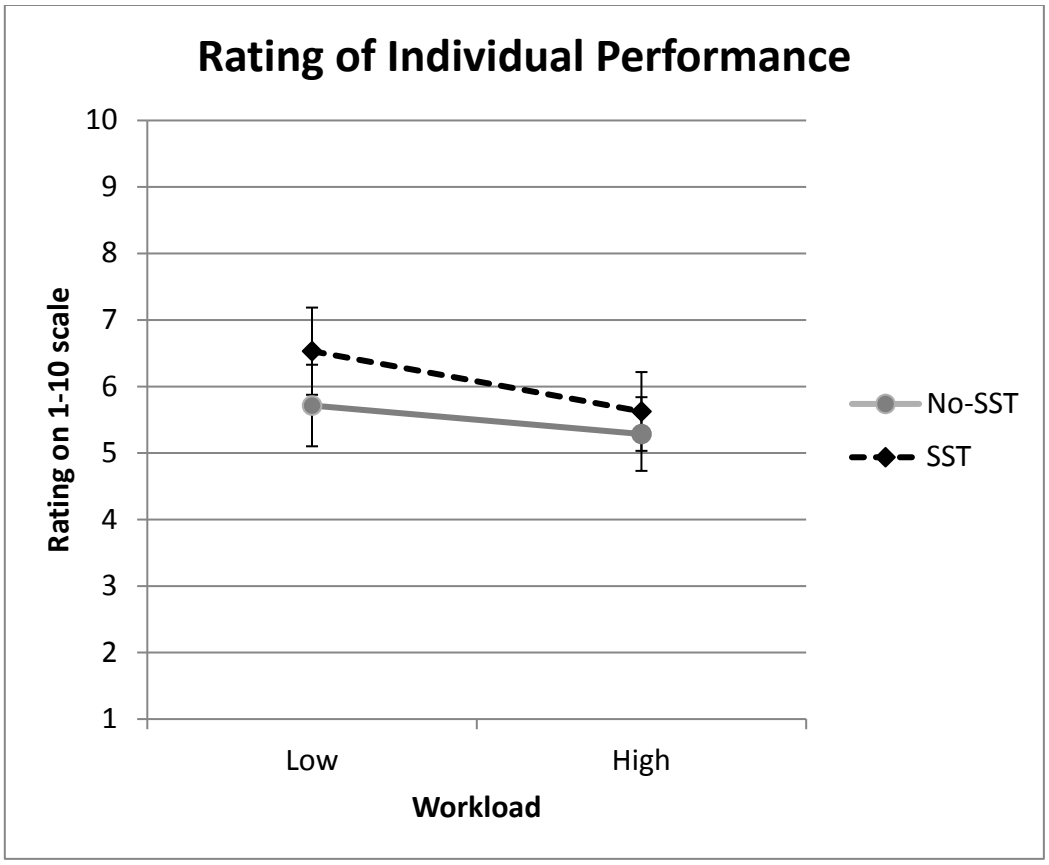

Figure 18. Interaction of workload and SST on ratings of individual performance. The error bars represent the standard error of the mean.

7. *Rate the RRT's performance in suggesting appropriate reroutes for this scenario.*

- Participants rated the RRT's performance higher for more highly reliable scenarios [$F(1,13) = 15.25$, $p = 0.002$]. The mean ratings were 4.39 (2.13) and 6.33 (1.77) for low- and high-reliability scenarios, respectively.
- Participants rated the RRT's performance higher when it made one recommendation rather than three recommendations [$F(1,13) = 5.23$, $p = 0.040$]. The mean ratings were 5.65 (1.94) and 5.06 (2.02) for one and three recommendations, respectively.
- There was a significant interaction between automation reliability and SST training. The SST participants showed a larger increase in their ratings between the low- and high-reliability scenarios (7.19 > 4.03) than the no-SST group (5.46 > 4.75), SST by Reliability interaction [$F(1,13) = 6.07$, $p = 0.028$] (Figure 19).

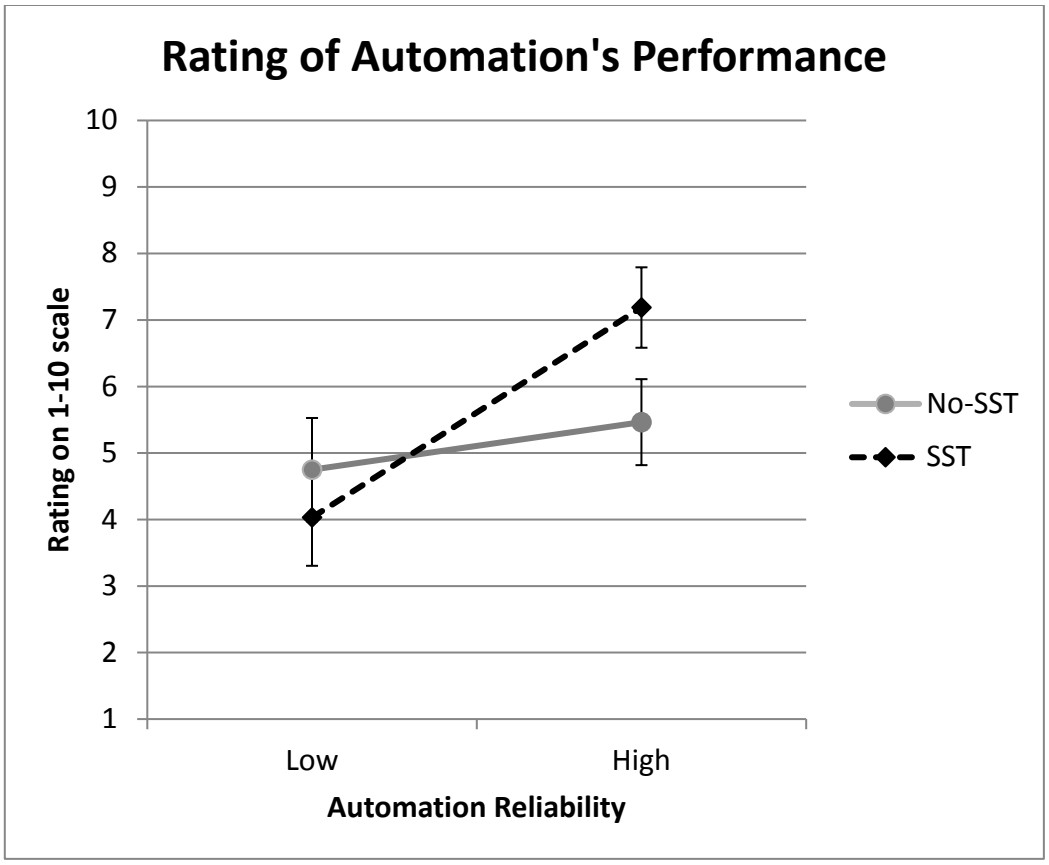

Figure 19. Interaction of automation reliability and SST on ratings of the automation's performance. The error bars represent the standard error of the mean.

8. *Rate the overall system performance (you and the RRT as a "team") in making appropriate rerouting decisions for this scenario.*

    - There was no effect of any of the independent variables on participant's rating of system performance other than a three-way interaction of NREC x Reliability x SST [$F(1,13) = 5.03$, $p = 0.043$]. In the absence of any main effects, the interaction is not interpretable.

3.4.3 Post-Experiment Survey

At the conclusion of the experimental scenarios, participants completed a final set of nine questions that asked them to rate various aspects of their own performance and the automation's performance. We submitted the ratings for the first seven questions to a two-tailed T-test to examine the differences between the training groups. The associated statistics are presented in Table 6. The SST group rated the RRTs performance in suggesting appropriate reroutes higher than the no-SST group. There was no other statistically significant difference between training groups in the first seven questions. The last two questions had oppositional wording, both related to using an RRT suggested reroute and busyness. Therefore, we were able to use a more powerful and sensitive statistical test for these two questions. We used a 2x2 mixed-model ANOVA (Training group x Busyness repeated-measure factor) to take advantage of their relatedness and shared variance. The SST group rated that they were more likely to use the automation's suggested reroute ($M = 6.56$, $SD = 2.23$) than the no-SST group ($M = 4.81$, $SD = 2.23$) [$F(1,14) = 4.94$, $p = 0.043$]. In addition, participants rated an

increased likelihood to use a suggested reroute when they were busy (M = 6.63, SD = 1.94) than when they were not busy (M = 4.75, SD = 1.63) [$F(1,14) = 19.33$, $p < 0.001$]. There was no interaction between training group and busyness.

Table 6 Results of Post-Experiment Survey Analyses with Means.

Question	SST	No-SST	Significant
1. *Rate your own performance in choosing reroute options without relying on the RRT.*	6.75 (1.58)	6.00 (1.69)	(N.S) $p = 0.375$
2. *Rate the RRT's performance in suggesting appropriate reroutes.*	**7.13 (0.64)**	**5.00 (1.93)**	**$t(14) = 2.96$ $p = 0.010$**
3. *Rate the overall system performance (you and the RRT as a "team") in making the appropriate rerouting decisions.*	7.25 (0.89)	6.63 (2.13)	(N.S) $p = 0.457$
4. *Overall, how much did you trust the RRT to provide a good choice(s)?*	6.25 (0.89	4.88 (1.96)	(N.S) $p = 0.092$
5. *Rate your level of performance in the NAS Monitor and NTML tasks.*	7.50 (1.85)	7.25 (1.85)	(N.S) $p = 0.843$
6. *Rate your overall workload for this experiment.*	6.25 (1.75)	7.13 (2.36)	(N.S) $p = 0.413$
7. *How much did workload of the NAS Monitor and NTML tasks affect your performance on the rerouting task?*	5.63 (1.75)	6.63 (3.02)	(N.S) $p = 0.448$
8. *How likely were you to use a RRT suggested reroute when you were busy?*	**6.56 (2.23)**	**4.81 (2.23)**	**$F(1,14) = 4.94$ $p = 0.043$**
9. *How likely were you to use a RRT suggested reroute when you were not busy?*			

4. SUMMARY AND RECOMMENDATIONS

The goal of this part-task experiment was to investigate factors that contribute to or influence the use of DSTs. We examined three conditions in this experiment: workload, automation reliability, and the number of recommendations provided by the automation. In addition, we examined the effects of supplemental DST training. Half of the participants completed the additional SST and half did not. We used a simplified weather rerouting task and two secondary tasks to obtain objective measures of behavior. We also presented several questionnaires throughout the experiment to assess the participant's subjective experience with DST use. In this section of the report, we present our conclusions of the results and provide speculations on the potential relevance to operational situations for each IV.

4.1 Workload

We manipulated workload by changing the number of responses needed for participants to accomplish secondary tasks. Our manipulation was effective, in that high workload negatively

impacted objective measures of performance and subjective ratings of workload. Rerouting performance decreased when workload increased. It is not surprising that high workload would negatively impact performance. However, the interpretation of the interactions between workload and the other experimental factors, discussed later, depended on our workload manipulation being effective.

We found workload to be an important factor contributing to DST use in our task. The development and deployment of DSTs for operational use must account for user workload to be maximally effective.

4.2 Number of Recommendations

The inspiration to manipulate the NREC made by the automation comes from the levels of automation theory (Sheridan & Verplank, 1978). This theory predicts different outcomes from situations when automation makes one versus multiple recommendations. Our manipulation of the NREC had no direct impact on objective measures of performance. Rerouting performance did not differ between conditions that provided one recommendation or three recommendations. However there was an interaction between the NREC and automation reliability. When the automation was highly reliable, the participants performed better when the automation made one recommendation than when it made three recommendations. Conversely, when the automation was not as reliable, participants performed better when the automation made three recommendations than when it made one recommendation. We also found statistical interactions between NREC and workload on several subjective measures of workload. Participants rated the RRT's performance higher overall when it made one recommendation than when it made three recommendations. This leads to a complicated picture regarding the ideal number of recommendations for a DST to provide. The ideal number of recommendation for a DST likely depends upon the task at hand and the current workload of the user.

We can use an analogy with GPS navigation while driving a car to illustrate this point. When we are driving and our preferred route closes, we need to find an alternate route. As we are likely very busy with the tasks involved in driving, we would prefer our DST (e.g., a navigation app on our smartphone) to make *one* reliable recommendation. In this high-workload context, a DST that proposes *multiple* options will force the user to use valuable time and attention resources to evaluate and choose among many options. In contrast, if we are planning a future trip, our workload is not as high, and we can evaluate multiple options.

4.3 Automation Reliability

We manipulated the automation's reliability by changing how likely it was to provide a high-scoring route. Scenarios with highly reliable recommendations led to increased performance on the rerouting task. In addition, participants subjectively rated the RRT's performance higher for high-reliability scenarios. They also reported "relying" more on the automation in scenarios when the automation was more reliable. Therefore, we can be confident that our reliability manipulation was effective. Reliability and the interaction of reliability with workload provide the most insight into the application of this study's results to the operational environment.

When the automation was not reliable, performance on the rerouting task suffered. A high workload also decreased performance on the rerouting task (Figure 7). When the route recommendations provided by the automation were not optimal, there was an increased need to evaluate alternative routes. When workload was low, the participants had the necessary time and cognitive resources to perform that evaluation. However, when the participants were under high

workload, they did not have the time to evaluate the alternatives fully and were forced to rely on the automation's recommendations. When the DST recommendations were not reliable, this led to poor performance.

When workload is low, there is enough time to evaluate all alternative options including those the DST has generated as well as the participant's own. In these situations, even a low-reliability DST may be helpful. When traffic managers are busy (which is probably most of the time), there is less time to fully evaluate all options, and the DSTs need to be more reliable. A low-reliability DST might be less helpful (or even harmful) to performance in a high-workload environment. Therefore, the users' workload context should be considered when evaluating how reliable an individual DST needs to be in order to be most effective. Our study used only two levels each of reliability and workload: low and high. It is unclear at what exact workload level an unreliable DST recommendation becomes useful. Future research is needed to get a better sense of the actual levels of workload and reliability that would have an impact in an operational setting.

The reliability and number of recommendations had an interactive effect on performance (Figure 9). When reliability was high, one recommendation led to better performance. However, when reliability was low, three recommendations led to better performance. This suggests that a DST that provides only one recommended course of action should be provided if the reliability of the suggestion is high. If the reliability of the tool is not as good, it may be better for it to provide multiple alternatives.

If we return to our example of rerouting a car while driving, the single recommendation needs to be reliable. If it is not a good reroute, time and fuel are wasted, and frustration increases. If the automation's recommendations are unreliable, pulling over and evaluating more options may be preferable. It may make a decision more difficult in the short term, but it may lead to more positive outcomes overall by avoiding poor reroutes.

4.4 Training

We distributed our participants into two groups. Both groups received a training session about how to perform the task. However, one group received additional SST that provided details about how the automation worked and the conditions under which it would be more or less reliable. The other group (i.e., no SST) did not receive this additional training. The rest of the experiment was identical for the two groups. We wanted the two groups to be as similar as possible, because we wanted any differences in performance between the groups to be due to our training manipulation and not to random pre-experimental variables. Our two groups were matched with regard to age, gender, and human factors research experience but were otherwise randomly assigned. The two groups scored similarly on the Complacency Rating Scale, a survey designed to assess individual differences in attitudes toward automation and susceptibility to overreliance in automation. In addition, the two groups did not perform differently on the rerouting task in the scenarios with no automation recommendations (Figure 6). Therefore, we can be confident that any effect the SST had on performance in the scenarios with automation was not due to a random difference in ability between the two groups.

The group that received the additional training scored numerically higher than the group that did not, although we did not find a statistically significant difference. This was likely due to the interaction between workload and training (Figure 8). When workload was low, the two groups performed equally. However, when workload was high, the SST group performed much better than the no-SST group. The additional training helped mitigate the performance decline caused by high workload. Knowing the situations when the automation could be relied upon, meant the SST

participants did not have to evaluate all of the options as thoroughly as the untrained participants. This improved performance in high-workload situations when evaluation time was reduced. The SST participants also rated that they were more likely to use the automation's suggested reroute than the no-SST group.

5. CONCLUSION

In summary, decision-support tools (DSTs) can improve performance, but care should be taken in the areas of reliability, training, and tool design. DSTs are most needed by traffic managers when they are dealing with system stressors such as weather, equipment outages, and congestion. These are high-workload situations during which they are attempting to deal with the current constraints and meet the demands of multiple tasks. Based on the findings of this part-task study, the reliability of the DST may be a critical factor in whether they get any benefit out of its use. If the DST is going to provide a single recommended solution, it should be a reliable one, otherwise performance may suffer. Finally, a training that addresses the strengths and weaknesses of the DST has a big impact on performance when workload is high. Based on the findings reported here, when workload increases and performance starts to decline, the proper training can counteract this decline by helping the traffic managers make the best use of the DST.

This study was conducted with novices, but experienced Traffic Flow Management (TFM) personnel have developed solution sets based on strategies that have worked for them in the past. If the automated recommendations are not in line with those proven strategies, they may be less likely to use the DST, even when workload is high. The downfall of this behavior is that the automated recommendations may be better solutions than the ones they have relied on in the past. Therefore, to succeed, the automation will have to be reliable, and the experienced TFM personnel will need training that effectively convinces them that the automation can be trusted in specific situations.

It will be important to conduct a study, similar to the one presented here, using TFM personnel with a range of experience levels to determine whether the same results we found in novices apply to experienced users. It is likely that experienced TFM personnel and less-experienced TFM personnel differ from one another in how they make use of the DST and the type of training that is most effective. DST training for more experienced users may need to be targeted more specifically to helping them determine where benefits from the automation can be gained, such as by providing information as to when the tool provides a faster resolution, or identifying and providing solutions to situations that these users encounter less frequently.

Another important consideration for experienced TFM personnel is the level of effort required to use the DST. If they find that it takes significantly more time, or requires more steps to generate and implement an automated recommendation, they may quickly revert to their former solution sets and processes. The design of the DST, its usability, and its integration into their workflows become very important.

Less-experienced TFM personnel may behave differently. They have fewer situations to draw from and may be less certain about their decisions. They may be more likely to rely on automation recommendations. Training for these users may be effective at a more general level than training for experienced users because the less experienced users are still learning the effects of the different conditions that influence outcomes. Their results may be more similar to those of the novices in the current study. It would be useful to evaluate whether less experienced and more experienced TFM personnel differ from one another in their level of reliance on automation and in their level of complacency regarding automation. It is also important to assess whether users with different

experience levels demonstrate different workload thresholds at which they begin to rely on automation and under which conditions their workload is most affected. Understanding these distinctions will be useful in developing more targeted training solutions for a range of users and will ultimately enable them to make more effective use of the DST.

References

Davison Reynolds, H. J., & DeLaura, R. A. (2011). *Concept of operations for the Integrated Departure Route Planning (IDRP) tool* (Project Report No. ATC-379). Lexington, MA: Lincoln Laboratory, Massachusetts Institute of Technology.

DeLaura, R. A., Underhill, N. K., Hall, L. M., & Rodriguez, Y. G. (2012). *Evaluation of the Integrated Departure Route Planning (IDRP) tool 2011 prototype* (Project Report No. ATC-388). Lexington, MA: Lincoln Laboratory, Massachusetts Institute of Technology.

Dixon, S.R. & Wickens, C.D. (2006). Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high workload. Human Factors, 48, 474-486. doi: 10.1518/001872006778606822

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243–1270. doi: 10.1080/00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153–184. doi: 10.1080/00140139.2016.1261187

Masalonis, A. J. (2000). *Effects of situation-specific reliability on trust and usage of automated decision aids* (Unpublished doctoral dissertation). Washington, DC: The Catholic University of America.

Masalonis, A. J., & Parasuraman, R. (1999). Trust as a construct for evaluation of automated aids: Past and future theory and research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, 184–188. doi: 10.1177/154193129904300312

Masalonis, A. J., Zingale, C. M., & Puzen, G. M. (2016). *NextGen Traffic Flow Management (TFM) Tools: Guidance for use, integration, and training: Annotated bibliography.* Unpublished internal document. Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Masalonis, A. J., Zingale, C. M., Puzen, G. M., Thomas, W. J., & Yuditsky, T. (2016). *NextGen Traffic Flow Management (TFM) Tools: NextGen Tools Assessment.* Unpublished internal document. Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Rein, J. R., Masalonis, A. J., Messina, J., & Willems, B. (2013). Meta-analysis of the effect of imperfect alert automation on system performance. *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*, 280–284. doi: 10.1177/1541931213571062

Sheridan, T.B., & Verplank, W. (1978). Human and computer control of undersea teleoperators. Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *International Journal of Aviation Psychology, 3*(2), 111–122. doi: 10.1207/s15327108ijap0302_2

Sorkin, R.D., Kantowitz, B.H., and Kantowitz, S.C. (1988). Likelihood Alarm Displays. Human Factors, 30(4), 445–459. doi: 10.1177/001872088803000406

Trapsilawati, F., Qu, X., Wickens, C. D., & Chen, C.-H. (2015). Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control. *Ergonomics, 58*(6), 897–908. doi: 10.1080/00140139.2014.997301

Wickens, C. D. & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201–212. doi: 10.1080/14639220500370105

Wiegmann, D.A., (2002). Agreeing with Automated Diagnostic Aids: A study of users; concurrence strategies, Human Factors and Ergonomics Society, 44 (1), pp. 44-50. doi: 10.1518/0018720024494847

Acronyms

ANOVA	Analysis of variance
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control
ATCSCC	Air Traffic Control System Command Center
ATCT	Air Traffic Control Tower
ATM	Air Traffic Management
CPRS	Complacency-Potential Rating Scale
CRA	Conflict Resolution Advisory
DST	Decision-support tool
IV	Independent variable
JBU	JetBlue Airways
NAS	National Airspace System
NASA	National Aeronautics and Space Administration
NASA-TLX	NASA Task Load Index
NREC	Number of recommendations
NTML	National Traffic Management Log
RDHFL	Research Development and Human Factors Laboratory
RRT	Route recommendation tool
RT	Response time
SME	Subject matter expert
SST	Situation-Specific Training
TFM	Traffic Flow Management
TMC	Traffic Management Coordinator
TMS	Traffic Management Specialist
TRACON	Terminal Radar Approach Control
UAL	United Airline
WJHTC	William J. Hughes Technical Center

Appendix A: Informed Consent Statement

Informed Consent Statement

I, _____, understand that this study, entitled NextGen Traffic Flow Management (TFM) Tools: Guidance for Use, Integration, and Training, is sponsored by the Federal Aviation Administration (FAA).

Nature and Purpose:

I have been recruited to volunteer as a participant in this project. The purpose of this study is to develop a better understanding of human behavior when using the types of Decision Support Tools planned for the Traffic Flow Management domain. The results of this study will generate considerations and suggestions for adding Decision Support Tools in the TFM environment.

Study Procedures:

Approximately sixteen (16) volunteers, primarily from the William J. Hughes Technical Center, will participate in this experiment, designed to simulate some of the demands of TFM. The experiment plus all training and questionnaires will take approximately 2 hours. After the conclusion of the experiment, participants and researchers will conduct a final debriefing session to share questions, comments, and feedback.

Anonymity and Confidentiality:

My participation in this simulation is strictly confidential. Any information I provide will remain anonymous; no individual names or identities will be associated with the data or released in any reports.

Benefits:

I understand that the only benefit to me is that I will be able to provide valuable feedback and insight into the effectiveness of potential ATC tools and procedures. My contribution will help the FAA to determine the benefits and feasibility of these modifications.

Participant Responsibilities:

I will perform the tasks presented during this study to the best of my ability and will answer all questions asked during the study truthfully. I will not discuss the content of the study with anyone until the study is completed.

Participant Assurances:

I understand that my participation in this study is completely voluntary and I can withdraw at any time without penalty. I also understand that the researchers in this study may terminate my participation if they feel this to be in my best interest. I understand that if new findings develop during the course of this study that may relate to my decision to continue participation, I will be informed. I have not given up any of my legal rights or released any individual or institution from liability for negligence.

The research team has adequately answered all the questions I have asked about this study, my participation, and the procedures involved. I understand that Carolina Zingale or another member of the research team will be available to answer any questions concerning procedures throughout this study. If I have questions about this study or need to report any adverse effects from research procedures, I will contact Carolina Zingale at 609-485-8629.

Discomfort and Risks:

I understand that I will not be exposed to any known risks or intrusive measurement techniques. I agree to immediately report any injury or suspected adverse effect to Carolina Zingale.

Signature Lines:

I have read this informed consent form. I understand its contents, and I freely consent to participate in this study under the conditions described. I understand that, if I want to, I may have a copy of this form.

Participant:_____ Date:_____

Investigator:_____ Date:_____

Witness:_____ Date:_____

Appendix B: Demographics Questionnaire

Demographics Questionnaire

Participant #_____ Age _____ Gender_____

Do you have normal color vision? (Y / N)

If no, please explain:_____

Have you ever worked as an Air Traffic Controller? (Y / N)

_____

Have you ever worked as a Traffic Manager? (Y / N)

_____

Are you or have you ever been a licensed pilot? (Y / N)

If yes, please indicate license and experience_____

What other experience (e.g., R&D) do you have with Air Traffic Control, aircraft piloting, or airline operations?

_____

_____

_____

What experience (e.g., R&D) do you have with Traffic Flow Management?

_____

_____

_____

Have you discussed this experiment with any of your colleagues who have already participated in it?

YES / NO

What do you know about the tasks you will be performing?

_____

Appendix C: Modified Complacency Rating Scale

Modified Complacency Rating Scale

INSTRUCTIONS Participant #_____

Read each statement carefully and check one response out of five alternatives in the appropriate box that you feel most accurately describes your views or experiences. The responses vary on a scale of agreement/disagreement, from "strongly agree" to "strongly disagree". For example:

Statement: Doing research in a library has been made easier by the introduction of computerized card cataloging systems.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

Give your answer for each statement and be sure to place your response in the correct place. Remember, this is an opinion survey and not a test of intelligence or ability. There are no right or wrong answers, only answers that fit your views accurately. Do not skip any question. Time is limited. Do you have any questions?

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

   ☐ ☐ ☐ ☐ ☐

   Strongly Agree Agree Undecided Disagree Strongly Disagree

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

   ☐ ☐ ☐ ☐ ☐

   Strongly Agree Agree Undecided Disagree Strongly Disagree

3. People save time by using automatic teller machines (ATMs) rather than a bank teller for banking transactions.

   ☐ ☐ ☐ ☐ ☐

   Strongly Agree Agree Undecided Disagree Strongly Disagree

4. I do not trust automated devices such as ATMs and computerized airline reservation systems.

   ☐ ☐ ☐ ☐ ☐

   Strongly Agree Agree Undecided Disagree Strongly Disagree

5. People who work frequently with automated devices have lower job satisfaction because they feel less involved in their job than those who work manually.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

7. 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.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

8. People whose jobs require them to work with automated systems are lonelier than people who do not have to work with such devices.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

12. I often use automated devices.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

13. People who work with automated devices have greater job satisfaction because they feel more involved than those who work manually.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

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

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

17. 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.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

18. Work has become more difficult with the increase of automation in aviation and banking.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

19. I do not like to use ATMs because I feel that they are sometimes unreliable.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

20. I think that technology used in medicine, such as CAT scans and ultrasound, help to provide very reliable medical diagnosis.

☐ ☐ ☐ ☐ ☐

Strongly Agree Agree Undecided Disagree Strongly Disagree

Appendix D: Screenshots of All Computer-Based Surveys

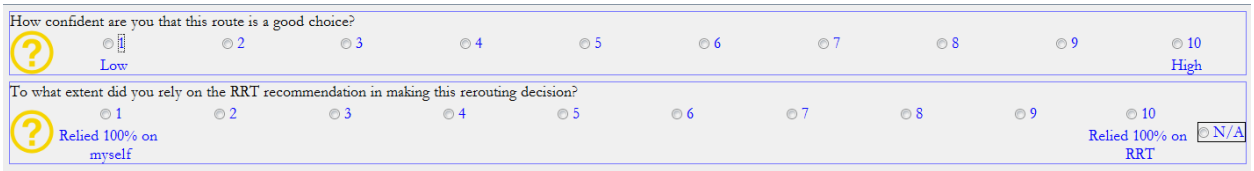

Figure D1. The survey questions presented after each aircraft reroute. N/A was given as an option when there was no automation.

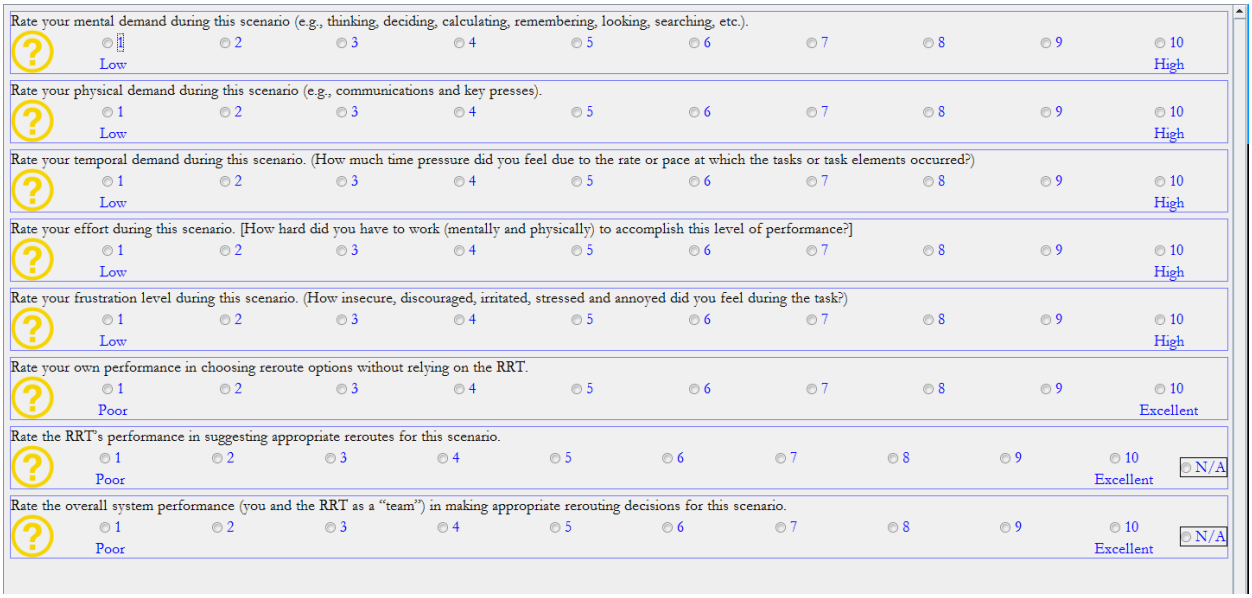

Figure D2. The survey questions presented after each scenario was completed. N/A was given as an option when there was no automation.

Figure D3. The survey questions presented at the conclusion of all ten test scenarios. N/A was given as an option when there was no automation.

Appendix E: Counterbalancing and Condition/Scenario Order

Counterbalancing and Condition Assignment

Instead of randomizing condition orders or systematically varying them in a Latin square or similar design, we explicitly specified certain orders for the combinations of independent variables in a way that attempted to avoid overly biasing any participant into a certain initial attitude about the automation or the task. Such biases might happen, for example, if too many consecutive high- or low-reliability scenarios or too many high- or low-workload scenarios occurred, especially at the beginning of the experimental trials. We determined that high reliability should always be seen first to avoid initially "souring" participants on the automation. Counterbalancing or explicitly manipulating the order of reliability levels is a potentially interesting avenue, but a complex question beyond the scope of this research. For example, the qualitative experience of several high-reliability followed by several low-reliability trials, and the resulting attitudes and behaviors, is likely not the direct inverse of experiencing these conditions in the reverse order or experiencing alternating reliability levels. Counterbalancing these orders would not necessarily have the desired effect of cancelling out order effects of other variables, but could interact with other variables in unexpected ways and confound the results. Varying the reliability order could make it hard to interpret the results and would become a whole separate manipulation/experiment. Therefore, over the course of the conditions containing automation, each participant was assigned the same order of reliability levels, one that began with a high-reliability scenario and then roughly, but not exactly, alternated between high and low reliability.

In addition, for any one participant, we always blocked all scenarios of each number of recommendations (NREC) condition—one recommendation (four scenarios), three recommendations (four scenarios), and no-automation control condition (two scenarios)—and ran both/all of the scenarios in each NREC level consecutively to avoid potentially disorienting rapid shifts between automation levels. We based all of the above condition ordering strategies on those used in past automation trust studies such as Lee and Moray (1992, 1994) and Masalonis (2000).

To derive condition orders for our 16 participants, we first constructed two initial prototype orders, each of which would be assigned to one participant. We built these according to the principles outlined above. Eight possible combinations of the within-participants independent variables (IVs) existed [Reliability (2) x NREC (2) x Workload (2)], plus two possible presentations of the control condition [the two levels of Workload]. Ten combinations of the within-participants IVs resulted and are depicted in Table E-1.

Table E-1 Ten combinations of the within-participants IVs.

Rel	NREC	Work
HI	3	lo
LO	3	Lo
HI	3	Hi
LO	3	Hi
HI	1	Lo
HI	1	Hi
LO	1	Hi
LO	1	Lo
--	--	Lo
--	--	Hi

We specified two initial prototype condition orders, using the exact order listed in Table E-1. The two differ from each other according to the automation algorithm (e.g., X or Y) and scenario type (e.g., 1 for eastbound and 2 for westbound) used in the scenario containing a specific combination of the other IVs. For example, one of the two initial orders used Algorithm X (scenario type 1) for the high-reliability, NREC3, low-workload scenario; the other of the two orders used Algorithm Y (scenario type 2) for that scenario.

The reason for this manipulation is that, although the variable-reliability rules we designed into the scenarios were arbitrary, we deemed it necessary to cross these manipulations as fully as possible with the meaningful IVs. The goal was to prevent any unanticipated expectations or attitudes participants might develop regarding one of the algorithms or one of the scenario types.

For the eight scenarios in the prototype condition order that contained automation, we assigned algorithms X and Y to four scenarios each, crossed as much as possible with the other IVs. The reliability level already in place for each scenario would then dictate the scenario type assigned based on the variable reliability rules. We did this assignment one way for the first prototype condition order and assigned the reverse combinations to the second prototype. For the two control scenarios, in the first prototype, we randomly assigned one of the two scenario types (1 or 2) to the high-workload condition, and the other to the low (last two rows of Table E-1). We assigned the opposite workload/scenario type combinations in the second prototype. The four prototype condition orders are shown in Table E-2.

Table E-2 The resulting four prototypes of conditions.

Prototype condition order 1						Prototype condition order 2				
Rel	NREC	Work	Algo	Type		Rel	NREC	Work	Algo	Type
HI	3	lo	X	1		HI	3	lo	Y	2
LO	3	lo	X	2		LO	3	lo	Y	1
HI	3	hi	Y	2		HI	3	hi	X	1
LO	3	hi	Y	1		LO	3	hi	X	2
HI	1	lo	Y	2		HI	1	lo	X	1
HI	1	hi	X	1		HI	1	hi	Y	2
LO	1	hi	X	2		LO	1	hi	Y	1
LO	1	lo	Y	1		LO	1	lo	X	2
--	--	hi	--	2		--	--	hi	--	1
--	--	lo	--	1		--	--	lo	--	2

Prototype condition order 3						Prototype condition order 4				
Rel	NREC	Work	Algo	Type		Rel	NREC	Work	Algo	Type
HI	3	hi	X	1		HI	3	hi	Y	2
LO	3	hi	X	2		LO	3	hi	Y	1
HI	3	lo	Y	2		HI	3	lo	X	1
LO	3	lo	Y	1		LO	3	lo	X	2
HI	1	hi	Y	2		HI	1	hi	X	1
HI	1	lo	X	1		HI	1	lo	Y	2
LO	1	lo	X	2		LO	1	lo	Y	1
LO	1	hi	Y	1		LO	1	hi	X	2
--	--	lo	--	2		--	--	lo	--	1
--	--	hi	--	1		--	--	hi	--	2

To vary the order in which participants ran the NREC conditions, while always keeping all scenarios for an NREC together in a block of scenarios, we cloned each of the initial four condition orders three additional times so that four versions existed of each of the original four condition orders. For each of the four versions of an original condition order, we modified the order in which the participants would experience the NREC scenario blocks, using the following possible orders:

- NREC 3, NREC 1, control (this was the order in the original four)
- control, NREC 3, NREC 1
- NREC 1, NREC 3, control
- control, NREC 1, NREC 3

Table E-3 summarizes this manipulation. Each column represents one of the original four condition orders. For the four orders within a column, the reliability-, workload-, algorithm-, and

scenario-type variables are in the same order for the eight automation conditions (as can be seen, the eight automation scenarios are always consecutive, the first eight or the last eight). The order of high and low workload among the two control scenarios also remains the same within a column. Only the order of the NREC variable varies within a column.

Table E-3 The four potential orders of scenarios.

NREC order	Condition order set number			
3, 1, --	1	2	3	4
--, 3, 1	5	6	7	8
1, 3, --	9	10	11	12
--, 1, 3	13	14	15	16

Within a row of the table, the order of the NREC conditions is the same, and the order of other conditions varies according to the differences between the four prototype condition orders described in the table above—separately for the eight consecutive automation scenarios and the two consecutive control scenarios. In each row, the first cell (i.e., condition order 1, 5, 9, 13) presents the combinations of the reliability-, workload-, algorithm-, and scenario-type variables in the order listed for prototype 1 in the four tables in Table E-2 for the eight automation scenarios, and the workload and scenario type variables in the order listed for prototype 1. The second cell of each row (i.e., 2, 6, 10, 14) presents the combinations in the order used for prototype 2, and so on.

To assign the between-participants variable of SST, we ensured that two participants in each column and two in each row would receive SST. We selected the two participants in any column assigned to SST so as to have opposite NREC orders. We selected the two participants in any row assigned to SST so as to have opposite workload orders within their eight automation scenarios and within their two control scenarios, and opposite combinations of algorithm and scenario type for each combination of reliability and NREC. Finally, we present the colors used in the route map in Table E-4.

Table E-4 Route table and route map R/G/B values

	R	G	B
Gray	166	166	166
Light Green	0	255	100
Dark Green	0	150	0
Yellow	255	255	0
Orange	235	160	0
Red	255	0	0