2011

Situation Awareness, Workload, and Performance in Midterm Nextgen: Effect of Variations in Aircraft Equipage Levels Between Scenarios

L. Paige Bacon
Thomas Z. Strybel
Kim-Phuong L. Vu
Joshua M. Kraut
Jimmy H. Nguyen

See next page for additional authors

Follow this and additional works at: https://corescholar.libraries.wright.edu/isap_2011

Part of the Other Psychiatry and Psychology Commons

Repository Citation

This Article is brought to you for free and open access by the International Symposium on Aviation Psychology at CORE Scholar. It has been accepted for inclusion in International Symposium on Aviation Psychology - 2011 by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
We investigated the impact of mixed equipage traffic on air traffic controller (ATC) performance, workload and situation awareness in an airspace simulation of an en route sector. Nine retired ATCs and seven student ATCs managed traffic and responded to probe questions designed to measure workload and situation awareness. ATC performance was measured with the AT-SAT OTS Performance Rating Form. Although workload decreased with increasing mixtures of equipped aircraft, the number of LOS appeared to be related more to the percentage of alerted conflicts in the scenario. The probe measure of workload and situation awareness was related to AT-SAT ratings.

The Next Generation of Air Transportation System (NextGen), is a program being developed under the guidance of several government agencies in order to expand the capacity of the National Airspace System (NAS), reduce the negative environmental impacts on the environment, improve efficiency and safety, and reduce delays on the ground (FAA Implementation Plan, 2010). To accomplish these goals, new tools, technologies and procedures for NextGen operations are being developed (e.g., Prevot, Homola, & Mercer, 2008). Because these tools, technologies, and concepts will change how air traffic controllers (ATCs), perform their tasks, it is important to evaluate NextGen developments in terms of their impact on human operator performance, workload, and situation awareness (SA), and investigations of these impacts have appeared in the literature (e.g., Willems and Schulz, 2010). Technologies such as Data Comm are expected to reduce ATC workload and enable controllers to manage higher traffic densities by reducing the need for ATC-pilot voice communications. Willems and Schulz (2010) determined that the number of voice communications, an indicator of ATC workload, was reduced only when the percentage of equipped AC exceeded 50%; at lower equipage levels, no change in workload was observed. In addition, tools such as automated conflict detection and resolution (CD&R), and trial planners should further reduce controller workload and improve their performance. However, because these tools may be useful only for separating pairs of appropriately equipped aircraft (AC), ATCs will need to rely on their current-day air traffic management skills for detecting and resolving conflicts involving pairs unequipped ACs in addition to using the advanced tools. Therefore, in the present study we determined the impact of air traffic containing different percentages of ACs equipped with both Data Comm and CD&R capabilities on ATC performance, workload, and situation awareness.

Evaluations of NextGen tools such as Data Comm and CD&R on operator performance require reliable and valid measures of performance. The need for reliable, valid, and diagnostic measures of human operator performance is not new, and remains a problem for the assessment airspace operators in both current-day and NextGen environments because the relationships between objective measures of airspace system performance and operator performance have yet to be determined, yet objective measures of sector outcomes are easily extracted from digital systems. Rantennan (2004), in a recent review of ATC performance metrics, summarized the problem as follows: “The problem is therefore not in the availability of data, but in derivation of valid, reliable, and meaningful measures from the abundance of data (p. 2).” This problem becomes more pronounced for measuring cognitive constructs such as workload and SA because they are not directly observable, but are considered important determinants of performance. Workload and SA have been shown to be related to operational errors (e.g., Jones & Endsley, 1996). Workload is commonly measured with standardized self-rating methods such as ATWIT (Stein, 1985) and NASA TLX (Hart & Staveland, 1988) both of which have been extensively tested and validated. Consensus has not been achieved, however, on the best approach to SA measurement (e.g., Stanton, et al., 2006).

In the present investigation, we attempted to identify the individual performance metrics that contribute the most to overall ATC performance. We first measured overall ATC performance using two standard ATC
assessment tools, the Behavioral and Event Checklist (BEC) and a modified AT-SAT High Fidelity Simulation Over the Shoulder (OTS) Rating Form (Ramos, Heil, & Manning, 2001). The modifications to the AT-SAT were needed to include an evaluation of ATC’s use of Data Comm. Both instruments have shown to be reliable and valid predictors of ATC performance (e.g., Manning and Stein, 2005). The overall measures then were used to determine a set of performance, workload, and SA measures that are significantly associated with the overall measure. We assessed operator SA, or the operators’ understanding of the task environment, by using a variant of the Situation Present Assessment Method (SPAM; Durso & Dattel, 2004) because a real-time probe method can measure SA when it is distributed in the task environment. (Chiappe, Strybel & Vu, submitted).

**Method**

**Participants**

Sixteen participants served as ATCs in the present study. Nine were radar-certified, retired air traffic controllers (6 TRACON and 3 ARTCC) and seven were students enrolled in the Aviation Sciences Program at Mount San Antonio College, an FAA CTI institution.

**Apparatus**

The simulation was run using the Multi Aircraft Control System (MACS), Aeronautical Datalink and Radar Simulator (ADRS), and VoiceIP software, developed by the Airspace Operations Laboratory at NASA Ames Research Center (e.g., Prevot et al., 2006). ATCs used MACS configured as a Digital System Replacement (DSR) display with integrated Data Comm, conflict alerting, and conflict probing for all aircraft (AC) designated as “equipped.” Online SA probes were presented on a touch-screen workstation located adjacent to the DSR display.

**Scenarios**

Six experimental trials were run, each lasting approximately 50 minutes, with two each containing 25%, 50%, or 75% equipped AC. The experimental scenarios were originally constructed with eight planned traffic conflicts, and the number of conflict alerts was manipulated in the scenario by setting AC equipage. Table 1 shows the intended number of conflicts and conflict alerts for each percentage of equipped AC. However, it was determined after data collection was finished that not all scenarios contained 8 conflicts (see Table 1 for the actual number of conflicts). Therefore, the scenario containing 2 conflicts (one of the 50% Data Comm scenarios) was not analyzed. Moreover, the variables Percent Equipped AC and Number of Alerted Conflicts were combined into one variable, that of Percentage of Alerted Conflicts, also shown in Table 1.

<table>
<thead>
<tr>
<th>Percent Equipped AC</th>
<th>Planned Number of Alerted Conflicts (out of 8 conflicts)</th>
<th>Number Conflicts</th>
<th>Number Alerts</th>
<th>Percent Alerted Conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>75 6</td>
<td>8</td>
<td>6</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>75 4</td>
<td>7</td>
<td>4</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>50 5</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>50 3</td>
<td>7</td>
<td>2</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>25 3</td>
<td>2</td>
<td>0</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>25 1</td>
<td>8</td>
<td>1</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

**Procedure**

Participants were given one day of training and practice prior to the experimental trials. During the experimental scenarios, 14 SA probe questions were presented, one every three minutes, beginning four minutes into the scenario. The sequence of events for each online probe question was as follows. The ATC received a “Ready for
Question” prompt on the touch screen accompanied by an audio alert. The participant was instructed to respond by touching a button on the screen only when he/she had sufficient capacity to answer a probe question. Once the participant responded affirmatively, a probe question and response alternatives were immediately displayed on the panel and the participant responded via the touch screen. If the participant did not respond to the “Ready” prompt within two minutes, the query was withdrawn.

The probe questions were developed beforehand, to ensure that equal numbers of questions were asked from each of four task- information categories: conflicts, sector status, command-communication and subjective assessment. Conflict questions queried information regarding current and future conflicts between an aircraft pair. Sector status questions were regarding the current state of the sector, such as number of equipped aircraft. Questions categorized as command-communications asked for information on previous or future commands and communications (both voice and Data Comm). Subjective assessment questions asked participants to rate their concern (1= low, 5=high) regarding a future conflict event. At the end of each scenario, participants completed the Situation Awareness Rating Technique (SART) and the NASA TLX workload measure.

**Expert Evaluation of Performance**

The performance of each participant was evaluated by two of three expert raters who reviewed screen recordings of a simulation run. Three experts (two retired en route ATCs, one retired TRACON ATC) were trained on the use of the BEC and AT-SAT OTS Rating. These raters independently watched a set of identical test recordings and recorded events on the BEC. At the end of the recorded scenario, the raters completed the AT-SAT Rating Form. The AT-SAT Rating Form consists of 8 rating dimensions. The evaluator assigned a number between 1 and 8 based on information from the BEC. The evaluators trained until an acceptable level of agreement between the raters was reached. Subsequently, each rater independently assessed 32 of the possible 96 recordings. Raters also completed the NASA TLX at the end of each scenario.

![Figure 1](image1.png) **Figure 1.** Mean AT-SAT OTS Rating as a function of Percent of Alerted Conflicts and Rating Dimension

![Figure 2](image2.png) **Figure 2.** Mean Number of LOS as a function of Percent of Alerted Conflicts and Conflict Pair Equipage.

**Results**

All 8 Rating Dimensions on the ATSAT-OTS form were significantly correlated with each other ($r_s = .34$ to .90; $p_s < .003$). A mixed-design ANOVA was run on ratings for each dimension, with the within-subject variable of percentage of alerted conflicts (4%, 13%, 28%, 57% and 75%) and between subject variable expertise (student vs. expert). A main effect of percentage of alerted conflicts was obtained on three ATSAT OTS dimensions ($p_s < .06$): Maintaining Separation, Maintaining Attention and Situation Awareness and Overall Performance, as shown in Figure 1. Maintain Attention and SA and Overall Performance Ratings were affected similarly by percentage of alerted conflicts. The highest average rating was given when 57% of the conflicts were alerted; lower ratings were made for higher and lower percentages. The effect of percentage of alerted conflicts on Maintaining Separation did not follow this pattern, however. AT-SAT ratings were lowest for scenarios having 13%, 28% and 68% alerted conflicts. The NASA TLX ratings were not significantly correlated with the OTS ratings. However, the TLX composite score was significantly affected by percentage of alerted conflicts.

A battery of performance metrics were also computed for each ATC and scenario. These were used to measure safety, efficiency, and workload. Measures of safety included number LOS, and average lateral and vertical distance between ACs. The number of LOS was counted separately for conflicts involving two equipped
AC, one equipped AC, and no equipped AC. The latter two conflict types were not alerted. For each variable, a mixed ANOVA was run with the factors percentage of alerted conflicts and expertise. A significant effect of percentage of alerts was obtained only for the number of LOS involving at least one unequipped AC ($p < .002$). In Figure 2, the average number of LOS for each equipage is shown. For scenarios having low percentages of alerted conflicts, most LOS occurred between two unequipped AC, which is not surprising because these scenarios contained relatively few equipped aircraft. At the highest percentages, more LOS occurred between equipped - unequipped AC pairs. A main effect of expertise was obtained for conflicts between equipped and unequipped ACs only ($p < .001$). Students, on average had nearly double the number of LOS compared with experts (student: $M = .64$; ATC: $M = .25$) when one AC was equipped and one AC was unequipped. Although experts had slightly more LOS for the remaining AC pairs, the differences were not significant.

The average lateral and vertical separations were significantly affected by percentage of alerts, and an interaction was obtained between expertise and percentage of alerted conflicts, was found for average vertical separation between AC ($p < .02$). As shown in Figure 3, expert ATCs typically maintained greater vertical separation between AC when the percentage of alerted conflicts was either very low or very high. In the mid range, the differences were minimal. Students maintained greater vertical separation in the scenario containing the 28% alerted conflicts. Performance measures of efficiency included mean and standard deviation of handoff latency, distance traveled, and time through the sector. All but mean handoff time were affected by the percentage of alerts ($ps < .01$). No effects of expertise were observed on these metrics.

Two performance measures of workload were computed, total number of voice communications and percentage of time on voice. A significant effect of percentage of alerts on total number of voice communications was obtained, most likely because the percentage of alerted conflicts increased with percentage of equipped aircraft. The proportion of voice communications decreased from .85 when the percentage of alerts was 13% (and 25% of AC were equipped) to .38 when the percentage of alerts was at least 57% (and 75% of AC were equipped). Thus, the total number of voice communications was related to the percentage of equipped AC. Unlike the OTS-TLX composite scores, participant self-ratings of TLX workload were not significantly affected by percentage of alerts.

Accuracy and response latency for probe questions overall, and by information category, were analyzed with the same mixed design ANOVA. All effects on probe accuracy were nonsignificant. Overall, response latency was affected expertise ($p < .01$). When probe latencies were evaluated for each probe information category, differences in response latencies between expert and student ATCs were significant for conflict and command-communication probes ($ps < .01$) and marginally significant for subjective and sector-status probes ($ps < .07$.) As shown in Figure 4, experts consistently responded faster than students, with the largest difference observed for conflict questions. Response latencies to the ready prompt, assumed to measure workload, was significantly affected by expertise ($p < .03$) and marginally affected by the percentage of alerts ($p < .10$).

### Predicting AT-SAT OTS Ratings

Hierarchical regressions were run to determine the combination of performance metrics that best describe ATC overall performance. A three level model was used. The experimental variables of percentage of alerted conflicts and expertise were entered first, followed by objective performance metrics, and then SA and Workload measures. The number of LOS variable was combined into two categories, alerted LOS (both Equipped AC) and non-alerted LOS (at least one AC unequipped). For situation awareness, overall probe latency and accuracy were
entered, and for workload the number of voice communications, participant rating of TLX, and latency to the probe ready prompt were entered. Regressions were performed on the three rating dimensions, in which percentage of alerted conflicts significantly affected expert ratings, shown in Figure 1.

AT-SAT OTS ratings of Overall Performance was predicted by experimental variables (percentage of alerted conflicts and expertise), objective performance metrics (number non-alerted LOS), situation awareness (probe latency), and workload (ready latency), as shown in Table 2. This model was significant \( F(10,65) = 10.92; p<.0001 \) and accounted 46% of the variability in overall performance ratings. The model for the Rating Dimension “Maintaining Attention and Situation Awareness” was identical to Overall Performance rating with the exception that the Ready Latency variable was not significant.

**Table 2**

**Significant predictors of AT-SAT OTS Overall Performance Rating (ps<.15).**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Est.</th>
<th>Partial r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Alerted Conflicts</td>
<td>-1.87</td>
<td>.07</td>
</tr>
<tr>
<td>Expertise</td>
<td>0.82</td>
<td>.03</td>
</tr>
<tr>
<td>Number non-alerted LOS</td>
<td>-0.43</td>
<td>.21</td>
</tr>
<tr>
<td>Handoff Latency</td>
<td>-4.17</td>
<td>.07</td>
</tr>
<tr>
<td>Probe Latency</td>
<td>-0.04</td>
<td>.03</td>
</tr>
<tr>
<td>Ready Latency</td>
<td>-0.05</td>
<td>.01</td>
</tr>
</tbody>
</table>

Table 3 shows the regression model for the Rating Dimension “Maintaining Separation.” Several additional factors were significant to the ones mentioned above. Neither expertise nor percentage of alerted conflicts entered into this model. Instead, the most highly predictive variables were the number of non-alerted LOS which accounted for 31% of the variance in ratings. The number of alerted LOS accounted for an additional 6%. This was followed by variables reflecting safety (e.g., average vertical distance) and efficiency (average handoff latency). Lastly, both situation awareness and workload entered in the model, with probe latency accounting for 4% of the variance in ratings, and participant TLX accounting for 2%.

**Table 3**

**Significant predictors of AT-SAT OTS Rating “Maintaining Separation” (ps<.15)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter</th>
<th>Partial r²</th>
<th>Variable</th>
<th>Parameter</th>
<th>Partial r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number non-alerted LOS</td>
<td>-0.83</td>
<td>.31</td>
<td>Average Lateral Distance</td>
<td>0.19</td>
<td>.02</td>
</tr>
<tr>
<td>Number alerted LOS</td>
<td>-1.09</td>
<td>.06</td>
<td>Std Dev Distance Traveled</td>
<td>-5.94</td>
<td>.02</td>
</tr>
<tr>
<td>Handoff Latency</td>
<td>-6.31</td>
<td>.06</td>
<td>Std Dev Time Thru Sector</td>
<td>0.71</td>
<td>.03</td>
</tr>
<tr>
<td>Probe Latency</td>
<td>-0.07</td>
<td>.04</td>
<td>Std Dev Handoff Accept Latency</td>
<td>-0.05</td>
<td>.01</td>
</tr>
<tr>
<td>Average Vertical Distance</td>
<td>-0.001</td>
<td>.02</td>
<td>TLX – Participant Rating</td>
<td>-0.02</td>
<td>.02</td>
</tr>
</tbody>
</table>

**Discussion**

This preliminary investigation indicated that ATC workload decreased with increased percentages of equipped AC, similar to the finding of Willem et al. (2008). However, in our scenarios, we included CD&R tools for the management of equipped AC pairs, and the percentage of conflict alerts affected the number and types of LOS made: more conflicts between equipped and unequipped aircraft were found as the percentage of equipped AC increased. Moreover, experienced ATCs were more capable of detecting and resolving these non-alerted conflicts in a mixed equipage environment compared with student ATCs, suggesting that students may have relied more heavily on automated CD&R tools.

AT-SAT OTS Rating dimensions of Maintaining Attention and SA and Overall Performance were significantly affected by the percentage of alerted conflicts, with the lowest ratings made for 75% alerted conflicts. Considering that LOS has the most impact on the rating of overall performance (i.e., overall performance rating
cannot exceed 3.0 when an LOS occurs), these ratings are consistent with our findings regarding LOS. However, our SA measure of probe latency significantly reduced the variance in AT-SAT ratings for overall performance, maintaining attention and situation awareness, and maintaining separation. We believe this provides strong support for the validity of an online probe technique as a measure of SA and workload. Probe latencies also discriminated between expert and student ATCs. However, the probe latencies to questions categorized by task information did not predict AT-SAT ratings.

Acknowledgements

This simulation was partially supported by NASA cooperative agreement NNA06CN30A: NRA: Metrics for Situation Awareness, Workload, and Performance in Separation Assurance Systems (Walter Johnson, Technical Monitor). Preparation of this paper was supported by NASA cooperative agreement NNX09AU66A, Group 5 University Research Center: Center for Human Factors in Advanced Aeronautics Technologies (Brenda Collins, Technical Monitor).

References


