Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013)

Usability Evaluation of the Spot and Runway Departure Advisor (SARDA) Concept in a Dallas/Fort Worth Airport Tower Simulation

Miwa Hayashi, Ty Hoang, Yoon C. Jung
NASA Ames Research Center
Moffett Field, CA, USA

Gautam Gupta, Waqar Malik
University of California, Santa Cruz
NASA Ames Research Center
Moffett Field, CA, USA

Victoria L. Dulchinos
San Jose State University
NASA Ames Research Center
Moffett Field, CA, USA

***Abstract*—Spot and Runway Departure Advisor (SARDA) is a decision-support tool proposed to aid air traffic control tower controllers in reducing taxi delay and optimizing the runway sequence. The purpose of the present paper was to evaluate the tool's usability to ensure that its claimed performance benefits are not being realized at the cost of increasing the work burden on controllers. The study analyzed workload ratings and questionnaire responses collected during a human-in-the-loop simulation experiment and assessed the effects of the SARDA advisories on the controllers' cognitive resources (e.g., workload, spare attention) and satisfaction. The results showed that SARDA reduced the controllers' workload and increased their perceived spare attention. SARDA also made workload and attention levels less susceptible to the effects of increases in the traffic load. The questionnaire responses suggested that the controllers generally were satisfied with the ease of use of the tool and the intended benefits of the SARDA concept, but with slight reservations. Sharing high-level reasoning behind SARDA's optimization process with the controllers may help the concept to gain more trust from them.**

Keywords—workload; attention; user interface; Electronic Flight Strips; Traffic Management Initiatives

I. INTRODUCTION

A. Background

In most U.S. airports, departure flights are released to taxi out primarily on a first-come-first-served (FCFS) basis. Even when the airport is already crowded and the anticipated wait time by the runway is long, pilots still need to call in and taxi out to obtain a departure slot. This scheme often worsens the congestion on taxiways and in runway queues. It also causes frequent stop-and-go movements and long waits in runway queues that waste onboard fuel and pollute the surrounding air. Furthermore, in this FCFS taxi-out scheme, the actual departure order is not known until the moment each aircraft takes off. This reduces the precision of en-route and arrival traffic flow management planning throughout the entire airspace and potentially leads to missed slots.

B. Goals of SARDA

To mitigate these inefficiencies, NASA Ames Research Center (ARC) has been developing Spot and Runway Departure Advisor (SARDA), a decision-support tool for air traffic control tower (ATCT) controllers [1]. It calculates the optimal runway-use sequence and assists the controllers in metering the departure flights accordingly at the *spots*, the points where each departure aircraft enters from a ramp to an airport movement area. After queues are formed at the runway, the controller releases each aircraft for takeoff in the order prescribed by SARDA. Small queues are always maintained at the runway so as to prevent the runway from being underutilized, yet keep the wait times in the queue reasonably short.

The runway sequence that SARDA calculates is conditioned to comply with the minimum wake-turbulence separation constraints and all the Traffic Management Initiative (TMI) constraints in effect, such as Miles-in-Trail (MIT) restrictions, Expect Departure Clearance Time (EDCT), and Call for Release (CFR). It also intentionally creates gaps between departures to allow ground traffic to cross the departure runway. For traffic control on an active runway, the controller receives only the sequence advisory from SARDA. The timing advisory accompanying the sequence is hidden to avoid interfering with the controller's responsibility to release each aircraft for takeoff or for crossing the runway only when it is safe for the aircraft to do so.

The key system performance goals of SARDA are to:

- Reduce total taxi delay and total fuel burn (from gates to takeoff)
- Improve conformance with the assigned TMI takeoff-roll times
- Maximize runway throughput

C. Purpose of the Paper

The above goals should not come at the expense of increased controllers' work burden. If the new concept negatively impacts controllers' workload and process, the controllers' and stakeholders' communities may not accept it, regardless of the magnitude of the system-performance benefits it may offer.

In May 2012, a SARDA human-in-the-loop (HITL) simulation experiment was conducted at a tower simulator

facility at NASA ARC with the participation of retired ATCT controllers to evaluate the feasibility of the concept. The results of initial analyses [2] showed that use of the technology led to improvement in the system performance. The present paper reports the usability assessment of the SARDA tool from the controllers' point of view. Subjective data collected during the HITL simulation study were analyzed for this purpose.

D. Previous Work for SARDA Concept Development

SARDA was implemented within the Surface Management System (SMS) [3]. SARDA uses the Spot Release Planner (SRP) as the core scheduler engine [4]. It consists of two stages. The first stage, the runway scheduler [5-6], computes the optimal departure sequence, release time from the runway, and runway-crossing schedule. The optimization results are then sent to the second stage, which calculates the optimal spot-release time and the runway-queue selection (if there are multiple runway queues). The second stage presents the advisory information to the Ground controller (who controls all traffic in the movement areas not being controlled by the Local controller). The advisories for the Local controller (who controls all traffic that uses the active runways) are supplied by the runway scheduler.

The HITL simulation study reported in this paper was a follow-up study of the initial SARDA HITL simulation conducted in April 2010 [7-8]. In this 2010 study, the Dallas/Fort Worth International Airport (DFW) East tower operations were simulated with participation of two retired DFW ATCT controllers. The results validated the initial concept: when the SARDA advisories were provided, the taxi delay in the movement area was reduced by 64%, and fuel consumption was decreased by 38% in heavy-traffic scenarios. However, the advisories simply moved the delays from the runway side to the ramps, and caused significant ramp-area congestion. This result suggested that, for SARDA's benefits to be realized, a departure aircraft that is still early for its taxi-out time needs to be held at its gate with its engines off. The study also compared two display formats for the SARDA advisories: data tag and timeline. The results showed that the Ground controller preferred the timeline format. (No format preference was found in the Local controller.) The controllers also suggested switching to an Electronic Flight Strips (EFS) format that resembles the paper flight progress strips commonly used at many ATCTs.

In the present HITL simulation, the following significant upgrades from the initial simulation [7-8] were made:

- A rudimentary gate-holding scheme was implemented: the aircraft pushed back from their gates so that they reached the spots at most a minute before the SARDA-scheduled spot-release time [9].
- A computer-generated out-the-window (OTW) view was added to the simulation to increase realism, allowing realistic scan patterns for the controllers.
- An EFS format was developed and provided to the controllers on a 24-inch touch-screen display.
- Taxi-speed uncertainty (12 to 17 knots) was added to increase realism in the traffic data.
- The number of controller participants was increased from two to six to reduce the influence of potential individual biases.

E. The Electronic Flight Strips (EFS) Format

The SARDA concept requires advisories to be dynamically updated as the state of airport traffic evolves and the advisories are re-calculated. SARDA also can use additional inputs from the controller, such as issuance of a taxi clearance or a takeoff clearance, to reflect the most recent situation in its scheduling optimization. These requirements necessitate the use of the Electronic Flight Data System (EFDS) [10] as the user interface. The EFS, data tag, and timeline formats are all examples of EFDS. The EFS format was chosen based on feedback in the previous study. Using an electronic format also makes it easy to share the information with other positions within the tower, as well as with other air traffic control (ATC) facilities [11], providing steps toward better traffic planning and coordination in the national airspace [12].

The EFS format of the Surface Decision Support System (SDSS) developed by Mosaic ATM, Inc., was used as the base design for the EFS format for the current HITL simulation, and the strip design was adjusted to fit the study's purpose. The strip design was, then, polished heuristically through multiple rounds of test sessions with ATCT controller subject matter experts (SMEs).

Strictly speaking, the EFS developed for this study was not a part of the official SARDA package, but rather a necessary technology to conduct a HITL simulation evaluation. (In future, SARDA could be using a different EFS system.) Still, the current EFS design may affect the evaluation results. To minimize the effects, the same EFS was used in all runs. In this way, the effects of the EFS would be mostly canceled out. Questions about the EFS usability were included in the post-study questionnaire to obtain controllers' feedback after they completed all the runs.

F. Usability Evaluation Approach

To assess usability, first the product's users, the goals, and the environment need to be specified. These elements are relatively well defined in the SARDA concept. For general SARDA operations, the users are certified ATCT controllers, the goal is to achieve the SARDA goals (listed in Section B) while performing regular ATC operations at the normally expected safety and efficiency levels, and the environment is regular ATCT operations at a major airport. However, due to simulation resource limitations, the actual scope of the study was more limited. In the present study, the users were retired ATCT controllers, the goal was the same as the above, and the environments were four nominal traffic scenarios at a single airport. One should use caution when extending the usability-evaluation results derived from this study to users, goals, or environments substantially different from those used in the simulation (e.g., much younger controllers, off-nominal operations, other airports).

There are several definitions of usability ([13] provides a good review of various definitions). In this paper, the definition used by the International Organization for Standardization (ISO) 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* [14]. For the effectiveness aspect (i.e., accuracy and completeness), the study referred to the prior analyses results of the system performance data [2] to examine the intended performance goals were achieved. The primary foci of the present study were the other two aspects: efficiency (i.e., resource availability to achieve the goals) and satisfaction. The subjective data, such as workload ratings and questionnaire response, were analyzed to investigate these aspects of SARDA usability.

II. METHODS

A. Simulation Facility

The FutureFlight Central (FFC) ATCT simulation facility [15] was used for the study. This tower cab simulator is equipped with a 360-degree computer-generated out-the-window view projected onto twelve 10-foot by 7.5-foot screens (Fig. 1). NASA's in-house software called Airspace Traffic Generator (ATG) was used to generate aircraft track and flight-plan information, based on which the SMS generated the SARDA advisories using the SRP algorithms. The ATG updated the track information on a 1 Hz cycle, and the SMS updated the advisories every ten seconds. The advisories then were presented to the controller participants on the EFS. The controllers gave instructions to confederate pseudo-pilot participants, who controlled multiple aircraft on the airport surface areas, via voice radio. The pseudo pilots then input the new aircraft-movement commands into the ATG via its pseudo-pilot interface so the next aircraft track update would reflect these changes.

Figure 1. FFC tower cab simulator
(Ground position on the left, and Local position on the right)

B. Traffic Scenarios

The East-side traffic of DFW in the South-flow configuration was simulated, where runway 17R was used for departures and 17C for arrivals (Fig. 2). Arrival traffic on runways 17L and 13L was not simulated to avoid a need for staffing another Local position. Bridge traffic to and from the West side of the airport also was excluded to simplify concept feasibility evaluation without loss of generality.

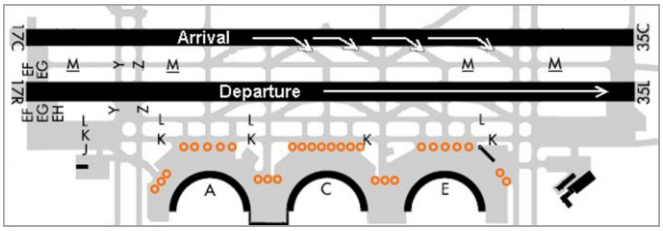

Figure 2. Simulated traffic flows and spots (orange circles)

Four traffic scenarios used in the study—called *M1*, *M2*, *H3*, and *H4* in this paper—are listed in Table I. Note that the runs were usually terminated before all of the departures took off, so, the actual run lengths were shorter—about 35 and 45 minutes in the medium- and heavy-traffic scenarios, respectively.

TABLE I. FOUR TRAFFIC SCENARIOS

Scenario Labels	Traffic Level	Total Length	Departures	TMI Flights
M1, M2	Medium (1.2× current-day[a])	45 min	40	5
H3, H4	Heavy (1.5× current-day[a])	50 min	50	7

a. Estimated using the live DFW traffic data recorded in January 2012.

C. Participants

Six retired DFW ATCT controllers were recruited for the study. The total ATCT experience per participant ranged from 27 to 39 years (mean = 31.3 years). The total DFW ATCT experience ranged from 12 to 30 years (mean = 19.0 years), and the number of years since the last control at the DFW tower was 6 or less. All controllers had experience in handling flights with TMI constraints, including MIT, EDCT, and CFR. None had prior knowledge of the SARDA concept. In each week, two controllers participated in the simulation and took turns playing the role of the Ground or Local controller. Aside from the six participant controllers, two retired ATCT controllers joined the study as confederate SME observers, who helped participant controllers in the hands-on training sessions and, once the data-collection runs started, closely observed the controllers' performance. Six pseudo-pilot positions were staffed by commercial or private pilots.

D. Controller Interface

The DFW East tower is located across runway 17C from the terminals. The Ground controller was positioned directly facing the airport's C terminal, and the Local controller was positioned on the right-hand side of the Ground controller, facing toward the runway 17R queues. This section describes the controller interface provided at each of these positions. EFS were the main controller interface for SARDA and will be described in detail.

The Ground controller was provided with two displays (Fig. 3): a map display (left) and an EFS display (right). The map display mimicked the Airport Surface Detection Equipment, Model X (ASDE-X) display and showed the aircraft-location symbols superimposed on the map diagram, with data tags indicating the aircraft call sign (e.g., "AAL123"). Departure aircraft symbols were colored green and arrivals were in white for easier distinction.

The EFS were displayed on a 24-inch touch-screen monitor. An optional stylus was provided to improve touch accuracy, and all controllers decided to use it instead of their fingertips. Strips were organized in virtual bays, each of which represented a clearance status and/or associated runway to use. The Ground EFS contained the following five bays:

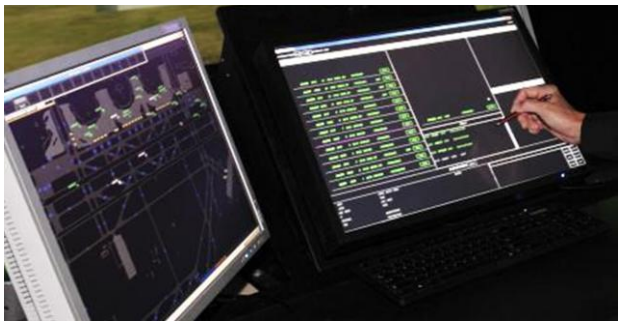

Figure 3. Map (left) and EFS (right) displays for the Ground controller position

i) *East Ramps – Departure* (departures not yet cleared to taxi out from the spots)
ii) *Taxi – Departure* (departures cleared to taxi out but not yet handed off to the Local)
iii) *Arrival* (arrivals that have crossed the departure runway 17R and have been handed off from the Local but not yet cleared to taxi back to the ramps)
iv) *Taxi – Arrival* (arrivals cleared to taxi back to the ramps)
v) *Drop List* (a list of both departures and arrivals recently handed off to the Local or ramps)

In nominal operation simulated in this study, all departures moved from i), to ii), then v), and all arrivals moved from iii), to iv), then v). In each bay, new strips were added to the top of the stack, and the controller worked from the bottom of the stack. Each strip had a *default action button* on the right side, and touching that button automatically moved the strip to the next bay. Strips also could be moved manually to any position as long as the move was not prohibited by the system (e.g., a departure was prohibited from moving to an arrival bay). The controller moved a strip in two steps: (1) touching the strip to *select* it, and (2) selecting a destination. Drag-and-drop strip-move capability was deactivated to prevent accidentally dropping a strip in the middle of a drag movement.

Fig. 4 shows example strips in the *East Ramps – Departure* bay. All strips in this bay are departures. Starting from the left, each strip showed the aircraft call sign (e.g., "DAL859"); the aircraft type ("B737"); the SARDA advisories, including a spot-release sequence ("3") and the count-down timer ("02:15"); the spot number/taxi route ("S45/K…EH," where "…" meant the taxi routes were too long to be shown, and only the first and last taxiways were displayed); the departure runway/first fix/destination ("17R/CLR/BTR"); the assigned TMI takeoff-roll time (if any) highlighted in an amber solid box ("2245"); and the default action button, which automatically moves the strip to the next bay, in this case the *Taxi – Departure* bay ("TX-D"). The SARDA sequence and the countdown timer were surrounded by a blue box outline when the countdown time was dropped below "01:00" (1 minute). They were then highlighted by a green solid box when the time fell within "00:30" (30 seconds). A negative countdown time meant it had passed "00:00," i.e., the advised spot-release time. In the runs during which the SARDA advisory was not provided, the SARDA advisory columns were

Figure 4. Ground EFS, East Ramps – Departure bay

left blank. The strips were designed to show minimum amount of information to reduce visual clutter. All information about the flight was available in the readout pane in the lower left corner of the display when the strip was selected.

The Local controller used three displays: the map display (left), the EFS display (center), and the radar display (right). The map display was similar to the one used by the Ground controller. The radar display emulated the same radar displays used in the Terminal Radar Approach Control (TRACON) facilities. The Local controller used this display to see the locations of airborne arriving and departing traffic.

The EFS for the Local controller also were presented on a 24-inch touch-screen display and contained the following bays:

i) *17R* (departures from 17R not yet cleared for takeoff, and arrivals on the ground that need to cross 17R but are not yet cleared to do so)
ii) *17R – Clear for Takeoff* (departures from 17R cleared for takeoff but not yet handed off to the TRACON)
iii) *17C* (arrivals landing on 17C)
iv) *Drop List* (both departures and arrivals recently handed off to the Ground or TRACON).

All departures moved from bays i), to ii), then iv), and all arrivals moved from iii), to i), then iv). Notice that both departures and arrivals entered bay i), the *17R* bay, where the arrivals crossed the departure runway.

Fig. 5 shows example strips in the 17R bay. The strips with green text (or black text on a colored strip) are departures, and those with white text are arrivals, using the same color-coding scheme as in the map display. The departure strips are also longer than the arrival strips. Both departure and arrival strips contain the aircraft call sign, the aircraft type, and the SARDA sequence advisory information. If multiple arrivals showed the same sequence number, they were assigned to cross during the same departure gap. Notice that there was no countdown timer advisory information on the Local EFS—the Local controller

Figure 5. Local EFS, 17R bay

was responsible for deciding safe timings. The remaining information on the departure strips included the assigned runway queue (e.g., "EH"), the first fix/destination, the assigned TMI takeoff-roll time (if any) highlighted by an amber solid box, and two default action buttons: *Line Up and Wait* ("LUAW") and *Clear for Takeoff* ("CFTO"). When the aircraft was instructed to line up and wait, the controller pressed the LUAW button, and that highlighted the strip in pale-green color as a visual reminder for the controller (see DAL138 in Fig. 5). When takeoff clearance was issued, the controller pressed the CFTO button, and the strip was sent to the next bay, *17R – Clear for Takeoff*. In the arrival strips, the spot to go to (e.g., "S24") and a default action button, "E GND," which automatically sent the strip to the (East) Ground controller's EFS, were provided.

E. Experimental Design

The study conducted 48 data-collection runs in three weeks. The primary independent variables were Advisory (Advisory vs. Baseline runs), Position (Ground vs. Local), Scenario (four scenarios with two levels of traffic volume), and Participant (six controllers). For the real-time workload rating data only, Phase (a 10-minute segment within a scenario; three segments in a medium-traffic run, or four in a heavy-traffic run) also was included in the independent variable. The test matrix was designed to counterbalance potential learning or fatigue effects within each participant.

The subjective data analyzed to evaluate the SARDA tool's usability were as follows. For the efficiency aspect, real-time workload ratings measured every five minutes (by a method similar to the Air Traffic Workload Input Technique [16] using the Workload Assessment Keypads [17]), NASA Task Load Index (TLX) workload ratings [18] collected at the end of each run, and other post-run questionnaire responses were analyzed to assess the amount of internal resources (e.g., workload, spare attention) the controllers *felt* they had during each run. For the satisfaction aspect, the controllers' responses to the post-run and post-study questionnaires regarding their subjective judgment on the helpfulness of the SARDA advisories, ease of use of the user interface, etc., were examined. Aside from the subjective data, as mentioned before, the previous analysis results of system-performance data [2] needed to evaluate the effectiveness aspect of usability also are summarized briefly in the next Results section.

The comparison between the Advisory and Baseline runs (i.e., the Advisory effect) was the main interest of this study. It should be noted that the Advisory runs provided not only the SARDA advisories that were unavailable in the Baseline runs, but also additional capabilities that were afforded by the presence of the target departure schedule and, thus, could not be included in the Baseline runs. These capabilities included i) gate holding, ii) automatic strip sorting by the SARDA sequence advisory, and iii) automatic assignment of a departure runway queue (i.e., EF, EG, or EH). Such capabilities likely offer some convenience to the controllers, and one may argue that they may add unfair advantages in the Advisory runs. The SARDA team considered it still fair to include them in the evaluation because these were all natural extensions of the utility of the target departure sequence computed in SARDA.

In the Baseline runs, the controllers were asked to control the traffic in a way they normally would. On the other hand, in the Advisory runs, they were asked to follow the SARDA advisories as closely as they could. (The actual SARDA concept of operation allows controllers to deviate from the advisories, but in this study, the researchers needed to study what would happen if all advisories were followed. This point will be revisited later.)

In both Advisory and Baseline runs, the controllers were instructed to try to meet the assigned TMI takeoff-roll time within a ±1-minute window. Even when the time window was going to be missed, no new time was assigned to the flight, and the controllers were asked to do their best to have the aircraft depart as close to the assigned time as possible.

III. RESULTS

A. Real-Time Workload Ratings

The real-time workload rating data were analyzed with a repeated-measures analysis of variance (ANOVA). A rating scale of 1 (lowest workload) through 7 (highest) was used. The main effects included in the model were Advisory, Position, Phase, Scenario, and Participant. Two-way interaction effects involving Advisory effect (e.g., Advisory × Position, etc.) were also included in the model. Scenario and Participant effects were treated as random effects, as the particular participants and scenarios selected for the study were considered to be sampled from a larger population [19]. The other effects were treated as fixed effects. Since there were two random effects in the model, quasi-*F*-ratios (denoted as *F**) were used for *F*-tests for all the fixed effects. Data collected from medium-traffic runs and heavy-traffic runs were analyzed separately because of different numbers of phases.

Advisory effect was found to be only marginally significant (i.e., had a *p*-value slightly greater than the conventional threshold of 0.05) in both the medium- and heavy-traffic runs (in medium traffic, $F^*_{1.04,\ 5.46} = 5.03$, $p = 0.070$; in heavy traffic, $F^*_{1.03,\ 5.92} = 5.47$, $p = 0.058$). Figs. 6 and 7 show that the workload ratings in both traffic-level scenarios tended to be lower in the Advisory runs than in the Baseline runs. The plots also show that the mean workload ratings stayed about the same (~2.2) between the two traffic levels in the Advisory runs, whereas in the Baseline runs, the ratings slightly increased when traffic was increased (from 2.6 to 2.9). The magnitudes of these differences were relatively small, however.

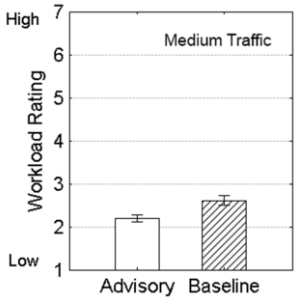

Figure 6. Means and standard errors of real-time workload ratings in medium traffic

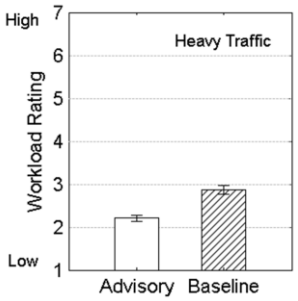

Figure 7. Means and standard errors of real-time workload ratings in heavy traffic

Position effect was found to be statistically significant only in the medium-traffic runs ($F^*_{1.12, 5.88}$ = 8.41, p = 0.026); the ratings tended to be higher in the Local position than in the Ground position, though the absolute differences in the mean ratings between the Ground and Local positions were small (2.3 vs. 2.5, respectively). The ratings also showed statistically significant Participant main and interaction effects in both traffic levels (in medium traffic: $F_{5, 5}$ = 175, p < 0.001 in Participant; $F_{5, 5}$ = 10.3, p = 0.011 in Advisory × Participant; in heavy traffic: $F_{5, 5}$ = 336, p < 0.001 in Participant, $F_{5, 5}$ = 9.75, p = 0.013 in Advisory × Participant).

B. NASA TLX Workload Ratings

NASA TLX workload ratings included the following ratings: Temporal Demand, Frustration, Performance, Effort, Physical Demand, and Mental Demand. The scales were 1 through 10, where 1 corresponded with the lowest workload and 10 the highest, except for the scale for Performance, where 1 corresponded with the poorest performance and 10 the most successful performance. The data were subjected to a repeated-measures ANOVA similar to the one used for the real-time workload rating analysis. The model included Advisory, Position, Scenario, and Participant main effects, and two-way interaction effects involving Advisory effect. Scenario and Participant effects were treated as random effects. In this analysis, the data from the medium- and heavy-traffic runs were analyzed together.

Advisory effect was found to be statistically significant in Temporal Demand ($F^*_{1.01, 6.91}$ = 6.35, p = 0.040), Effort ($F^*_{1.01, 6.89}$ = 6.04, p = 0.044), Physical Demand ($F^*_{1.02, 6.80}$ = 5.97, p = 0.045), and Mental Demand ($F^*_{1.02, 7.07}$ = 5.78, p = 0.046). Fig. 8 shows the means and standard errors of these ratings. In all of the four types of ratings, the mean ratings were lower in the Advisory runs than in the Baseline runs by about 2 points (2.0 vs. 4.0 in Temporal Demand, 2.3 vs. 4.3 in Effort, 2.1 vs. 3.9 in Physical Demand, and 2.5 vs. 4.5 in Mental Demand).

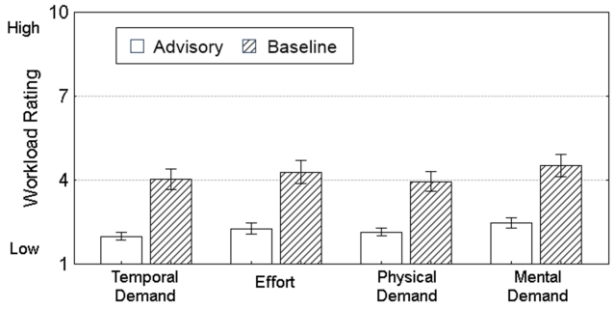

Figure 8. Means and standard errors of TLX ratings by Advisory

Advisory × Scenario interaction effect also was found to be statistically significant in Temporal Demand ($F_{3, 15}$ = 4.39, p = 0.021) and Effort ($F_{3, 15}$ = 4.24, p = 0.023), and marginally significant in Physical Demand ($F_{3, 15}$ = 3.17, p = 0.055) and Mental Demand ($F_{3, 15}$ = 3.19, p = 0.055). Planned-comparison analyses [19] revealed that the contrasts in the Advisory × Scenario effects in all of these four ratings were statistically significant between the medium- and heavy-traffic scenarios ($F_{1, 15}$ = 12.4, p = 0.003 in Temporal Demand; $F_{1, 15}$ = 12.6, p = 0.003 in Effort; $F_{1, 15}$ = 4.89, p = 0.043 in Physical Demand; and $F_{1, 15}$ = 8.57, p < 0.001 in Mental Demand). Since no significant contrast was found in the Advisory × Scenario effect within the two medium-traffic scenarios (M1 vs. M2) or within the two heavy-traffic scenarios (H3 vs. H4), the ratings from the two scenarios of the same traffic levels were consolidated. Fig. 9 plots the means of the four types of ratings by Advisory and by traffic level. The graph indicates steeper slopes in the Baseline data (dotted lines) than in the Advisory data (solid lines). Notice also that the Advisory data were all almost flat between the medium- and heavy-traffic levels.

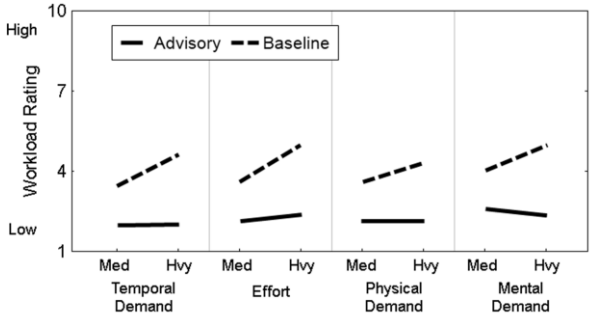

Figure 9. Means of TLX ratings by Advisory × Traffic Level (Med = Medium traffic, Hvy = Heavy traffic)

Position effect was found to be statistically significant only in the Temporal Demand ratings; the controllers felt more time pressure in the Local position than in the Ground position ($F^*_{1.23, 7.84}$ = 5.10, p = 0.050). Scenario effect was found to be statistically significant in Frustration ($F_{3, 15}$ = 4.38, p = 0.021) and Effort ($F_{3, 15}$ = 5.62, p = 0.009), and planned-comparison analyses showed that, in both ratings, the contrasts were statistically significant between the medium- and heavy-traffic scenarios ($F_{1, 15}$ = 12.8, p = 0.003 in Frustration; $F_{1, 15}$ = 14.4, p = 0.002 in Effort), but not within the same traffic-level scenarios. Participant effect was found to be statistically significant in all of the six ratings ($F_{5, 15}$ = 55.7 in Temporal Demand, 72.2 in Frustration, 15.3 in Performance, 45.8 in Effort, 50.5 in Physical Demand, and 52.4 in Mental Demand; all p < 0.001). Advisory × Participant effect also was found to be statistically significant in all ratings but Performance ($F_{5, 15}$ = 19.6 in Temporal Demand, 49.1 in Frustration, 19.2 in Effort, 15.3 in Physical Demand, and 12.8 in Mental Demand; all p < 0.001).

C. Post-Run Questionnaire Responses

In the post-run questionnaire form, 13 to 15 questions were asked, depending on the position and advisory availability in the run. The following are findings relevant to usability evaluation in the efficiency and satisfaction aspects.

Responses showed that the controllers felt they had more spare attention in the Advisory runs than in the Baseline runs for both the peak-traffic time ($F^*_{1.01, 7.90}$ = 10.6, p = 0.012) and the majority of the time ($F^*_{1.01, 7.82}$ = 11.9, p = 0.009). The mean ratings for the peak-time periods were 4.3 in the Advisory runs and 3.0 in the Baseline runs (Fig. 10). The mean ratings for the majority of the time were 4.3 in the Advisory runs and 3.3 in the Baseline runs (not shown). Advisory × Scenario effect also was found to be statistically significant ($F_{3, 15}$ = 6.95, p = 0.004 in peak time; $F_{3, 15}$ = 3.48, p = 0.043 the majority of the time), and planned-comparison analyses

indicated that the contrasts were statistically significant between the two traffic levels ($F_{1, 15}$ = 16.3, p = 0.002 for peak time; $F_{1, 15}$ = 8.14, p = 0.012 for the majority of the time). Fig. 11 shows the means of the amount of the perceived spare attention during the peak time. The plot, again, shows a steeper drop in the Baseline runs (dotted line) than in the Advisory runs (solid line) when the traffic load increased. The plot for the majority of the time is not shown here but the trends are similar to Fig. 11.

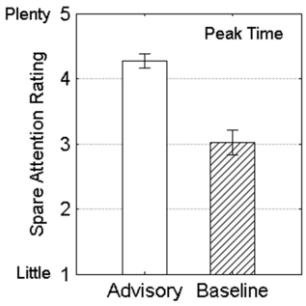

Figure 10. Means and standard errors of perceived spare attention amounts during peak time

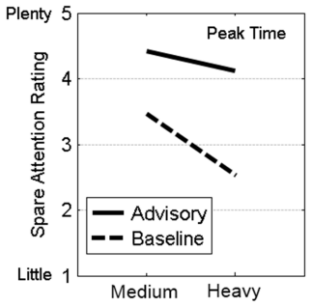

Figure 11. Means of perceived spare attention amounts during peak time

The controllers also perceived that the complexity was higher in the Baseline runs than in the Advisory runs for both determining the optimal departure sequence ($F^*_{1.03, 7.76}$ = 9.26, p = 0.016) and handling TMI-constrained flights ($F^*_{1.04, 6.46}$ = 7.85, p = 0.028). On the 10-point scale for complexity (1 being not complex at all, and 10 being very complex), the mean ratings for determining the optimal departure sequence were 2.4 and 4.2 in Advisory and Baseline runs, respectively, and those for handling TMI-constrained flights were 2.1 and 4.2 in the Advisory and Baseline runs, respectively.

After the Advisory runs only, the controllers were asked what percent of the time they agreed with the SARDA advisories and what percent of the time they understood the reasoning behind the advisory. The mean ratings were 78.9% and 82.2%, respectively. Fig. 12 shows a 3D histogram plot of their responses. It shows that the majority of the points fell in the bins along the diagonal line ([0%, 0%] to [100%, 100%]) or bins adjacent to them, suggesting a correlation between them.

D. Post-Study Questionnaire Responses

In the post-study questionnaire form administered at the end of each week, a set of 12 Advisory-related questions were

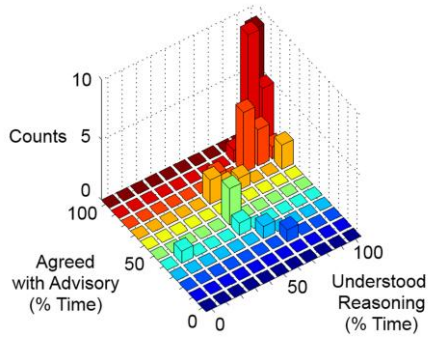

Figure 12. 3D histogram of controller responses

asked for the Local position, and then repeated for the Ground position. These were followed by seven user-interface-related questions, and then six general open-ended questions. The following are the findings pertinent to assessment of the tool's satisfaction aspect.

Table II lists the means and standard deviations of the responses to five Advisory-related questions for Local and Ground positions. Responses were made on a seven-point Likert scale, with labels noted in the table. The lower the rating was, the more favorable the response was for the SARDA advisories. The neutral point was 4.

TABLE II. POST-STUDY RESPONSES TO ADVISORY QUESTIONS

Question	Local Mean (Sdv)	Ground Mean (Sdv)
Overall, how easy/difficult was it to use the advisories? (1 = Easy, 7 = Difficult)	1.7 (0.8)	1.2 (0.4)
How easy/difficult do you think actual controllers in the field would likely find using the advisories? (1 = Easy, 7 = Difficult)	2.3 (1.4)	2.0 (1.6)
How much did the advisories help/interfere with your management of the traffic with a TMI restriction? (1 = Helped, 7 = Interfered)	2.2 (1.5)	3.0 (1.3)
How much did you trust the advisories to help you in making better decisions? (1 = Trusted, 7 = Did not trust)	3.2 (1.7)	2.8 (1.3)
Given the choice, would you prefer or not prefer to have the advisories? (1 = Prefer to have, 7 = Prefer not to have)	3.2 (1.9)	2.8 (1.9)

One of the six controllers tended to mark unfavorable responses for SARDA, and another tended to mark neutral. The remaining four controllers tended to mark favorable responses. As a result, as Table II indicates, all of the ten means fell on the favorable side for the SARDA advisories (< 4), though some were close to 4.

The responses to the three EFS-related questions were made on a seven-point Likert scale, as well, where 1 corresponded with the most favorable choice for the SARDA EFS (i.e., *easy*, *helped*) and 7 with the least favorable choice (i.e., *difficult*, *interfered*). The neutral point was 4. The results are listed in Table III. All of the three means fell on the favorable side for SARDA (< 4), but the second and third questions resulted in the response means being close to 4.

TABLE III. POST-STUDY RESPONSES TO EFS QUESTIONS

Question	Mean (Sdv)
How easy/difficult was it to understand the information on the EFS? (1 = Easy, 7 = Difficult)	1.8 (1.0)
How easy/difficult was it to manage the strips on the EFS? (1 = Easy, 7 = Difficult)	3.3 (1.9)
How much did the EFS help/interfere with your management of traffic? (1 = Helped, 7 = Interfered)	3.5 (1.8)

Lastly, the responses to the general open-ended questions showed that the controllers thought the strengths of the SARDA advisories were: i) planning and sequence-decision-

making assistance (in the Local position, especially for planning for TMI-constrained flights), ii) stress reduction, and iii) optimized departure queue. They thought weaknesses of the advisories were: i) handling of runway-crossing traffic (Local only), ii) little room for controllers' discretion, iii) potentially degrading performance of highly skilled controllers, and iv) controllers put out of the loop.

E. System-Performance Results

A brief summary of the system-performance data analysis results is provided in this section. Further details of the analyses are found in [2].

The analyses showed that SARDA successfully reduced the taxi delay and fuel consumption. In the medium-traffic scenarios (M1 and M2), taxi delay per aircraft (defined as the actual taxi time minus the unimpeded taxi time) was reduced by 45%, and extra fuel burn per aircraft (computed in a similar manner) was decreased by 23% in the Advisory runs in comparison to the Baseline runs. In the heavy-traffic scenarios (H3 and H4), taxi delay per aircraft was reduced by 60%, and extra fuel burn was decreased by 33% in the Advisory runs.

On the other hand, the TMI takeoff-roll conformance rates showed only weak trends of improvement in the Advisory runs, and the results remained inconclusive due to the small sample size of the nonconformance cases. Also, no runway throughput increases or concessions were observed in runs with or without the advisories.

IV. Discussion

As noted, the ISO describes three aspects of usability: effectiveness, efficiency, and satisfaction [14].

A. Effectiveness Aspect

The system-performance data demonstrated that the SARDA tool successfully reduced taxi delay and fuel consumption. The TMI-conformance and runway-throughput performances, whose improvements were aimed yet not explicitly demonstrated in this study, also at least did not show any deterioration. The results suggest that, with respect to these two system performances, the controllers may have been able to perform well enough even without the advisories, leaving only little room for improvements.

The difference between the Advisory and Baseline runs were detected more distinctly in the subjective data relevant to the other two usability aspects—efficiency and satisfaction. The following subsections discuss about each of these aspects.

B. Efficiency Aspect

The efficiency discussion examines the resources internally available to the controllers, and for this, the workload ratings and some of the post-run questionnaire responses are examined.

The NASA TLX workload rating results exhibited clear reductions of workload levels in terms of Temporal Demand (time pressure), Effort (how hard controllers had to work physically and mentally), Physical Demand (e.g., using EFS, communicating on the radio), and Mental Demand (e.g., thinking, deciding, calculating, remembering, looking). In all four ratings, the magnitude of the mean-score reductions from the Baseline runs to the Advisory runs was approximately 2 points, which may have been large enough to be sensed by the controllers. Temporal Demand and Mental Demand may have been reduced mainly by SARDA's scheduling function. The additional capabilities enabled by the availability of the target departure time in the Advisory runs—such as gate holding and automatic strip sorting—also may have contributed to the reduced time pressure and mental workload in the Advisory runs. The reduced need for manual strip-sorting may have helped lower the Physical Demand ratings, as well. The Effort ratings can be considered as a combination of Temporal, Physical, and Mental Demands, and, indeed, exhibited similar effects from the advisories in these three ratings. The results of planned-comparison analyses on Advisory × Scenario effects, plotted in Fig. 9, show that the controllers felt the task became more difficult as the traffic load increased when the advisory was not provided, whereas they felt the perceived difficulty of the tasks was almost unaffected by the traffic-load increase when the advisories were provided.

The real-time workload ratings showed only slight trends in workload reduction when the SARDA advisories were provided. The trend was weak in both the statistical sense (i.e., only marginally significant) and the magnitude of the workload reduction. The relative lack of strength may have been caused by the narrower range of the response scale (a 7-point scale in contrast to the 10-point scale used in the TLX ratings) and the tendency for the controllers to use only the lower range of the scale (1-3) when self-assessing their workload level in real time. The overall direction of the effect agreed with those in the TLX ratings, however.

According to the post-run questionnaire responses, the controllers felt they had more spare attention during both the peak time and the majority of the time when the advisories were provided. They also felt that the complexity they perceived in determining the optimal departure sequence and handling of the TMI-constrained flights was reduced when the advisories were available. SARDA's scheduling function may have helped reduce the perceived complexity of the tasks, which may have relieved some of the attention needed for the tasks. Increasing controllers' spare attentional capacity may be one of the major benefits of SARDA, as the controllers can use the extra cognitive resources to monitor more traffic, plan ahead better, resolve more complex problems, and so forth. The planned-comparison analysis results suggested that the effects of the SARDA advisories on the spare attention amount became more evident in heavier traffic. Likewise, in the TLX ratings, the advisories made the spare attention amount less susceptible to being affected by the traffic load increase.

To conclude the efficiency discussion, the SARDA advisories reduced the perceived complexity of the traffic management tasks, increased the controllers' spare attention capacity, and reduced the controller workload, especially in terms of time pressure, physical workload, and mental workload. With SARDA advisories provided, the controllers could retain a similar amount of cognitive resources (i.e., spare attention, spare workload capacity) even when the traffic volume increased.

Besides the Advisory effect, the ANOVA also detected strong Participant-related effects in the workload ratings and questionnaire responses. This means these subjective data contained large variations among the controllers (as expected). The repeated-measures ANOVA used in the above inferences can handle such variations to a certain extent (in fact, it exploits the large variations among the controllers). Given an appropriate analytical tool, having more controller participants in the study can make the results more robust and, thus, is a critical improvement from the previous HITL simulation study.

C. Satisfaction Aspect

Next, the satisfaction aspect is examined based on the post-study questionnaire and some of the post-run questionnaire responses.

In Table II, the relatively low scores of the responses to the first and second questions suggest that the controllers felt the SARDA advisories were easy enough to use and believed that controllers in the field also could use them with ease. They thought the advisories helped the Local controller in managing the TMI-constrained flights (the third question, Local). The responses to the open-ended questions showed that the controllers acknowledged SARDA's strength in its ability to assist controllers in planning and decision making, reduce workload, and generate optimal departure sequences.

The fourth and fifth questions in Table II resulted in slightly high scores in both positions (> 2.5), though still on the favorable side for the SARDA concept (< 4). The questions were related to controllers' trust and acceptance. These results may suggest the controllers have slight reservations about the SARDA tool. To consider what would make the advisories more trustworthy and acceptable, let us go back to one of the post-run questions. Fig. 12 shows a correlation between the percent time that the controllers understood the reasoning behind the advisories and the percent time that they agreed with the advisories. One possible conjecture is that more understanding of the reasoning behind the advisories increases the likelihood of a controller's agreement. If that is true, then explaining to them some high-level reasoning of the SARDA optimization processes, such as how multiple competing priorities are handled, may help them to understand the advisories more and raise the chance that they will agree with the advisories. (Note that there are other possible conjectures, such as that agreement with an advisory causes understanding it, and that both agreement and understanding are driven by something else.)

Providing more information about the reasoning of the SARDA optimization processes also would improve the controllers' situation awareness about the SARDA operations, which could, in turn, address some of the out-of-the-loop controller problems raised in the post-study open-ended question responses. The problem of a human operator being out of the loop is actually a by-product of the reduced workload benefit, and not necessarily a negative consequence. However, it becomes a problem when the controller suddenly needs to take over the operation. Lack of situation awareness is often noted as a cause of an out-of-the-loop controller [20]. Thus, learning the high-level reasoning behind how the advisories were generated may help controllers to stay in the loop or smoothly take over the operation when needed.

Another known negative consequence of out-of-the-loop operators is skill degradation [20]. It is usually a long-term consequence, and cannot be easily demonstrated in a short-term simulator study like the current one. However, it is a valid concern that, after prolonged use of SARDA, controllers may lose some manual ATC skills, or new controllers may never develop these skills. Additional controller training may be required to help them to retain the necessary skills. The automation also should be made reasonably reliable, so that the controllers seldom need to take over the operation, and, if such an instance occurred, should assist smooth transition to a manual mode (e.g., via *graceful degradation* of functionality). This is a common concern with many advanced automation tools, not only with SARDA.

In the open-ended question responses, the controllers noted little room for controllers' discretion (ii) and potentially degrading performance of highly skilled controllers (iii) as SARDA's weaknesses. These were primarily caused by the current simulation's artificial constraint that the controller participants had to follow the advisories. In the actual SARDA concept of operation, controllers are allowed to deviate from the advisories whenever they wish—in fact, it is the controller's responsibility to assess each advisory and reject those that seem unsafe or inappropriate. This flexibility would address the above two weaknesses. The controllers' deviations would be absorbed in the frequent SRP refresh cycles. Effects of this process on system performance and usability need to be researched further, however.

The results of the EFS-related question responses shown in Table III suggest that the controllers were able to understand the information on the EFS relatively easily (the first question). However, the scores for the responses to the second and third questions being closer to 4 imply that they may have had some reservations. An unresponsive user-interface could seriously hinder traffic-management performance if it happens in the middle of a high-workload time. This is not a problem of the SARDA concept itself, but since the EFS responsiveness could affect results of a HITL simulation evaluation, this problem needs to be resolved before the next evaluation.

To conclude the satisfaction discussion, it appeared that the controllers thought the SARDA tool was easy enough to use and acknowledged its ability to assist them in planning and decision making, optimizing the runway sequence, improving the TMI takeoff-roll time conformance, and reducing their workload. However, the data also suggested that the controllers had slight reservations about giving full endorsement just yet. The SARDA concept still needs to win their trust and acceptance by, for instance, explaining the high-level reasoning behind advisories to them. It is also recommended to improve the responsiveness of the EFS touch screen.

V. Conclusion

The previous analyses of the system-performance data demonstrated that SARDA helped reducing taxi delays and fuel burn. The subjective-data analyses conducted in the current

study shed additional light on the effects of the SARDA advisories on the ATCT controllers.

The efficiency-aspect analyses revealed that the SARDA advisories helped lower the controllers' workload, especially with respect to time pressure, physical workload, and mental workload. When the SARDA advisories were provided, the workload also became less sensitive to be affected by the traffic load increase. It was considered that the SARDA's scheduling function helped reducing task complexity and relieved their attention, which, in turn, reduced their workload. Additional capabilities afforded by the presence of the SARDA's target departure sequence, such as gate holding and automatic strip sorting, may have also contributed to reducing the workload.

The satisfaction-aspect analysis showed that the controllers were generally favorable toward the SARDA concept, with slight reservations: They thought the tool was easy enough to use. They also valued SARDA's ability to help them in attaining the optimal runway sequence, improving TMI takeoff-roll time conformance, and reducing their workload. More work is needed for the concept to gain more trust and acceptance from controllers. Explaining the high-level reasoning behind the advisories to the controllers may help them to feel more comfortable in accepting the advisories, maintain situation awareness, and remain in the loop.

VI. Future Work

The current HITL simulation study evaluated relatively simple cases of nominal routine operations with retired ATCT controllers as a proof of the SARDA concept. Consequently, the usability assessment results reported in this paper are limited to these contexts. Future work will need to look at the tool's usability in other contexts, such as off-nominal operations and other airport operations. In addition, as pointed out in Section IV-C, evaluating the effects of controllers' deviations from the SARDA advisories on the system performance and the usability is another proposed future work.

Acknowledgments

Special thanks go to the SARDA simulator development staff members for their hard work and patience: Easter Wang, Dan Pietrasik, Cynthia Freedman, Kenny Ray, Carla Ingram, Boris Rabin, Betty Silva, and David Chin. The authors thank, for valuable input received, our fellow SARDA research team members: Bob Windhorst, Justin Montoya, Len Tobias, and Jon Holbrook. The authors are also grateful for expert advice received from our controller subject matter experts: Mark, Mitch, Barbara, Fred, Rebecca, and Kari. Last but not least, a big thank you goes to all of our controller and pilot participants.

References

[1] Y. C. Jung, T. Hoang, J. Montoya, G. Gupta, W. Malik, and L. Tobias, "A concept and implementation of optimized operations of airport surface traffic," 10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference, Fort Worth, TX, September 2010.

[2] G. Gupta, W. Malik, L. Tobias, Y. Jung, T. Hoang, and M. Hayashi, "Performance evaluation of individual aircraft based advisory concept for surface management," the Tenth USA/Europe Air Traffic Management Research and Development Seminars, Chicago, IL, June 2013 (submitted).

[3] S. Atkins, Y. Jung, C. Brinton, L. Stell, T. Carniol, and S. Rogowski, "Surface Management System field trial results," AIAA 4th Aviation Technology, Integration, and Operations (ATIO) Forum, Chicago, IL, September 2004.

[4] W. Malik, G. Gupta, and Y. C. Jung, "Managing departure aircraft release for efficient airport surface operations," AIAA Guidance, Navigation, and Control (GNC) Conference and Modeling and Simulation Technologies (MST) Conference, Toronto, Canada, August 2010.

[5] G. Gupta, W. Malik, and Y. C. Jung, "Incorporating active runway crossings in airport departure scheduling," AIAA Guidance, Navigation, and Control (GNC) Conference and Modeling and Simulation Technologies (MST) Conference, Toronto, Canada, August 2010.

[6] J. Montoya, Z. Wood, and S. Rathinam, "Runway Scheduling Using Generalized Dynamic Programming," AIAA Guidance, Navigation, and Control (GNC) Conference, Portland, OR, August 2011.

[7] Y. Jung, T. Hoang, J. Montoya, G. Gupta, W. Malik, L. Tobias, and H. Wang, "Performance evaluation of a surface traffic management tool for Dallas/Fort Worth International Airport," Ninth USA/Europe Air Traffic Management Research and Development Seminar, Berlin, Germany, June 2011.

[8] T. Hoang, Y. C. Jung, and J. B. Holbrook, "Tower controllers' assessment of the Spot and Runway Departure Advisor (SARDA) concept," Ninth USA/Europe Air Traffic Management Research and Development Seminar, Berlin, Germany, June 2011.

[9] G. Gupta, W. Malik, and Y. C. Jung, "An integrated collaborative decision making and tactical advisory concept for airport surface operations management," 12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference, Indianapolis, IN, September 2012.

[10] T. R. Truitt, "Electronic flight data in airport traffic control towers: literature review," Technical Report, DOT/FAA/CT-05/13, Federal Aviation Administration, Atlantic City, NJ, 2006.

[11] T. J. J. Bos, M. Shuver-van Blanken, and H. Huisman, "Towards a paperless air traffic control tower," NLR-TP-2011-192, National Aerospace Laboratory NLR, Amsterdam, the Netherlands, May 2011.

[12] F. T. Durso, and C. A. Manning, "Spinning paper into glass: transforming flight progress strips," Human Factors and Aerospace Safety, vol. 2(1), pp. 1-31, Ashgate Publishing, Hampshire, UK, 2002.

[13] J. Jeng, "Usability assessment of academic digital libraries: effectiveness, efficiency, satisfaction, and learnability," International Journal of Libraries and Information Services, vol. 55, pp. 96–121, 2005.

[14] International Organization for Standardization, "Ergonomic requirements for office work with visual display terminals (VDTs)—part 11: guidance on usability," ISO 9241-11, Geneva, Switzerland, 1998.

[15] N. Dorighi and B. Sullivan, "FutureFlight Central: a revolutionary air traffic control tower simulation facility," AIAA Modeling and Simulation Technologies Conference and Exhibit, Austin, TX, August 2003.

[16] E. S. Stein, "Air Traffic Controller Workload: An Examination of Workload Probe," DOT/AFF/CT-TN84/24, Federal Aviation Administration, Atlantic City Airport, NJ, April 1985.

[17] C. A. Manning, S. H. Mills, C. Fox, E. Pfleiderer, and H. J. Mogilka, "Investigating the validity of Performance and Objective Workload Evaluation Research (POWER)," DOT/FAA/AM-01/10, Federal Aviation Administration, Washington, D.C., July 2001.

[18] S. G. Hart and L. E. Staveland, (1988). "Development of a NASA-TLX (task load index): results of empirical and theoretical research," in Human Mental Workload, P. S. Hancock & N. Meshkati, Eds. Amsterdam: Elsevier Science Publishers B. V., 1988, pp. 139-183.

[19] H. R. Lindman, Analysis of Variance in Complex Experimental Designs. San Francisco, CA: W. H. Freeman and Company, 1974.

[20] M. R. Endsley and E. O. Kiris, "The out-of-the-loop performance problem and level of control in automation," Human Factors, vol. 37(2), pp. 381-394, 1995.