Evaluation of Usability and Workload Associated with Paper Strips as Compared to Virtual Flight Strips Used for Ramp Operations

Victoria Dulchinos

SJSURF/NASA Ames Research Center
Moffett Field, CA, USA
victoria.l.dulchinos@nasa.gov

Abstract. This paper describes a study comparing the use of paper strips with virtual flight strips depicted on a new user interface, the Ramp Traffic Console (RTC), designed for use by ramp controllers to be used in place of paper strips. A Human-In-the-Loop (HITL) experiment was performed as the fifth in a series of six HITL simulation studies designed to evaluate a pushback Decision Support Tool (DST) concept for Charlotte Douglas International Airport (CLT). Workload and usability were assessed in post-run and post-study questionnaires. In the RTC virtual flight strip condition, post-run questionnaire results show lower workload ratings across all aspects of workload; additionally, a trend is found toward increased usability ratings. Post-study questionnaire results indicate a preference for RTC over paper strips. Additional research is suggested with more training runs and a greater number of participants to increase statistical power. It is also suggested that this new technology be re-evaluated as a part of the ATD-2 field testing activities.

Keywords: Human Factors · Human-systems Integration · Decision support tool · Usability · Workload

1 Introduction

New technologies developed for use by Air Traffic Controllers (ATC) and airline ramp operators are studied in a Human in the Loop (HITL) simulation study. The Ramp Traffic Console (RTC), shown in figure 1 below, was designed along with the Spot and Runway Departure Advisor (SARDA) Decision Support Tool (DST) proposed to aid ramp controllers in reducing taxi delay. SARDA was first evaluated as a decision support tool for air traffic controllers to meter flights from the spot to the runway (Hayashi et al, 2013).

Air Traffic Control Towers (ATCT) are equipped with multiple electronic systems that have been developed over time to facilitate controllers in the management of air traffic. Advanced Electronic Flight Strips (AEFS) is one such technology that is likely to be subsumed into Terminal Flight Data Management (TFDM) as a part of a larger effort to integrate multiple existing electronic systems. In a 2012 study of a prototype ATCT TFDM system, Controller-Pilot communications were used to measure

cognitive workload (Lockande, 2016). This study found that controllers utilizing the prototype TFDM system reported lower workload than the control group. While RTC is designed for use by airline operators, like AEFS and TFDM, RTC is intended to replace paper strips with a digitally integrated information source to present integrated flight data. In the current study, SARDA advisories are presented to the ramp controller as a tactical surface scheduler (DST) designed to meter flights from the gate. The RTC has a novel user interface displayed on a 27" multi-touch screen monitor, used by ramp controllers in place of paper strips and paper maps, and includes the SARDA pushback advisories.

During simulated operations, ramp controllers gave instructions to pilots via radio communications to manage traffic and ensure airplanes were safely separated while efficiently taxiing to their destination. This task required the controllers to engage in a variety of high-level cognitive functions, including planning, managing, monitoring, problem solving, and coordinating with other ramp controllers, pilots, and air traffic controllers. The CLT ramp is divided into four sectors, North, East, South and West, with most airplanes needing to taxi through multiple sectors. Ramp controllers hand off airplanes to each other at the sector boundaries. Handoffs are also made to air traffic controllers at various points, called spots, intersecting with the Federal Aviation Administration controlled active movement area on their way to and from the arrival or departure runway. Outbound departure flights are handed off to the Air Traffic Controller (ATC) at the spots and inbound arrival flights are received from the ATC at the spots and directed to their gate. Consequently, the ramp controllers were required to communicate with other sector controllers as well as air traffic controllers and multiple pilots to efficiently manage all the departure and arrival flights to and from their gates on the ramp. The RTC and SARDA concept were developed initially for use at the Charlotte Douglas International Airport (CLT).

The simulation study reported in this paper is one in a series of studies to evaluate SARDA and RTC from the ramp controller's point of view. Human-in-the-Loop (HITL) simulations are used as a safe and controlled environment to evaluate new concepts and decision support tools. The goal of the present study was to evaluate virtual flight strips on RTC as compared to the use of paper strips in ramp traffic management.

The research questions explored here are regarding the effect of using virtual flight strips on RTC as compared to using paper strips shown in Figure 2 below, on the workload and usability ratings of the ramp controller participants.

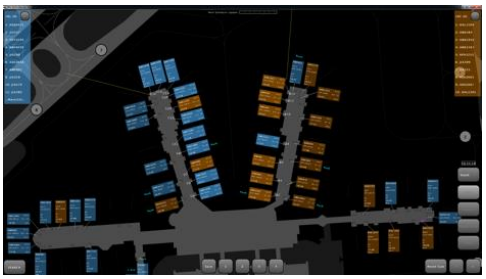

Fig. 1. RTC with virtual strips

Fig. 2. Paper strips and paper map

2 Methods

The virtual flight strips as presented on RTC were tested in a HITL simulation study in Future Flight Central (FFC), a high-fidelity tower simulator at NASA Ames Research Center. This study included eight 90-minute data collection runs over three days. There were two RTC training sessions for a total of 3 hours and 20 minutes of controller training using RTC. There were four ramp controller participants. In four of the data collection runs, the ramp controllers used paper strips and paper maps while controlling ramp traffic, and in the other four runs, the ramp controllers used the virtual flight strips on RTC. There were two traffic scenarios used in the simulation and each was repeated twice in the paper condition and in the RTC condition. Two participants were active ramp controllers from CLT, a third was a retired FAA controller, and the fourth participant was an active ramp controller from another airport. The four ramp controller participants used the RTC in the simulated ramp operations environment while usability and workload data was collected from the users under the two different conditions. In one condition the participants used paper strips and paper map, while in the second condition participants used the virtual flight strips and movable map on RTC. The two ramp controllers who were current CLT controllers were rotated through sector assignments such that each worked both scenarios in the paper and RTC conditions. The other two ramp controllers who were not active CLT controllers remained in one of the "less busy" sectors that were deemed to have less impact on the operation. Post-run and post-study workload and usability questionnaires were administered to all four of the sector controllers.

User workload is commonly assessed with subjective measures, which require the participants to report on their subjective psychological experience. These measures include self-reported subjective ratings on certain scales, such as the NASA Task Load Index (TLX) (Hart & Staveland, 1988). Workload for the purposes of the present study is defined by four components of the NASA-TLX (Task Load Index). The four components include Mental Demand (Thinking, deciding, calculating, searching, etc.), Physical Demand (Hands and arm movement, force), Temporal Demand (Time pressure), and Frustration (Stress, annoyance, irritation). Controllers were asked to rate each of the four components of their workload after every run on a scale of 1-10. For example, see Figure 3 for the "mental demand" question response format. A performance sub-scale was not included.

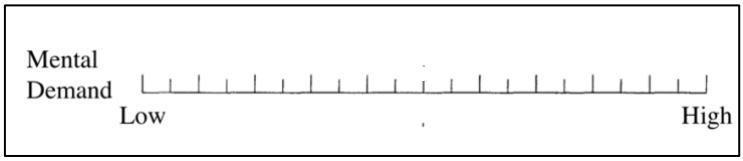

Fig. 3. Post Run Questionnaire Workload question format

Along with workload, usability of the RTC was also assessed. There are several definitions of usability (J. Jeng, 2005, provides a good review of various definitions). In this paper, the definition used by the International Organization for Standardization (ISO, 1998) will be followed. It defines usability as the extent to which the users of a product are able to work *effectively*, *efficiently*, and with *satisfaction.* Following the definition used by the International Organization for Standardization (ISO, 1998), usability for the purposes of this paper is defined by three aspects of usability, effectiveness, efficiency, and satisfaction. Traffic management performance questions were included in the post run questionnaire with the aim of determining the "effectiveness" aspect of usability. Resources and efficiency questions were included in the post-run questionnaire with the aim of determining the "efficiency" aspect of usability. The post-study survey questions were designed to assess the "satisfaction" aspect of usability. After each run, the controllers were asked questions regarding their traffic management performance and resources and efficiency using a response format with a scale of 1 "Referred to Always" to 7 "Referred to Never." For example, one Traffic Management and Performance aspect of Usability is assessed by the controller's response to the question shown in Figure 4 below:

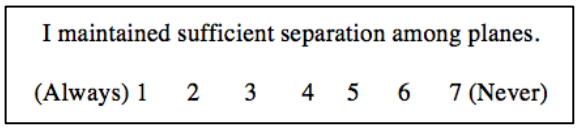

Fig. 4. Post-Run Questionnaire Usability question format

Post Run and Post Study questionnaire responses were gathered and the results were analyzed to assess controller workload and usability ratings under both conditions, virtual flight strips on RTC and paper strips. To determine the effect of condition (Paper or RTC) on controller workload and usability ratings, mean post run responses on the workload and usability related questions were collected from all four sector controllers and a two-way repeated measures Analysis of Variance (ANOVA) was performed to determine effect of flight strip type on participant workload and usability.

3 Results

The mean post run workload ratings and ANOVA results shown in Table 1 and are graphed with standard error bars at a 95% confidence level in Figure 5 below. These

results show that the mean workload ratings for the RTC condition are lower than the mean ratings for the Paper condition across all four components of workload. With respect to the Mental Demand aspect of workload, the participants reported a higher mean workload rating of 5.7 for the Paper condition as compared to a mean workload rating of 3.9 in the RTC condition however, as can be seen in Table 1, this was not a statistically significant main effect. There was a statistically significant main effect across the other three aspects of workload. With respect to the Time Pressure aspect of workload, the participants reported a higher mean workload rating of 4.9 in the Paper condition as compared to a mean rating of 2.4 in the RTC condition. With respect to the Physical Demand aspect of workload, the participants reported a higher mean workload rating of 4.6 in the Paper condition, and 2.8 in the RTC condition. Finally, looking at the Frustration aspect of workload, the participants reported a higher mean workload rating of 3.6 in the Paper condition, and 1.3 in the RTC condition.

Table 1. Mean Workload response all sectors

Mean Participant Workload Ratings Across Four Aspects of Workload					
Aspect of Workload	Mean Response Paper	SE Paper	Mean Response RTC	SE RTC	F(1,3)=
Mental Demand	5.7	0.82	3.9	1.67	3.59, p=.155
Time Pressure	4.9	0.57	2.4	0.50	48.46, *p=.006
Physical Demand	4.6	1.32	2.8	1.43	84.26, *p=.003
Frustration	3.6	0.31	1.3	0.34	29.73, *p=.012

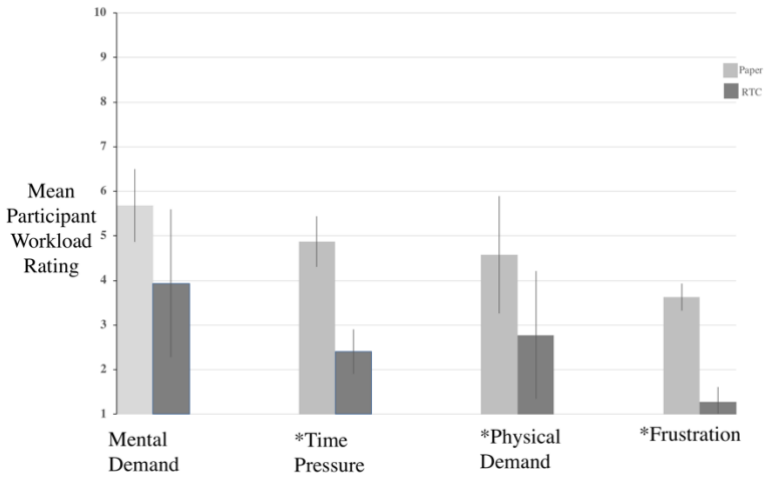

Fig. 5. Mean Participant Workload Rating

Because the response scale for the post run usability questions was presented in reverse order such that “Always” is the lower anchor (1) on the scale, and “Never” is the upper

anchor (7) on the scale, for ease of discussion, an inverse scale of the means is reported in this paper to account for the opposite phrasing of the questions.

The mean usability ratings of the post run traffic management and performance questions, meant to assess the "effectiveness" aspect of usability, were higher in the RTC condition as compared to the Paper condition for all of the seven questions. The means and standard errors are shown in Table 2 and graphed in Figure 6 below. The results of the analysis showed a statistically significant main effect of condition for questions 2, 3, and 5 that asked about "maintaining organized traffic flow," "minimizing taxi delay," and "maintaining pressure on the runways" respectively (see Table 2). Looking at question 2 which asked if the participant "maintained well organized traffic flows," the participants reported a higher rating of 6.6 for RTC as compared to a mean rating of 6.1 in the Paper condition. Looking at question number 3 which asked if the participant "minimized taxi delay of each aircraft," the participants reported a higher mean rating of 6.5 in the RTC condition than the mean rating of 5.9 in the paper condition. For question number 5 which asked if the participant "maintained pressure on the departure runways," the participants reported a higher mean rating of 6.7 in the RTC condition than the mean rating 5.9 in the paper condition.

All of the other traffic management questions had higher mean usability ratings in the RTC condition as compared to the paper condition, although this difference was not statistically significant (see Table 2). For question number 1 which asked if the participants "maintained sufficient separation among planes," the participants reported a higher mean rating of 6.9 for the RTC condition than the mean rating of 6.7 for the Paper condition. For question number 4 which asked if the participant "avoided sending airplanes into head on course or gridlock", the participants reported a higher mean response of 6.9 in the RTC condition than the mean rating of 6.6 in the Paper condition. For question number 6 which asked if the participant "metered their departures", the participants reported a higher mean response of 6.6 in the RTC condition than the mean response of 6.2 in the paper condition. Finally, for question number 7 which asked if the participant "responded to the pilots call promptly", the participants reported a higher mean response of 6.9 in the RTC condition than the mean response of 6.8 in the Paper condition. Looking at the results overall for the Traffic Management questions, there is a trend toward increased mean usability ratings in the RTC condition as compared to the paper condition for the traffic management and performance questions which were meant to assess the "effectiveness" aspect of usability, with the mean participants rating being higher in the RTC than the paper condition for all of these questions.

Table 2. Participant ratings of Traffic Management and Performance

Traffic Management Performance Questions Mean Response with Standard Error and F values					
Question	Mean Paper	S.E.	Mean RTC	S.E.	F (1,3)=
1. Maintained Separation	6.7	0.157	6.9	0.125	9, p=.058
2. Maintained Flow	6.1	0.373	6.6	0.295	12, *p=.04
3. Minimized Delay	5.9	0.329	6.5	0.25	22.09, *p=.018
4. Avoided Grid-lock	6.6	0.161	6.9	0.063	6.82, p=.088
5. Maintained Pressure on Runway	5.9	0.258	6.7	0.237	54, *p=.005

6. Metered Departures	6.2	0.493	6.6	0.12	.73, p=.456
7. Responded Promptly	6.8	0.188	6.9	0.063	.33, p=.604

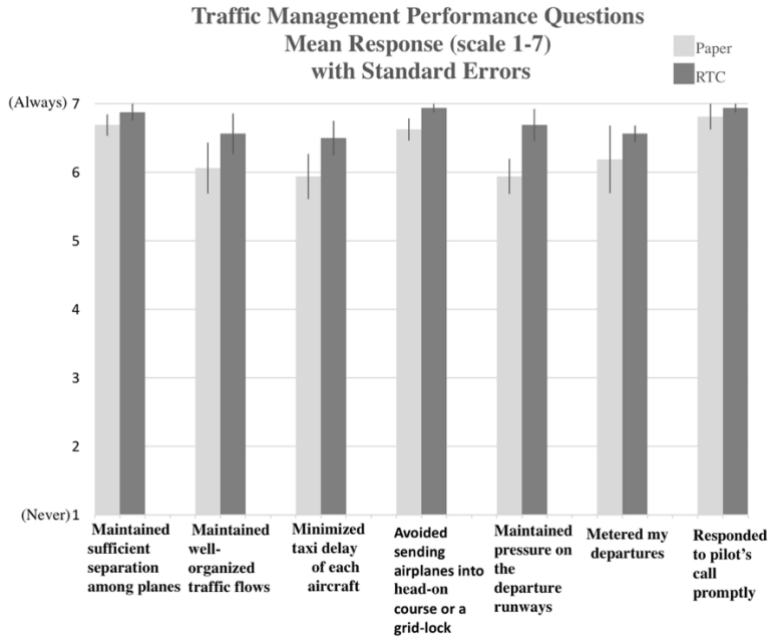

Fig. 6. Mean participant ratings of Traffic Management Performance

The mean participant response values for the post run usability resources and efficiency questions meant to assess the "efficiency" aspect of usability are shown Table 3 and graphed in Figure 7 below. The mean rating was higher in the Paper condition for questions 3 and 4, and the mean was the same for RTC and Paper conditions for question 6. However, none of these results demonstrated a statistically significant main effect of condition on participant usability ratings (See Table 3).

Questions 1, 2, and 5 of the resources and efficiency questions resulted in a higher mean rating in the RTC virtual strip condition as compared to the paper strip condition. Looking at the Resources and Efficiency question 1 which asked if "the information needed was easily accessible," the participants reported a higher rating of 6.0 for RTC virtual strips as compared to a mean rating of 5.4 in the paper strip condition. Similarly, looking at question 2 which asked if "the information was available but required some work to get to it," the participants reported a mean rating of 3.4 for RTC and 3.3 for Paper. Question number 5 asked the participants if "they collaborated with other controllers and took action to help them," the participants reported a higher rating of 6.9 in the RTC virtual strip condition as compared to a rating of 6.8 in the Paper condition. Questions 3 and 4 of the Resources and Efficiency questions the results show a higher mean rating in the Paper condition as compared to the RTC condition. Looking at question 3 which asked "if information need to keep track of held aircraft was

available," the participants reported a higher mean rating of 5.5 in the Paper condition as compared to the mean rating of 5.4 in the RTC condition. The Resources and Efficiency question 4 asked "if the actions required the minimum number of steps," with a higher mean participant rating of 5.4 in the Paper condition as compared to the mean RTC rating of 4.9. Finally, for question 6 which asked "if other controllers handled traffic in the way it was requested," the mean participant rating was the same in both Paper and RTC conditions with a mean rating of 6.88 for both RTC and Paper.

Table 3. Resources and Efficiency Mean Participant Resonse

Resources and Efficiency Questions Mean Response with Standard Error and F values					
Question	Mean Paper	S.E.	Mean RTC	S.E.	F (1,3)=
1. Information Accessible	5.4	0.904	6.0	0.654	1.86, p=.266
2. Information Available, but required work	3.3	0.753	3.4	0.74	.03, p=.878
3. Held Aircraft Information Available	5.5	0.729	5.4	0.582	.16, p=.718
4. Actions Required Minimum Steps	5.4	0.439	4.9	0.161	1.85, p=.267
5. Collaborated	6.8	0.25	6.9	0.125	.27, p=.638
6. Others Handled Traffic as Expected	6.9	0.125	6.9	0.125	0, p=1.0

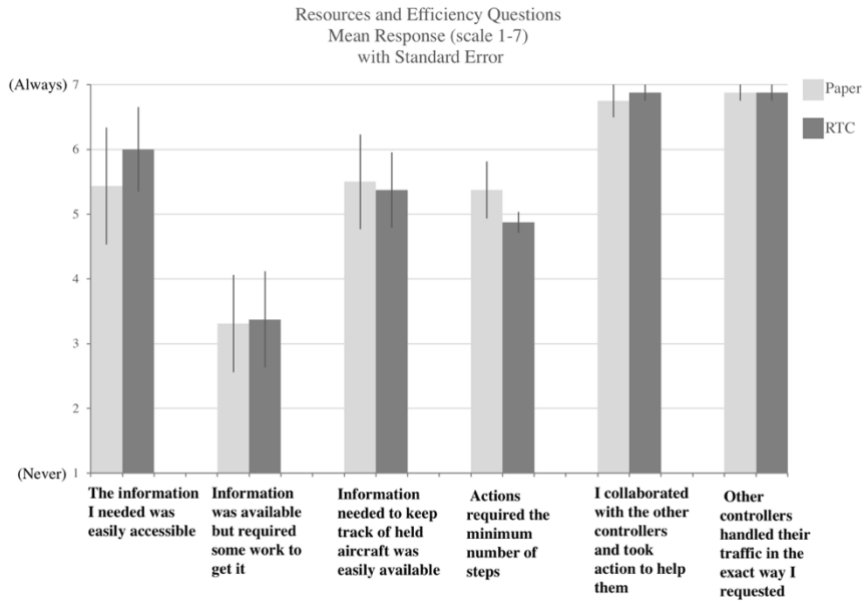

Fig. 7. Resources and Efficiency Participant Mean Response

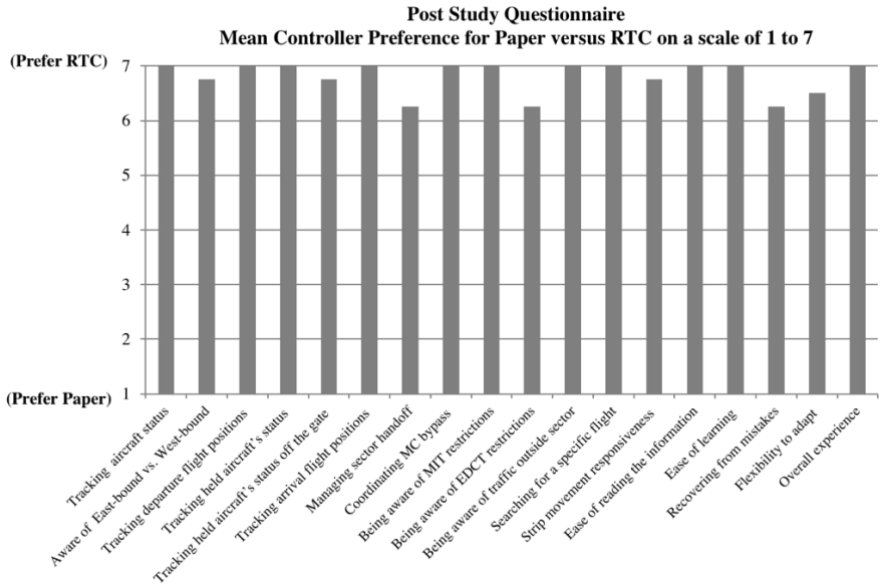

Fig. 8. Post study questionnaire mean participant satisfaction ratings

To assess the satisfaction aspect of usability, a set of 18 specific preference questions were included in the post study questionnaire. The responses were collected from all four controller participants with responses on a scale of 1 (Prefer Paper) to 7 (Prefer RTC). The results shown in Figure 8 above indicate that very high level of satisfaction ratings were achieved for all the questions ranging from tracking aircraft status, and being aware of the direction of the flight, to managing sector handoff to ease of reading of information.

In sum, results from the Post Run questionnaire indicate lower workload ratings for RTC condition, with only one of the workload elements not statistically significantly lower. Usability ratings for Traffic management performance questions are lower in the RTC condition than in the paper condition showing a preference for RTC over Paper, with not all of the questions showing a statistically significant difference. Usability ratings for Resources and efficiency questions showed mixed results. Post Study Usability responses and satisfaction ratings indicated a clear preference for RTC.

4 Discussion

The mean participant ratings for workload were lower in the RTC virtual strips condition as compared to the Paper condition for all four aspects of workload. There was a statistically significant main effect of condition for all aspects of workload measured except for the mental demand aspect of workload, which was similar for paper and virtual strips. It is possible that this mental workload result would decrease

with increased training and increased familiarity. The participants had a minimal amount of training with the RTC virtual strips prior to the data collection. The total amount of time spent training with RTC was 3 hours and 20 minutes; it is possible that with more time training the participants might have reported lower mean mental demand workload rating for RTC condition as compared to the Paper condition resulting in a statistically significant main effect. The participants in this study had been using only traditional paper strips to manage traffic in their experience as professional ramp and air traffic controllers, and RTC was a new tool. The participant ratings for mental demand aspect of workload were lower in the RTC condition than in the paper condition, however this was not a statistically significant difference, perhaps more time training in preparation for the data collection runs, or a greater number of data collection runs might have allowed the participants to gain more experience with the tool resulting in a decrease in the mental demand aspect of workload of using the RTC virtual strips to perform their role as ramp controllers in the HITL. Also, due to the nature of the simulation study with a limited number of controller positions and a limited number of data collection runs, there were only four participants and only eight 90-minute data collection runs. Perhaps, future studies might include a greater number of participants and or data collection runs, thereby increasing the statistical power of the study.

The participant ratings for the "effectiveness" aspect of usability were higher in the RTC virtual strip condition than the Paper condition for all of the Traffic Management Performance questions, with statistically significant results for some of these questions. The trend shows that RTC was more efficient than paper on all questions except for two. The lower RTC rating regarding managing the strips was possibly due to lack of familiarity and usage; potentially the participants did not perceive a difference in the efficiency between the two conditions (RTC virtual strips and Paper strips) or the lack of sufficient data in this study.

Looking at the results of the Resources and Efficiency questions in relation to the results of the Traffic Management questions, the Traffic Management questions received a more consistently favorable and statistically significant positive rating for RTC than the Resources and Efficiency questions, perhaps the participants found using the RTC virtual strips to be more effective than using the paper strips. At the same time, these results might be interpreted to indicate that for some aspects of efficiency, the results were not a clear indication of a preference for RTC. Again, perhaps this is a function of the participants being new to the RTC virtual strips and given more time and experience using the RTC virtual strips, the participants rating of the efficiency aspect of usability might improve. Participants' ratings from the post study questionnaire for the "satisfaction" aspect of usability indicate a definite preference for the RTC over the Paper condition. Overall these results indicate a trend towards increased mean participant Usability ratings when using the RTC virtual strips as compared to using the paper strips across the three aspects of Usability assessed: effectiveness, efficiency and satisfaction.

As in the TFDM prototype system study by Lockande (2012), the workload results from the current study indicate reduced workload in the RTC virtual strip condition as compared to corresponding baseline or paper strip condition. Similar to the Lockande (2012) study, one possibility is that a reduction in workload is a function of the RTC displaying data on the virtual flight strips that is digitally updated. Like the TBFM prototype used by Lockande, the RTC also integrates other operational data and

presents it to the ramp controller in real time such that the ramp controller is not seeking out and verifying information regarding, for instance, Traffic Management Initiatives, or airport configuration, thereby reducing overall workload. The workload results indicating reduced Workload when using RTC along with the Usability results indicating a trend toward increased Usability when using RTC seem to indicate that the participants favored the RTC virtual strips as compared to the Paper condition. Future studies of the RTC may benefit from more training runs, as well as having either a greater number of participants or a greater number of data collection runs to increase the statistical power of the analyses.

Recently, RTC has undergone a design refactoring, removing the touch capability, and going to a mouse only design. This refactoring was prompted by a couple of reasons. During the HITL testing of RTC, feedback from some of the controllers indicated that they prefer using the mouse over touch screen functionality. Also, it was decided to a larger 32' screen size for screen sharing with another technology in the field. Going to a larger screen meant possible degradation of touch screen precision along with possible increased fatigue while using the larger display. The controller feedback information along with deciding to go to a larger screen size resulted in the decision to go to a mouse only design. The SARDA tactical surface scheduler has also undergone some development and maturation as it has been integrated along with the RTC with a set of other Air Traffic Management Technologies as a part of NASA's ATD-2 effort (Malik et al, 2016). The ATD-2 Phase One field testing began in September of 2017 where RTC is currently in use by ramp controllers at CLT. Given that additional development and maturation has been completed on the RTC and the tactical scheduler tool, it will be important to follow up on this study to determine the impact of this refactoring on ramp controller user workload and usability ratings.

Acknowledgements

The author acknowledges the work of the team of people who made this research possible. I express my special thanks to Miwa Hayashi, Yoon Jung, Savita Verma, Katherine Lee and Victoriana Delosantos.

References

1. Hart, S.G., and Staveland, L.E. (1988). Development of a NASA-TLX (task load index): results of empirical and theoretical research. In P.S. Hancock, & N. Meshkati, *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier Science Publishers B.V.
2. International Organization for Standardization (1998), "Ergonomic requirements for office work with visual display terminals (VDTs)—part 11: guidance on usability," ISO 9241-11, Geneva, Switzerland.
3. Jeng, J., "Usability assessment of academic digital libraries: effectiveness, efficiency, satisfaction, and learnability," International Journal of Libraries and Information Services, vol. 55, pp. 96–121, 2005.
4. Lokhande, K., Reynolds, H.J., "Cognitive workload and visual attention analyses of the air traffic control tower flight data manager (TFDM) prototype demonstration", 2012 Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting.

5. Hayashi, M., et al (2013). “Usability Evaluation of the Spot and Runway Departure Advisor (SARDA) Concept in a Dallas/Fort Worth Airport Tower Simulation,” ATM Seminar 2013.
6. Malik, W.A., Lee, H., and Jung, Y.C., "Runway Scheduling for Charlotte Douglas International Airport," AIAA-2016-4073, 2016 AIAA Aviation and Aeronautics Forum and Exposition, Washington D.C., 13-17 June 2016.