Investigating Sources of Mental Workload Using a High-Fidelity ATC Simulator

Penelope M. Sanderson
Martijn Mooij
Andrew Neal

Follow this and additional works at: https://corescholar.libraries.wright.edu/isap_2007

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 - 2007 by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
INVESTIGATING SOURCES OF MENTAL WORKLOAD USING
A HIGH-FIDELITY ATC SIMULATOR

Penelope M. Sanderson, Martijn Mooij, Andrew Neal
ARC Key Centre for Human Factors and School of ITEE
The University of Queensland, St Lucia, Queensland, Australia

In this study we present results from a high-fidelity simulator study of Air Traffic Controller (ATCo) workload, in which routine and non-routine scenario events are examined. Specifically, we test the ability of a task load metric (Airservices Australia’s Workload Assessment Tool: WAT) to predict controller workload in these situations. In the study, sector was a between-subjects variable and task load a within-subjects variable. Twenty-one professional ATCos worked three 30-minute segments in Airservices Australia’s TAAATS simulator, and rated their workload on a scale from 1 to 10 every two minutes. After each scenario, ATCos reviewed the video and explained their ratings. For the moderately high to high workload scenarios we used, non-routine events did not have the predicted effect on rated workload. The WAT measure predicted 38% of the variance in workload ratings but the best fit of factors in the WAT task load algorithm to rated workload accounted for only 40% of the variance in rated workload. Video-cued recall data indicated a strong influence on workload of anticipated traffic rather than actual traffic on frequency, and a direction of activity to proactive rather than reactive control.

Introduction

There is a long tradition of research into the sources of Air Traffic Controller (ATCo) workload (Mogford, Gutman, Morrow, & Kopardekar, 1995; Hillburn, 2004). Being able to predict ATCo workload is important for various reasons. First, it helps planners determine sector sizes and traffic load that can be handled by an individual controller. Second, it helps planners determine when and how to combine sectors when traffic is low or decombine sectors when traffic is high. Third, reliable and valid ways to predict ATCo workload could help in the design and deployment of advanced air traffic management regimes, in which new roles and new forms of responsibility will emerge.

Many researchers have sought to identify properties of the air traffic situation that—either singly or in combination—will reliably predict the subjective mental workload experienced by the ATCo. Research in the above tradition has shown that it is difficult to predict more than 50% of the variance in workload ratings by using linear models of current air traffic situations (Kopardekar & Magyarits, 2003; Laudeman, Shelden, Branstron, & Brasil, 1998). In addition, it has proven very difficult to predict workload accurately ahead of time in order to put dynamic workload management measures into place, such as combining or decombining sectors or introducing flow restrictions (Majumdar & Ochieng, 2002; Masalonis, Callaham, & Wanke, 2003). Moreover, the applicability of such linear models for different sectors and for different kinds of sectors is largely unknown.

Loft, Sanderson, Neal and Mooij (in press) reviewed literature investigating the relationship between task demands and mental workload, and between ATCo capacity and mental workload. They pointed out that multicollinearity makes it difficult to use linear models to determine which aspects of a traffic situation drive subjective workload. Moreover, the relationship between subjective workload and performance—especially in connection with safety—is unknown because ATCos will adapt their air traffic management strategy. Indeed, Loft et al. (in press) found that much of the research on the relation between ATCo capacity and workload focuses on ATCo use of strategies that vary subjective workload while preserving safety, orderliness, and expeditiousness. ATCo capacity is therefore dynamic, rather than static.

Drawing from the work of Sperandio (1978) and others, Loft et al. (in press) proposed that a focus on how ATCos proactively control their own workload through the selection of strategies might be a more fruitful way of understanding ATCo workload and its relation to airspace safety.

In this paper we discuss a workload assessment tool (WAT) proposed by Airservices Australia personnel. Then we describe an empirical study conducted in a high-fidelity ATC simulator with professional ATCos in which we collected in-the-loop workload ratings from ATCos and compared how well they are predicted by various traffic load and density metrics, including the WAT tool. Finally we describe how an analysis of video-cued recall interviews has started to reveal sources of workload not easily captured in traffic load and density metrics.
The Workload Assessment Tool (WAT) is a linear model developed by Airservices Australia personnel to estimate the amount of work an ATCo must do under different traffic situations. Given sectors, routes and flight plans, the WAT calculates flight trajectories and estimates the amount of work the controller of a sector would have to do at each point. Therefore the WAT tool indicates “work (activity) to be done” rather than subjective workload. A question we wished to answer was how well the WAT measure estimates ATCo subjective mental workload.

The WAT’s profile of work to be done is a weighted linear function of six aspects of sector air traffic, each divided by the average distance between aircraft in the sector and weighted according to subject matter expert judgments of the probable contribution of each factor to work to be done.

$$\text{WAT} = 0.2 \times \frac{\text{AC in sector}}{\text{average distance}} + 0.1 \times \frac{\text{AC entering sector}}{\text{average distance}} + 0.15 \times \frac{\text{AC at top of climb}}{\text{average distance}} + 0.25 \times \frac{\text{AC at top of descent}}{\text{average distance}} + 0.2 \times \frac{\text{AC exiting sector}}{\text{average distance}} + 0.35 \times \frac{\text{AC in conflict (15 nm, A020)}}{\text{average distance}}$$

Subject matter experts have verified that WAT ratings reflect probable ATCo workload in sample scenarios, but in the present study we test this.

**Goals of present study**

Although estimating ATCo workload through linear combinations of task load factors has formal limitations, described above, there have also been empirical limitations in much work of this kind. For example, many studies have used ratings made by subject matter experts after the event rather than by the ATCos at the time of control (eg. Kopardekar & Magyarits*, 2003, dynamic density metric and the Manning et al., 2001, POWER metrics). Moreover, the dynamic density ratings were of traffic complexity rather than workload. Examining the relation between task load and traffic configuration with ATCo in-the-loop workload ratings will be informative. Moreover, it was important to test the effectiveness of the WAT approach—based on flight plans and traffic configurations—for predicting subjective workload in Australian airspace.

The first goal of the present study was to establish which task load factors would predict ATCo in-the-loop subjective workload in the TAAATS (The Australian Advanced Air Traffic System) environment. The predictive power of such a metric could then be tested on independent data. In the present report we focus on the WAT metric and we discuss prospects for using DDM measures.

The second goal was to examine the effect of non-routine events on ATCo workload. Most prior studies have predicted workload on the basis of relatively normal traffic patterns of different levels of intensity, with no systematic study of abnormal events. We wished to see whether non-routine events would increase ATCo workload disproportionally with respect to task load factors measured at the same time. If so, it would be important to pinpoint the cause of mismatches in an attempt to represent such factors in models predicting ATCo workload.

The third goal was to acquire a data set of rich ATCo behaviour in audiovisual recordings for in depth examinations of the relationship between traffic situations, ATCo strategies, and rated ATCo workload, with a view to understanding ATCos’ strategic management of workload.

To achieve the above goals, we sought a combination of high fidelity context, high expertise and good validity for workload ratings. We used Airservices’ TAAATS simulator. Professional ATCos worked only on sectors for which they were currently rated. For workload ratings we used the ATWIT procedure, collecting ratings from the ATCos while in the control loop. Finally, we conducted video-cued recall interviews after each scenario to provide qualitative insights into ATCo strategy and workload.

**Method**

**Test environment**

The test environment was the TAAATS high fidelity air traffic simulator at Airservices Australia in Brisbane. TAAATS is a fully electronic air traffic management system used for managing Australian airspace. The simulator itself is a full air traffic control environment that is capable of supporting real air traffic operations in an emergency, and so has an exceptionally high level of fidelity. TAAATS workstations contain a large Air Situation Display (ASD) with auxiliary screens to provide displays of the airspace at a different resolution, meteorological radar, and to support voice switching and control. The principal tools provided by TAAATS are electronic flight progress strips, a graphic air traffic picture supported by highly interactive tools, and integrated software applications for managing flow.
By using different sectors we could examine whether task load factors causing workload can be generalized across sectors or whether they are specific to particular sectors. The sectors selected were radar-controlled feeder sectors in the northern region of the Australian airspace. Burnett (BUR) is a feeder sector on the north side of Brisbane (volume is 90802 nm$^3$). Gold Coast (GOL) is a feeder sector on the south side of Brisbane (volume is 49304 nm$^3$). For BUR and GOL, flow is controlled through a software tool (Maestro) that displays aircraft sequencing in a ladder display). Daintree (DAT) is a much larger feeder sector with dimensions (volume is 468668 nm$^3$) that encircles Cairns further north in Queensland. For DAT, flow information is not displayed on Maestro, but provided by a flow controller.

Table 1. Experimental design

<table>
<thead>
<tr>
<th>Scenario Profiles</th>
<th>N</th>
<th>Medium Non Routine</th>
<th>High Routine</th>
<th>High Non Routine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold Coast</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burnett</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daintree</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participants

Participants were 18 qualified air traffic controllers from Airservices Australia in Brisbane. Participants volunteered to participate and completed their session during a rostered shift as an alternative to working in the operational environment. Participants performed the experiment on the sector they regularly worked in the operational environment at the time of the experiment. Therefore participants were very familiar with the sector they experienced in the experiment.

Design

Sector (BUR, DAT and GOL) was a within-subjects factor and scenario (Medium Non-Routine, High Routine and High Non-Routine) within subjects in our experimental design, as shown in Table 1. Non-routine events were included as non-routine events are more likely to lead to safety compromises than high traffic alone. Participants experienced three 30-minute air traffic scenarios in the order shown.

Scenarios

The scenarios were created in close consultation with subject-matter experts from the BUR, DAT and GOL sectors. The first scenario (Medium Non-Routine) was intended to produce a medium level of task load, punctuated by two abnormal events. The second scenario (High Routine) was intended to produce a high level of task load, with no abnormal events. The third scenario (High Non-Routine) was intended to produce a high level of task load punctuated by two abnormal events. The specific abnormal events were different for each sector.

Procedure

In each 30-minute scenario, ATCos controlled traffic as they normally would. Every 2 minutes (plus or minus 10 seconds, depend on ongoing ATCo communications) the experimenter, sitting nearby, asked “workload?” The participant responded with a number from 1 to 10. A scale indicating the mapping of numbers to experienced workload was in front of the ATCo at all times for reference (see Figure 1).

After each 30-minute scenario was completed, a video-cued recall session followed, lasting up to 40 minutes. The experimenter replayed a video recording of the ASD screen and asked ATCo “What is it about the situation now that made you rate it at x?” where x was a workload level from 1 to 10.

We collected scenario event data from the TAAATS simulator in two forms. First, from Eurocat Track Data we could extract a first approximation to the task load variables used by many other researchers studying mental workload in ATC. This let us estimate task load using the same traffic factors that are used in the WAT algorithm. Second, Eurocat Trace Data offered a more detailed view of human-system interaction during the sessions.

Results

Figure 2 shows results for ATWIT ratings. ATCos’ workload ratings are highly sensitive to changes in conditions from the start to the end of each scenario. Workload ratings are lower for the larger, more dispersed DAT sector (lowest line in each graph with square points) than for BUR and GOL (upper two lines).
Figure 2. ATWIT subjective workload ratings for (L to R) Medium Non-Routine, High Routine and High Non-Routine scenarios. Error bars are 95% confidence intervals. ATWIT ratings range 1-10.

Table 2. Correlations between ATWIT ratings and various measures of task load or workload. Small grey numbers are not significant. Bold numbers are the highest correlation for the row indicated.

<table>
<thead>
<tr>
<th>Scen/Sector</th>
<th>N</th>
<th>ACIN</th>
<th>AVDIST</th>
<th>TC/SV</th>
<th>S5</th>
<th>S10</th>
<th>S40</th>
<th>S70</th>
<th>WAT-U</th>
<th>WAT-W</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>739</td>
<td>.40</td>
<td>-.37</td>
<td>.59</td>
<td>.50</td>
<td>.07</td>
<td>.33</td>
<td>.58</td>
<td>.55</td>
<td>.57</td>
</tr>
<tr>
<td>BUR</td>
<td>297</td>
<td>.66</td>
<td>.15</td>
<td>.60</td>
<td>.66</td>
<td>-.06</td>
<td>.11</td>
<td>.62</td>
<td>.53</td>
<td>.57</td>
</tr>
<tr>
<td>DAT</td>
<td>264</td>
<td>.44</td>
<td>-.31</td>
<td>.42</td>
<td>.44</td>
<td>.09</td>
<td>.24</td>
<td>.30</td>
<td>.39</td>
<td>.42</td>
</tr>
<tr>
<td>GOL</td>
<td>178</td>
<td>.51</td>
<td>-.16</td>
<td>.44</td>
<td>.51</td>
<td>.15</td>
<td>.49</td>
<td>.48</td>
<td>.46</td>
<td>.41</td>
</tr>
<tr>
<td>MedNR</td>
<td>252</td>
<td>.24</td>
<td>.13</td>
<td>.54</td>
<td>.52</td>
<td>.06</td>
<td>.23</td>
<td>.69</td>
<td>.68</td>
<td>.58</td>
</tr>
<tr>
<td>HIR</td>
<td>236</td>
<td>.28</td>
<td>-.66</td>
<td>.72</td>
<td>.62</td>
<td>.19</td>
<td>.35</td>
<td>.63</td>
<td>.53</td>
<td>.68</td>
</tr>
<tr>
<td>HNR</td>
<td>251</td>
<td>.32</td>
<td>-.37</td>
<td>.48</td>
<td>.31</td>
<td>.10</td>
<td>.45</td>
<td>.51</td>
<td>.55</td>
<td>.48</td>
</tr>
</tbody>
</table>

Table 2 shows the pattern of Pearson product-moment correlations coefficients between different objective measures of task load and the ATWIT ratings. Row 1 shows the correlations when all sectors and scenarios are included in the analysis. Rows 2-4 show the correlations within each sector, and rows 5-7 show correlations within each scenario.

Simple traffic measures

Correlations of ATWIT with simple measures of task load are at left in Table 2. ACIN, the number of aircraft in the sector, is the best predictor of workload within sectors. The average distance (AVDIST) between aircraft in the sector correlates negatively with ATWIT ratings. The ratio of ACIN and AVDIST is a strong predictor of ATWIT ratings, as is traffic count divided by sector volume (TC/SV).

WAT algorithm

Table 2 shows the correlation of ATWIT ratings with WAT task load using the Airservices subject-matter expert weights (WAT-W) and unit weights (WAT-U). Figure 3 shows task load estimated every two minutes using the WAT-W algorithm with Eurocat Track Data from our scenarios.

Figure 3. WAT-W measure of task load for (L to R) Medium Non-Routine, High Routine and High Non-Routine sectors. Error bars are 95% CIs.

When compared with ATWIT results (Figure 2) the WAT algorithm clearly under-predicts the steep rise in ATWIT ratings in the MedNR and HiNR scenarios compared with the HiR scenario. Not only are non-routine events not greatly exacerbating workload ratings, but they are not reflected in WAT task load measures.

As can be seen in Table 2, WAT-U correlates slightly more strongly with ATWIT ratings than WAT-W does. In order to improve the fit of the WAT, we ran several multiple regression analyses that let us derive the best weights for predicting ATWIT ratings. The best model used simple load factors rather than ratios, plus the average distance between aircraft:

ATWIT-Raw = .4 * AC in sector (ACIN) + .11 * AC entering sector (ACEN) + -.01 * AC at top of climb (ACTOC) + .02 * AC at top of descent (ACTOD) + .13 * AC exiting sector (ACEX) + -.05 * AC in conflict (ACIC: 15 nm, A020) + -.41 * average distance (AVDIST)

ACIN is clearly more important than anticipated in the original weightings. AVDIST is a heavily weighted predictor of ATWIT ratings, especially when BUR, DAT and GOL sectors are used. The above equation leads to a multiple R of .63, which explains 40% of the variation in ATWIT ratings—better than the 27% with the original equation. However simpler regression with ACIN and AVDIST alone accounts for 37% of the variance.

Proactive control

The values S5, S10, S40 and S70 in Table 2 show the number of aircraft within 5, 10, 40 or 70 nm laterally of each other, whether or not under the ATCo’s jurisdiction. S5 has little correlation with ATWIT ratings, S10 somewhat more, and S40 and S70 even more. This suggests that workload is not driven by the number of aircraft in close lateral proximity—
indeed, separation is usually already assured. Instead, workload may be driven by the number of aircraft still at some distance from each other, when possible future breakdowns of separation must be estimated.

The video-cued recall provides further clues as to why S40 and S70 are better predictors of workload. Controllers report that the nature of their work changes as events unfold during a scenario. There seem to be four different stages to an event—e.g. an aircraft pair where a controller needs to intervene to maintain separation assurance—that requires a controller’s attention, summarized as follows:

1. “There is some stuff coming.” In this stage the aircraft involved in the event are still far apart and potentially still in an upstream, adjacent, sector.
2. “How is this going to work?” The potential for a problem is acknowledged by the controller but it is impossible to implement a solution yet, because it is still uncertain how the problem will pan out exactly. At this stage the aircraft are still well apart and there generally is no urgency of intervention.
3. “That should do it.” At this stage the aircraft are sufficiently close that the problem is more constrained and hence forming a solution is more viable.
4. “See how it works out.” This last stage is the actual implementation of the solution. This consists of providing the instruction to the pilots and then monitoring that the intervention actually has the desired effect.

These stages reflect the high value that controllers put on proactive control. They try to become aware of potential problems early—often before aircraft even call onto the frequency or are in a situation that would be captured by traffic load factors. At that stage, the details of the problem are still unclear, so ATCos hold off intervening until the event has unfolded far enough to warrant intervention.

Under low workload a controller will have sufficient time to monitor a given problem and see how the details unfold. When workload increases the ability to monitor problems can be severely affected.

Proactive control in KJN-AMR event

As an example of proactive control and some of its difficulties, see Figure 4. A participant gave AMR—the aircraft outlined just to the north of the KJN hold area—a direct route to its destination. At that time there were a couple of northbound aircraft coming into the sector that needed to hold, including KJN (KJN is the aircraft outlined to the south of the hold area). The participant had assigned FL180 to both KJN and AMR. At this stage he had not fully appreciated the trajectory KJN will take once it enters the hold.

![Figure 4. Start of the KJN-AMR event](image)

During the next three minutes he mostly attends to the aircraft that need to hold and he gives instructions to six other aircraft before he notices that separation is not assured between AMR and KJN. When KJN is about to enter the hold (Figure 5), he becomes fully aware that with the speed differential—KJN is a jet and AMR a propeller driven aircraft—KJN will catch up with AMR. He quickly instructs AMR to continue on its present heading, keeping it clear of the hold area of KJN. Then he waits almost two more minutes before AMR can resume its route direct to its destination.

![Figure 5. Three minutes later in the KJN-AMR event](image)
In this scenario, with multiple ongoing events, the controller had to monitor a number of problems simultaneously. When he diverted his attention to another event he could have missed critical developments in the problem between AMR and KJN. In the end he had to switch from a proactive approach—giving AMR a shorter route for fuel savings—to a more reactive approach to keep these two aircraft safely separated.

In the analysis of the video-cued recall it seems that stages 2 and 3, as described above, are critical in the problem solving process. This is when controllers try to unravel the inherent uncertainties of a problem before they intervene. Stages 2 and 3 seem to capture a large part of the subjective workload of controllers. For the sectors we studied, stages 2 and 3 seem to happen when aircraft are still between 40 to 70 nm apart. As shown in the example, this is also when they seem most susceptible to disruptions. However these factors are not accommodated in task load-based measures of ATCo workload.

Conclusions

Our high fidelity simulator experiment illustrates the difficulty of capturing workload using linear combinations of traffic load factors, even though we were able to collect in-the-loop workload ratings by ATCos on the sectors for which they were current. As for prior studies, the overall variance in workload ratings that could be accounted for with the task load factors at hand to date was 40%. While further analysis and the extraction of more sensitive factors may improve that result, we do not expect it to rise above the 50% commonly observed in this area.

Interestingly, our manipulation of routine vs. non-routine events across the scenarios did not have a measurable effect on subjective workload. When traffic load is already moderately high to high, the extra workload imposed by the non-routine event may be subjectively smaller than otherwise, give the high existing level of arousal and the probable selection of efficient strategies that preserve safety at the expense of maximizing expeditiousness or orderliness. The impact of non-routine events should be examined systematically with lower traffic load and lower probably ATCo arousal.

Our video-cued recall data gave ample evidence of ATCos' strong orientation towards the future. Traffic factors that are the objects of proactive control are not necessarily captured effectively in linear regression models. Our exploratory use of the S40 and S70 metrics, and their greater correlation with rated workload than S5 and S10, indicate the role of anticipation in ATCos' activity and its contribution to their workload. In ongoing work we are performing further detailed analyses of the video-cued recall data from the study with a view to clarifying further (1) the reasons that task load measures and traffic configuration measures fall short when estimating ATCo workload and (2) the relationship between subjective workload and air traffic safety. As suggested in Loft et al. (in press) if the ATCo's goal is to keep workload under control, then variation in strategy may reflect task load factors more than workload does.

References


