

HWORLOAD 2017 – The relationship between workload and performance in air traffic control: Exploring the influence of levels of automation and variation in task demand

Tamsyn Edwards¹, Lynne Martin², Nancy Bienert¹, Joey Mercer²

¹ San Jose State University at NASA Ames Research Center,
Moffett Field, CA, USA

² NASA Ames Research Center,
Moffett Field, CA, USA

{tamsyn.e.edwards, lynne.martin, nancy.bienert, joey.mercer}@nasa.gov

Abstract. In an air traffic environment, task demand is dynamic. However, previous research has largely considered the association of task demand and controller performance using conditions of stable task demand. Further, there is a comparatively restricted understanding of the influence of task demand transitions on workload and performance in association with different types and levels of automation that are available to controllers. This study used an air traffic control simulation to investigate the influence of task demand transitions, and two conditions of automation, on workload and efficiency-related performance. Findings showed that both the direction of the task demand variation and the amount of automation influenced the relationship between workload and performance. Findings are discussed in relation to capacity and arousal theories. Further research is needed to enhance understanding of demand transition and workload history effects on operator experience and performance, in both air traffic control and other safety-critical domains.

Keywords: Air traffic control, Workload transitions, Workload history, Time-based metering, Automation

1 Introduction

Within the safety critical domain of air traffic control (ATC), workload “is still considered one of the most important single factors influencing operators’ performance” [1 p639]. Workload has been defined within the ATC domain as the “activities, both mental and physical, which result from handling air traffic” [2, p3]. Air Traffic Controllers’ (ATCOs’) primary task is to ensure the safety of aircraft in their airspace [3]. They have to ensure at least standard separation between the aircraft in the airspace (sector) for which they are responsible, which includes changing the course of one or more aircraft if they predict that the paths of these aircraft will, in the future, come too close together (conflict). Secondly, controllers strive to efficiently manage their traffic, which, in airspace where aircraft are

descending to arrive at an airport, includes creating strings of evenly spaced aircraft to assist in maximizing landings. ATCO tasks can be thought of as a series of speed-time-distance trigonometry problems. Thus, their workload stems mainly from cognitive demands, and is “mental” in nature, although a sector that has many aircraft entering and exiting can have a high physical load, in terms of the communication required with pilots.

Although there are many factors that can increase the complexity of an event for a controller (e.g., sector structure, weather), amount of controller workload is closely related to traffic density¹. While there are procedures in place to limit traffic density becoming too great in any one sector, controllers also manage task demand by employing a range of strategies [4]. This behavior can be described by resource theory [5], which assumes that the human operator has a limited capacity of cognitive resources available to be allocated to a task. More tasks are understood to demand more processing resources. At some point, the number of tasks lead to demands greater than the resources available, and performance suffers, unless the operator (in this case the ATCO) can change the task demand on cognitive resources. In ATC, safety performance is paramount, and so ATCOs develop a range of strategies to manage the demands of the task and therefore, the available cognitive resources, as observed by [6, 7].

In ATC, as with many other safety critical environments, task demand and workload are dynamic. ATCOs frequently experience changes in traffic load and the complexity of the traffic situation. These changes in task demand can potentially result in changes to the cognitive complexity of managing the traffic and subsequently, ATCOs’ subjective experience of transitions between high and low workload. These transitions can be expected by the controller, such as when traffic load changes based on the time of day or known activities in surrounding sectors, or unexpected, for example, through increased complexity resulting from an emergency situation. Transitions may also be gradual or sudden [8]. Controllers, therefore, have to remain vigilant at all times when they are ‘on position’ to make sure they are aware of events as they build, even if the transition is sudden.

Research on task demand transitions, and the affect on both performance-influencing covariate factors (such as workload) and task performance is limited, with studies frequently utilizing a constant task demand [9] or changing demand only between experimental conditions. Of the research available on demand transitions, there appears to be conflicting findings. Some (e.g. [10]) have reported that overall performance efficiency on a vigilance task was not affected by task demand transitions, regardless of whether the transition was expected or unexpected. However, others (e.g. [10]) have found that performance on vigilance tasks was influenced by a low-to-high demand transition or high-to-low demand transition (e.g. [8]). Task demand and workload transition research specific to an ATC environment is particularly underrepresented. Consequently, there is limited understanding of the influence of demand transitions on workload and performance in air traffic environments.

To contribute to understanding in this domain, [12] reported on a study that investigated task demand transitions on workload, fatigue and an efficiency

¹ Traffic density refers to the number of aircraft an ATCO is managing in their sector.

performance measure, metering accuracy. Findings showed that a change in task demand appeared to affect both workload and fatigue ratings, although not necessarily performance. In addition, participants' workload and fatigue ratings in equivalent task demand periods appeared to change depending on the demand period preceding the time of the current ratings. However, the findings reported specifically focused on a scenario in which the controller had full manual control. In both the current and future planned (i.e., NextGen) air traffic systems, automation is increasingly present to both assist (i.e., the ground based separation assurance tools offered to air traffic controllers in studies reported by [13]), and in some cases, take over (e.g., automated handoffs), controller tasks. In order to increase National Airspace (NAS) capacity, it is therefore important to investigate the association of taskload variations, and taskload after-effects, with both current-day manual tasks and tasks with functions that will potentially be automated in the future.

As [14] discuss, there can be a tradeoff for the operator between the situation awareness (SA) that is generated by completing tasks and the accompanying workload and time pressure. Automation adds another layer to tradeoff considerations because, if implemented with the human/ automation system in mind, automation can offer situation awareness-enhancing qualities, such as predictability and integrated information [14], which together help the human to build and maintain situation awareness.

It is important to understand for which tasks air traffic controllers can continue to be an effective part of the separation assurance system and which tasks are now more suitable for automation. The tradeoffs between the amounts of automated aid with human involvement in air traffic management performance were explored in a series of three studies, of which the third is reported in detail below. The addition of automation (that redefines a human system as a human/automation system) is intended to aid human performance and increase system capacity.

The data reported in this paper was generated from a larger study reported in [9]. The authors extend the findings reported in [12] by investigating the association of differing levels of automation on workload and efficiency-related performance in an ATC simulation. The aim of the study reported here was to investigate the influence of expected and gradual task demand transitions (high-low-high and low-high-low) on workload and performance under two different levels of automation, within a high fidelity ATC simulation environment. Due to the quantity of measures and data generated from this study, only a subset of the measures and findings that are most relevant to this research aim are presented. Initial findings are reported in [12] which are extended in the current paper.

2 Method

2.1 Design Overview

A within measures, en-route ATC human in the loop (HITL) simulation was utilized to investigate task demand variation on workload and performance. Participants operated a combined low and high altitude sector in Albuquerque Center (ZAB) and

were assigned to meter aircraft into Phoenix (PHX) and manage overflights. Metering is a specific controller task of scheduling arrival traffic to meet a pre-planned schedule or time. Task demand was manipulated to create two scenarios. Efficiency-related performance was inferred from delay to metered aircraft (in seconds) at three nautical miles before a meter fix). Participants were eight retired-ATCOs who had previously worked in enroute airspace in Oakland Air Route Traffic Control Center (ARTCC). Pseudo pilots were paired with controllers, and completed standard pilot tasks such as controlling the aircraft in accordance with controller instructions and communicating with controllers. Each simulation session lasted for 90 minutes.

2.2 Airspace and Task Demand Scenarios

Participants operated a simulated, combined low and high altitude sector (segment of airspace, Figure 1), in Albuquerque Center (ZAB) that handles aircraft beginning their arrival descent into the Phoenix Sky Harbor International Airport (PHX). This airspace was selected for the complex mix of arrivals and overflights. Scenarios were designed to have the mix of traffic present in this sector – overflights passing through at level altitudes and transitioning aircraft either climbing out from PHX and other airports in the area, or on a metered descent into PHX. The scenarios included winds for the area, which were constant-at-altitude with a nominal forecast error.

Arrival traffic in both scenarios was metered through the HOMRR fix on the EAGUL6² arrival (Fig. 1). Aircraft were initiated in the scenario with up to two-minute delays ($M=76$ sec) as they entered the sector (on the right of Fig. 1). In addition, nine conflicts were created in each scenario where an overflight would lose separation with another overflight or an arrival if not adjusted. In the Start High scenario, four conflicts were built to occur in the first thirty minutes, two in the second thirty minutes, and three in the final thirty minutes. In the Start Low scenario, three conflicts were built to occur in each of the three thirty-minute segments.

² Lining aircraft up for a runway begins many miles out from an airport as aircraft begin their descent. Aircraft begin to fly on more formalized waypoint-to-waypoint routes that “channel” them to a runway. Each set of routes is given a name, in this case “EAGUL”

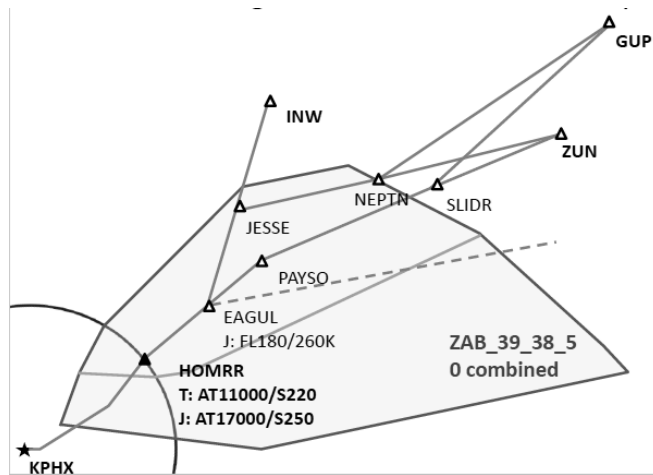


Fig. 1. Low-high altitude sector (shaded in grey) in ZAB with the routes that comprise the “EAGUL6 STAR” marked

The direction of the task demand transition was manipulated to create the two scenarios. Scenario 1 followed a high-low-high task demand pattern and scenario two followed a low-high-low task demand pattern. The creation of three task demand periods was implemented in order to better reflect the multiple task demand transitions that can be experienced within an operational environment. In addition, this permitted an extension of previous studies that had focused on the comparison of workload and performance for one transition period (e.g. [8]).

Each simulation session lasted for 90 minutes and consisted of three, 20 minute [15] periods of stable task demand which alternated between high and low traffic levels, interspersed with a total of three, 10 minute transition phases. Task demand was created by the number of aircraft under control [16] as well as the ratio of arrival aircraft and overflights. Arrival aircraft create complexity in the task, which also influences task demand. Task demand phases for equivalent stable task demand periods (i.e., high demand regardless of which scenario the high demand was positioned in) were created using the same aircraft counts and number of arrival aircraft, permitting comparability between demand variation scenarios. Scenarios followed a counterbalanced presentation.

2.3 Study Condition – Amount of Automated Support

Automation was introduced into the study to different extents to create three conditions: a manual condition, an arrival manager (AM) condition, and fully automated condition. The fully automated condition will not be reported during this paper as there was no measure for metering performance. Instead, the focus is on comparing subjective workload and controllers’ metering performance in the Manual and Arrival manager (AM) conditions only.

In order to compare the effects of different levels of automation on subjective measures and performance, key ATC tasks were identified and assigned to “the automation” (actually a suite of tools) or the controller. The key tasks were conflict detection, conflict resolution, arrival metering (schedule conformance), and monitoring the automation while it was completing these tasks. Other ATC housekeeping tasks, including handoffs, frequency changes, and climb and descent clearances, were automated for all conditions and the controller had to monitor these for all conditions.

The four key tasks were combined in the study conditions. The first “mostly manual” condition was close to current day operations where the participants worked all four key tasks (including monitoring the automated housekeeping tasks). In the second, mid-level decision support condition (Arrival manager or “AM”), participants were responsible for “metering” and monitoring the automation. Metering refers to the controller task of contributing to arrival traffic schedule conformance. In this case, controllers in this low-altitude en route sector are required to deliver the PHX arrival traffic to meet a schedule. The scheduler is spacing aircraft to assure well-spaced runway arrivals. The controller does not have to put the aircraft exactly on time but has to deliver them within a plus or minus (+/-) 30 sec window across a waypoint (HOMRR) that is at the lower left of the sector. The automation was allocated the tasks of conflict detection and resolution (CD&R) and housekeeping. The algorithms that alerted and resolved strategic conflicts (looking from 3 to 12 minutes ahead) were based on the Automated Airspace Concept [17, 18] and the tactical CD&R automation (looking 0-3 minutes ahead) was based on TSAFE [19].

During the study, each participant worked with each of the automation conditions for four runs. For half of the runs they worked the Start High traffic scenario and for the other half they worked the Start Low traffic scenario. Combined, this was a 3x2 design (level of automation by traffic density), which was repeated to give a data set of twelve 90-minute runs. It was predicted that the increased amount of automation would be reflected in lower workload ratings from the participants and increased efficiency performance, measured by greater schedule conformance (more arrival aircraft crossing the meter-fix in the +/- 30 sec window).

2.4 Participants

A total of eight male retired-controllers took part in the simulation. Age ranged from 50 years – 64 years. Participants responded to grouped age ranges and so an average age could not be calculated. Participants had worked as en-route controllers in the Oakland, California, ARTCC. Participants’ years of experience as active ATCOs (excluding training) ranged from 22 – 31 years (M=26.56, SD=3.90).

2.5 Procedure

Participants were asked to work the traffic, as they would normally do, ensuring separation and metering the arriving aircraft to deliver them within a +/- 30 sec delay

window across the HOMRR fix. It was emphasized that the participants could work any of the traffic at any time in any condition if they wanted. That is, for the conditions with greater amounts of automation, the controller could intervene if they did not think the automation was going to achieve the separation criteria. In addition to the primary tasks, the participants completed two other sets of tasks. Firstly, they were prompted to rate their workload and then answer a situation awareness question every three minutes for the duration of each run. Secondly, they were asked to verbalize whenever they saw a “glitch” in the software, e.g., an aircraft not behaving as directed or overcorrecting.

The study was run over five consecutive days. The first day and a half was devoted to training the participants on the study environment and procedures. After an initial briefing, six training scenarios were run with increasing levels of traffic and complexity (two 45 minute training runs and four 90 minute training runs). Beginning in the afternoon of the second day, participants worked 13 data collection runs (12 planned runs and one repeat). They completed workload and awareness scales during each run and questionnaires at the end of each run, as well as a post-simulation questionnaire. The last session on the fifth day was a debrief that provided an additional opportunity for participants to offer feedback. As four of the twelve runs were under the “fully automated” condition, which incorporated metering in a different fashion to the other two conditions, these four runs were removed from the data for the analysis presented below.

Data from workstation logs and controller responses were analyzed from eight runs for each participant. The results section below compares data across the levels of automation to describe the relationships between automation and performance for efficiency. The discussion explores relationships between the performance factors.

3 Results

3.1 Task Demand Variation Manipulation Check

A review of the descriptive statistics suggests that task demand did vary in the intended direction (Fig. 2). Figure 2 confirms that the number of aircraft in the controller’s sector were similar between equivalent task demand periods regardless of scenario (high-low-high demand or low-high-low demand). The number of arriving aircraft was also similar.

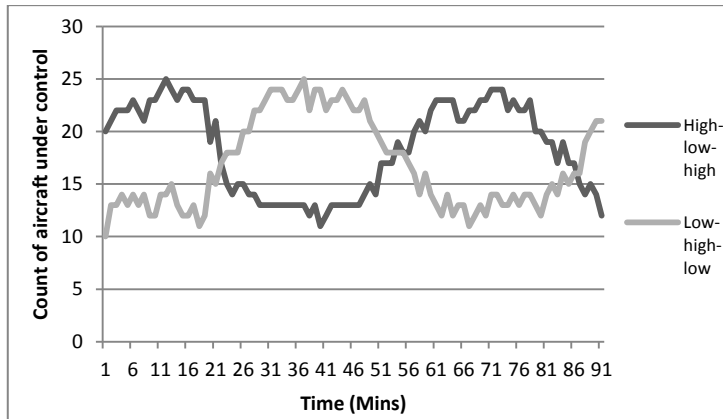


Fig. 2. Count of aircraft under control by minute for scenario 1 (high-low-high demand) and scenario 2 (low-high-low demand).

3.2 The Relationship between Taskload and Workload

Two sets of data were chosen for comparison – participants’ perceived workload, recorded through a real-time rating that indicated how controllers thought they were managing the scenario demands, and a task performance metric of schedule conformance that indicated how well the human-automation system was maintaining the delay goals for the sector.

Participants rated their workload in real time using an ISA-type rating scale and prompt. Every three minutes during a run, when the scale illuminated on the workstation banner, they rated their level of workload between 1 (very low) and 6 (very high). Fig. 3 and 4 show the mean perceived workload ratings at each time point during the runs split by the type of scenario and then plotted by the two task-sets that the controllers were given (Manual and Arrival manager). Overall, participants rated themselves as having low to moderate workload during the H-L-H scenario, with the lowest mean rating being 2.5, and the highest 4.1, out of a possible 6 (Fig. 3). Mean ratings for the Arrival manager task set were very similar to those given for the Manual task set. During the L-H-L scenario (Fig. 4), participants also rated their workload, on average, as moderate to low, with the lowest mean rating being 2.0, and the highest 3.6. Mean ratings for the two task sets were not so similar for this traffic scenario. The mean workload reported under the AM task set was consistently slightly lower than that given for the Manual task set.

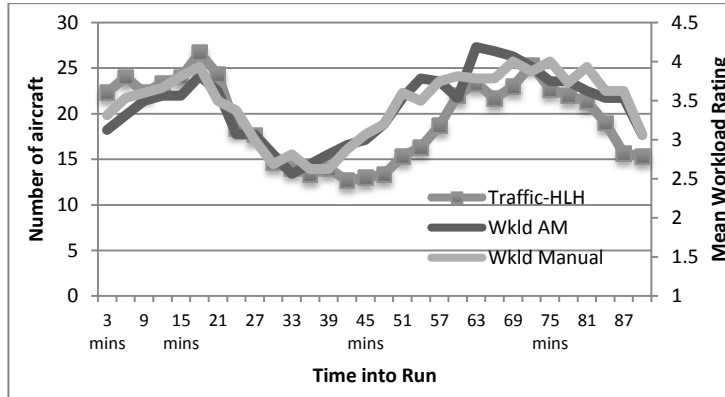


Fig. 3. Mean real time workload rating of the AM and manual conditions during the High-Low-High traffic scenario

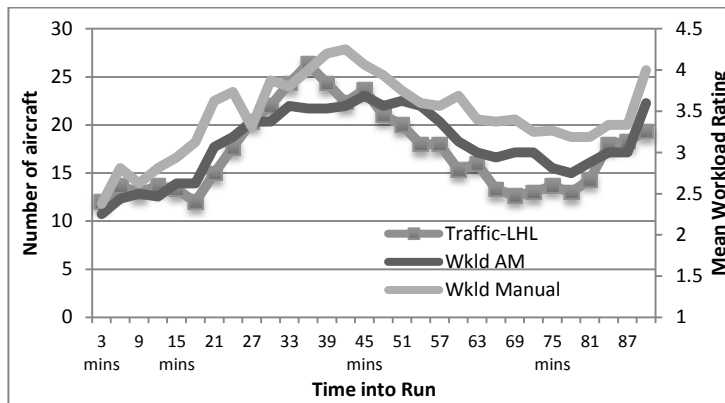


Fig. 4. Mean real time workload rating of the AM and manual conditions during the Low-High-Low traffic scenario

As the level of traffic in the scenario was assumed to be one of the main influences on workload, the number of aircraft in each scenario (traffic count) is also plotted in Fig. 3 and 4. The correspondence between workload ratings and traffic count is very high (please note the two y-axes in the figures). Significant, positive relationships were found between the traffic count and workload ratings for both the AM condition ($r=0.71$, $p<0.001$) and Manual condition ($r=0.79$, $p<0.001$) in the High-Low-High demand scenario and workload and the AM condition ($r=0.85$, $p<0.001$) and Manual condition ($r=0.81$, $p<0.001$) in the Low-High-Low demand scenario. One point of interest is that, although the curves of the mean workload and traffic lines are very similar for both traffic scenarios when the traffic is increasing to a “High” phase, the mean workload reported begins to rise slightly before the traffic (see 39-63 mins on Fig. 3 and 15-33 mins on Fig. 4). Conversely, when the traffic is decreasing to a “Low” phase, the mean workload reported begins to decline slightly after the traffic (see 72-90 mins on Fig. 3 and 42-69 mins on Fig. 4).

To further investigate any differences between the task demand and automation conditions on reported workload, a one-way repeated measures analysis of variance (ANOVA) was conducted for each scenario. [12] first reported the findings for the manual condition which are repeated below, but did not extend the analysis to comparison with the arrival manager (AM) condition. Therefore the following analysis extends previous findings.

3.2.1 The Relationship between Taskload and Workload in the AM condition

Workload ratings were averaged across the 20 minute periods of stable task demand to facilitate comparison between the separate task demand periods. A review of the descriptive statistics (Table 1) suggests that workload in both demand scenarios varied as expected with task demand. In the high-low-high demand scenario (scenario 1) workload appears to be rated slightly higher in the third task demand period (high demand) compared to the first task demand period (high demand). In the low-high-low demand scenario (scenario 2), workload was rated highest in the high demand phase. However, on average, participants perceived workload to increase in the second low demand period compared to the first. Comparing across low demand periods between conditions, workload is rated similarly in the first period of scenario 2 and the middle period of scenario 1. However, the low demand period in the third period of scenario 2 is rated as higher workload than either of the other low demand periods.

Table 1. Mean and standard deviation for workload (as rated by ISA) in both transition phases for the AM condition.

| Workload (ISA) | Task demand period 1 (0-20 minutes) | | Task demand period 2 (31-50 minutes) | | Task demand period 3 (61-80 minutes) | |
|-------------------------------------|-------------------------------------|------|--------------------------------------|------|--------------------------------------|------|
| | M | SD | M | SD | M | SD |
| Scenario 1 workload (High-low-high) | 3.48 | 0.82 | 2.85 | 0.77 | 3.98 | 0.76 |
| Scenario 2 workload (Low-high-low) | 2.59 | 0.62 | 3.60 | 0.69 | 3.01 | 0.78 |

To further examine the changes in perceived workload, a one-way repeated measures analysis of variance (ANOVA) was conducted for each scenario [5]. In the high-low-high demand AM condition, Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2) = 7.08, p < 0.05$); therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($E = 0.59$). The results show that there was a significant main effect of task demand period on self-reported workload $F(1.18, 8.27) = 28.79, p < 0.001$. Pairwise comparisons revealed that workload was significantly lower in task demand period 2 (low demand) than high task demand period one ($p < 0.005$) and three ($p < 0.001$). Workload was not rated significantly differently between high demand period 1 and high demand period 3 ($p = 0.2$). In scenario 2 (low-high-low demand) a significant main effect of task

demand period was found on self-reported workload $F(2,14) = 11.18, p < 0.005$. Pairwise comparisons revealed that workload was rated significantly higher in the high demand period than the first low demand period ($p < 0.05$) and was not significantly higher than the final low demand period ($p = 0.13$). Workload ratings in the second low demand period were not significantly higher than the first low demand period ($p = 0.061$).

3.2.1 The Relationship between Taskload and Workload in the Manual condition

Workload ratings were again averaged across the 20 minute periods of stable task demand (Table 2). A similar pattern of workload between demand scenarios was seen in the manual condition and the metering condition. A review of the descriptive statistics (Table 2) suggest that workload in both scenarios varied as expected with task demand. In scenario 1 (high-low-high demand) workload appears to be rated slightly higher in the third task demand period (high demand) compared to the first task demand period (high demand). In scenario 2 (low-high-low demand), workload was rated highest in the high demand, second task demand phase. However, on average, participants rated perceived workload to increase in the third task demand period (low demand) compared to the first low demand period. Comparing between scenario 1 and 2, the high demand period is perceived to generate the most workload for participants in the low-high-low demand scenario, although the high demand periods were objectively equivalent between scenarios. Comparing across low demand periods between conditions, workload is rated similarly in the first period of scenario 2 and the middle period of scenario 1. However, the low demand period in the third period of scenario 2 is rated as higher workload than either of the other low demand periods.

Table 2. Mean and standard deviation for workload (as rated by ISA) in both demand transition scenarios for the manual condition.

| Workload (ISA) | Task demand period 1 (0-20 minutes) | | Task demand period 2 (31-50 minutes) | | Task demand period 3 (61-80 minutes) | |
|-------------------------------------|-------------------------------------|------|--------------------------------------|------|--------------------------------------|------|
| | M | SD | M | SD | M | SD |
| Scenario 1 workload (High-low-high) | 3.67 | 0.77 | 2.87 | 0.61 | 3.85 | 0.62 |
| Scenario 2 workload (Low-high-low) | 2.78 | 0.64 | 4.06 | 0.71 | 3.33 | 0.61 |

A repeated measures ANOVA was applied to each scenario, to explore differences within-scenarios. In relation to scenario 1 (high-low-high demand) a significant main effect of task demand period was found on self-reported workload $F(2,14) = 44.23, p < 0.001$. Pairwise comparisons revealed that workload was significantly lower in task demand period 2 (low demand) than high task demand period one ($p < 0.005$) and three ($p < 0.001$). Workload was not rated significantly differently between high demand period 1 and high demand period 3 ($p = 0.68$). In scenario 2 (low-high-low demand) a

significant main effect of task demand period was found on self-reported workload $F(2,14) = 32.72, p < 0.001$. Pairwise comparisons revealed that workload was rated significantly higher in the high demand period than the first low demand period ($p < 0.001$) and second low demand period ($p < 0.005$). It was also identified that the workload ratings in the second low demand period were significantly higher than the first low demand period ($p < 0.05$).

3.2.3 Workload across Demand Scenarios and Automation Conditions

Figure 5 presents a comparison of the mean workloads for the task demand periods for the task demand transition direction variable (low-high-low and high-low-high) and the automation variable (AM or manual). It is interesting to note that based on the descriptive statistics, workload ratings in the low-high-low demand scenario overall are lower for the AM condition than the manual condition. The same pattern is not seen for the high-low-high scenario. In addition, the high workload period in the low-high-low manual condition is rated higher than either of the high workload periods in the metering and manual conditions for the high-low-high scenario,

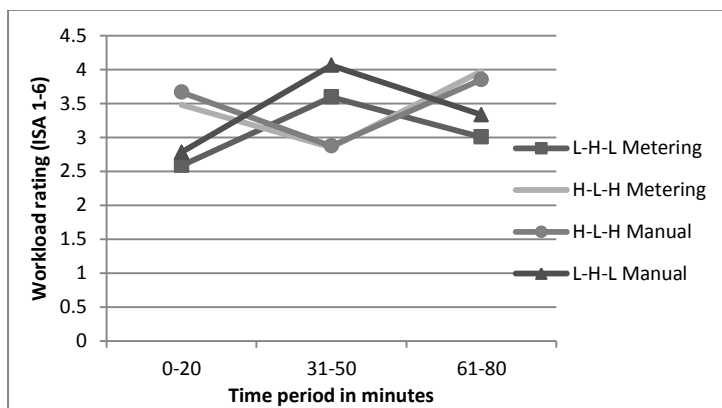


Fig. 5 Mean metering delay under two taskloads during the H-L-H traffic scenario

3.3 The Relationship between Taskload and Task Performance

The metering task was to reduce the scheduled delay on the arrival aircraft to meet the delay goal of being within +/- 30 sec of the scheduled time at the HOMRR waypoint. The controller was required to do the metering with only the help of a trial planning function – a tool that marked on the sector display predicted route of the aircraft. The meter-fix accuracy metric describes an aircraft’s successful delivery at HOMRR. Aircraft crossing the meter-fix were counted as successful if the aircraft arrived within +/- 30 seconds and crossed within the 3nmi gate around HOMRR.

Overall, 90.4% of the aircraft were successfully delivered across the HOMRR meter-point. Approximately the same percentage of flights was successfully delivered under the two task-sets (91.0% and 90.4%). On average (when the mean was calculated with absolute values), aircraft in the Manual condition were delivered with 11.9sec of delay and they had 10.7sec of delay in the Arrival manager condition. The mean delay over time per task-set was calculated and is charted in Fig. 6 and 7. The pattern of delay for both task sets during the H-L-H scenario (Fig. 6) is similar, with larger mean delays occurring when the traffic is High, and lower mean delays during the Low traffic in the middle of the runs. While there is a good amount of variation in the delay over the meter fix, the goal for the arrivals was to be within +/- 30 seconds, and at most of the time-points the average delay across the aircraft within that time bin is less than 30 seconds. It should be noted that individual aircraft within that time bin may not have been delivered successfully (under 30 sec) but that the group average is successful. For the AM task set, there is only one time point when mean delay is above 30 seconds, it is at 66 minutes into the run when the traffic load is High. For the Manual task set, there are two time points when mean delay is above 30 seconds, again traffic load is High – at 12 minutes and 66 minutes into the run. Since both sets of data show a marked increase in metering delay at the beginning of phase 3 (66 min) and in the middle of phase 1 (12 min), it is possible that the controllers were more focused on other tasks at those times and this caused their metering efficiency to reduce. In the H-L-H scenario, there were two planned conflicts between 10 and 14 minutes into the scenario, and the seventh planned conflict was at 62 minutes. It is suggested that, even when CD&R was allocated to the automation in the AM condition, the participants traded-off fine-tuning aircraft in their metering task to ensure these conflicts did not occur. However, an important difference between the Manual and AM delay is that the standard deviation of delay for phase 3 (61 to 90 min) under the Manual task set is much larger (at 21.44sec) than for the other three High phases represented in Fig. 7 (which are 13.11, 13.53 and 13.73sec respectively).

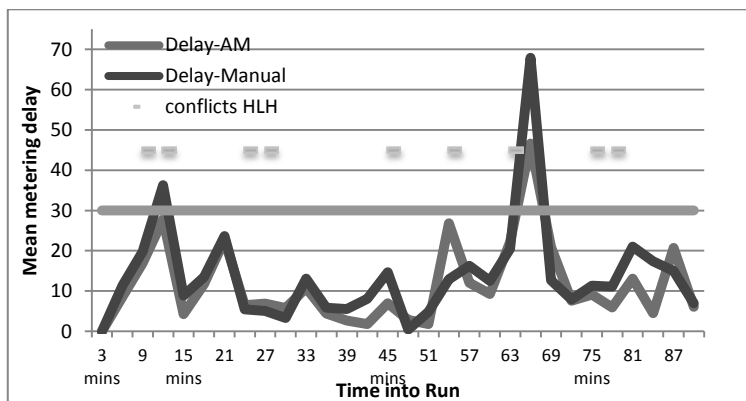


Fig. 6. Mean metering performance under two taskloads during the H-L-H traffic scenario

The pattern of delay for both task sets during the L-H-L traffic (Fig. 7) is also similar, with larger mean delays occurring in phase 3, when the traffic is Low. As for

the H-L-H traffic, there is a good amount of variation in the delay over the meter fix and, at most of the time-points, the average delay across the aircraft within each time bin is less than 30 seconds. However, although the delay patterns are similar, they seem slightly offset from each other, with delay rising or falling slightly sooner (by about 3 minutes) in the Manual condition compared to the Arrival manager condition.

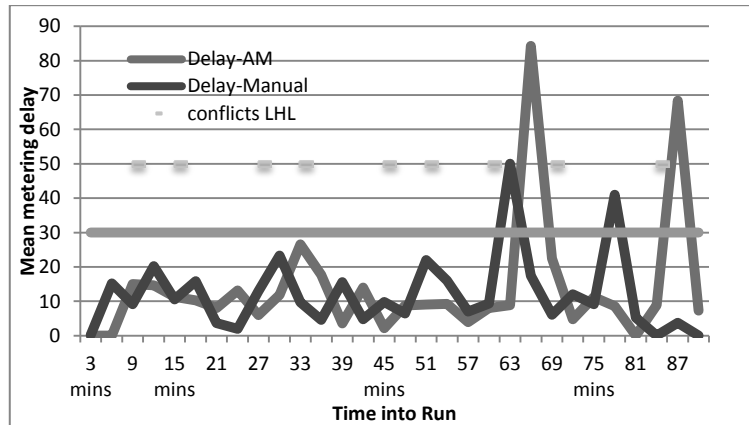


Fig. 7. Mean metering performance under two taskloads during the L-H-L traffic scenario

For both task sets, there are only two time points when mean delay is above 30 seconds; for the AM task set they are at 66 and 87 minutes into the run when the traffic load is Low or increasing; and, for the Manual task set, they are at 63 and 78 minutes into the run. In this L-H-L scenario, the last three planned conflicts were between 60 and 85 minutes into the scenario. Again, the observed decline in metering efficiency suggests that the participants traded accuracy on their metering task to ensure these conflicts did not occur. Since both sets of data show a marked increase in metering delay during phase 3 (61-90 min), the standard deviation of delay for this phase were compared. Under the Manual task set, the standard deviation of the delay in phase 3 is much larger (at 16.43sec) than for the other three Low phases represented in Fig. 7 (which are 10.15, 9.55 and 9.43sec respectively).

3.4 The Relationship between Task Performance and Workload

The main inquiry of this analysis was to explore the relationship between taskload, performance efficiency, and how these are related to perceived workload. The data that are shown in Fig. 3 to 7 above were combined to compare workload with task performance (represented by metering delay) under each traffic scenario and automation set. Table 3 compares the mean metering delay and mean workload during the H-L-H traffic. For both the AM and Manual conditions, the delay was correlated with workload across the whole scenario and then by the three phases of traffic load. There is a significant correlation of the Arrival manager task set workload with delay ($p < .01$), which is above 0.5 overall and across each phase of

traffic – broadly, as workload ratings rise and fall metering delay also rises and falls. The correlation between workload and delay in the Manual task set is lower, only 0.39 overall, but still significant ($p < .05$). While the correlations for phase 1 and 2 are close to the overall correlation, there is a noticeable reduction in the correlation during phase 3, down to 0.12 between workload and delay (Table 3).

Table 3. Correlation of Workload with Delay under H-L-H traffic load (** $p < .01$; * $p < .05$)

| | Overall correlation | Phase 1 (High traffic) | Phase 2 (Low traffic) | Phase 3 (High traffic) |
|-------------------|---------------------|------------------------|-----------------------|------------------------|
| Metering task set | 0.58** | 0.53 | 0.51 | 0.58 |
| Manual task set | 0.39* | 0.49 | 0.35 | 0.12 |

This process of correlation was completed for each task set under L-H-L traffic load (Table 4). The unexpected finding is that the correlations between workload with delay for both task sets are very low, the overall correlations are slightly negative for both task sets. Despite the correlations being so low, a slight trend similar to that in the H-L-H traffic load can be seen – the relationship between delay and workload reduces over the phases of the scenario. For both task sets, phase 3 shows the least correlation between workload and delay, which for the L-H-L traffic is negative.

Table 4. Correlation of Workload with Delay under L-H-L traffic load

| | Overall correlation | Phase 1 (High traffic) | Phase 2 (Low traffic) | Phase 3 (High traffic) |
|-------------------|---------------------|------------------------|-----------------------|------------------------|
| Metering task set | -0.10 | 0.09 | 0.10 | -0.13 |
| Manual task set | -0.09 | 0.15 | -0.26 | -0.26 |

4 Discussion

A within-measures design was used to investigate task demand variation and automation levels on subjective workload and efficiency performance, measured by delay accuracy of arrival aircraft. The direction of the task demand transition was manipulated to create two scenarios: H-L-H and L-H-L. Results showed that task demand varied as intended. Descriptive statistics confirmed that equivalent demand periods, regardless of scenario or position, were composed very similarly in terms of controlled aircraft count and arrival aircraft count. This suggests that changes in the covariates or dependent variable are unlikely to be attributed to demand differences between the created scenarios.

4.1 The Relationship between Taskload and Workload

In general, task demand and workload had high covariance for both H-L-H and L-H-L scenarios, across automation conditions. However, a key finding of interest is that perception of workload appears to differ depending on the demand period preceding the current ratings, in line with previous findings [5], and the level of automation in the control task. In the H-L-H condition, workload in the manual condition was reported on average a little higher than the Arrival manager condition, although this trend is reversed in the second high taskload period. This is an interesting data trend. As [14] discuss, more manual tasks can increase situation awareness (SA) for the operator. It may be possible that during the ramp up transition, the increased automation resulted in controllers requiring more cognitive effort to build the picture with the increasing traffic, creating a perception of higher workload.

In the L-H-L scenario there is still a high correlation between taskload and workload overall, but some differences can be observed compared to the H-L-H scenario. In the manual, L-H-L condition, workload starts low, as expected, at around an average rating of 2.5. This is similar to the workload ratings for the low taskload period in the H-L-H manual condition. However, as we get into the high workload period, the workload ratings appear to ramp up faster than in the comparison period of the H-L-H scenario. In addition, workload is rated higher than in either of the two high taskload periods in the H-L-H scenario suggesting that there is a difference in perceived workload in the ramp up phases of the L-H-L scenario compared to the ramp up phase of the H-L-H scenario. As the traffic counts were the same in all high taskload periods for all scenarios, this is likely not due objective differences in the traffic scenario. Workload is also perceived to be significantly greater in the second low demand period than the first, potentially suggesting that workload is perceived to be greater after the high demand period. This increased workload would not be the result of working to resolve delays from the previous period, as any remaining delays were absorbed in the 10-minute transition period between the stable demand periods. These findings indicate that the workload appears to be perceived differently depending on what precedes the time of rating. More specifically, results suggest that in this ATC task, a demand transition pattern of low-high-low demand may result in operators perceiving subsequent high and low demand periods after the initial low demand period as generating a greater workload than equivalent demand periods in a high-low-high demand transition pattern.

As expected, when comparing the workload ratings in the manual and Arrival manager conditions, reported workload appears to be lower in the AM condition than the manual condition in the L-H-L scenario. Interestingly, however, and unexpectedly, this finding was not replicated in the H-L-H scenario, where manual and AM conditions appear to have similar workload ratings. In addition, the workload ratings in the high taskload period of the L-H-L AM condition were lower than the high taskload periods of the H-L-H scenario. This suggests that in the L-H-L condition, the application of the automation, and associated removal of specific controller tasks, supported the controller, and possibly increased available resources [5], resulting in lower workload ratings. As the same effect of the metering task was not observed in the H-L-H scenario, it may be that the L-H-L scenario created higher demand on the controller overall, and so the removal of tasks in the AM condition had

a noticeable effect on reported workload. If controllers did not feel that same demand in the H-L-H scenario, then the AM condition may not have had a notable influence on subjective workload.

4.2 Taskload and Task Performance

Task performance was assessed by the accuracy of metering arriving aircraft. Overall performance was good, with most aircraft arriving within the task criterion (30 seconds of the metered time). As expected, accuracy seems to co-vary with taskload, with higher delay seen in High workload times, in the H-L-H condition. This relationship is less obvious in the L-H-L condition however, with accuracy unexpectedly decreasing in the last low taskload period, possibly due to fatigue or time on task effects. Another interesting finding is that, in general, the AM and manual conditions do not appear to be too different in terms of metering, although there appears to be more variation between the conditions in the L-H-L scenario. This may suggest that the influence of the Arrival manager condition on workload that was found in the L-H-L scenario did not extend to improve performance. Finally, the standard deviations of delay in phase 3 are larger than the equivalent phase 1 period, for both conditions and scenarios. Performance variability therefore appears to have increased across task demand period. Increases in performance variability over time have been documented previously, although for vigilance-based performance [10]. The increase in performance variability may suggest that controllers have to work harder to maintain efficiency performance, with this becoming harder to maintain.

4.3 Workload, Automation Level and Performance

Analysis of the correlations between workload and arrival aircraft metering provides further detail about the relationship between workload and performance under different automation levels and taskload variation scenarios. In the H-L-H condition, a significant correlation was found between workload and metering; as workload ratings rise and fall metering delay also rises and falls. The correlation between workload and delay in the Manual task set was lower, although still significant. There is a noticeable reduction in this correlation during phase 3. The lower correlation is not unexpected, as in the Manual condition, participants had to work on conflict detection and resolution tasks in addition to the metering task. Controllers are not passive in their environment. With a higher experienced workload, controllers may have applied strategies to ensure maintenance of performance even under a high workload [7]. This is not seen in the Arrival manager conditions, however. The added automation may have resulted in less strategic options for maintaining performance. The differential application of strategy and how the controller elected to control and manage the traffic could contribute to the reduce covariance.

An unexpected finding was the low correlation between performance and workload for both manual and AM conditions in the L-H-L scenario. In fact, there appears to be hardly any relationship. The small covariance that is observed is often negative, with delay increasing with low workload and decreasing in association with

high reported workload. There is therefore a clear effect of workload transition direction on the association between workload and performance. In the H-L-H scenario, the relationship between workload and performance is more predictable, although less so in the manual condition potentially due to application of individual control strategies, or just more choice in the control approach. In the L-H-L scenario, the transitions appear to influence the workload-performance relationship.

Although there is a lack of common agreement regarding the mechanisms by which task demand transitions may impact covariate factors [20], this collection of workload findings may be interpreted in the context of Limited Resource theory [5] and arousal theories. Potentially, in H-L-H scenario, the low demand period may have enabled controllers to use this time to recover resources and prepare for the next high task demand period. [6] has previously documented that this is an active control strategy that controllers use during low demand periods, when it is considered safe to do so. Arousal theories may provide some insight into why this effect may not be seen in the L-H-L demand transition pattern. Arousal theories suggest that low workload (or underload) may lead to lower arousal, which may limit attentional resources and create boredom and lack of motivation. If a human operator started a task from this point, it may be that the following demand periods are perceived to be more demanding. By the final low demand period, the operator may find it difficult to pay attention. Attentional resources theories suggest however that if preceded by a higher demand, lower demand periods can be utilized to replenish attentional resources, not necessarily reducing arousal to a level that would create negative effects. The application of these theories therefore potentially account for the disparate findings between the different task demand transition patterns. If this effect occurred in the L-H-L scenario but not the H-L-H scenario, this also may explain why the AM condition had noticeably lower workload rating in the L-H-L scenario but not the H-L-H scenario.

The finding of improved metering may also be the result of controllers applying strategies to support performance across the demand periods [7]. Although controller strategies were not a direct focus of this research, this finding highlights an important issue for future research considerations. Although this measure of performance (arrival metering) indicates that performance was maintained in the L-H-L scenario, controllers also reported greater perception of workload. It is therefore possible that controllers may have experienced having to work harder to maintain performance, even though this was not observable in the performance measure itself. This result emphasizes that in order to detect, and prevent, performance declines, further research should focus on measures that are sensitive to the operators' experience, and that can be monitored and utilized to detect potential performance decline prior to a performance related incident.

It is acknowledged that these results are provisional, and need to be interpreted within context. For example, in an air traffic environment, it is easier for the controller to build a picture of the traffic by ramping up with the traffic rather than just starting a session in a high demand period [6]. However, findings do have important implications for the prediction of controller performance in an operational environment. Findings suggest that high and low demand periods can affect controller perception of covariate factors such as workload differentially depending on what has happened prior to the current situation. Thus, supervisors may need to pay close

attention to the number and direction of transitions that a controller experiences per session to most effectively support controller performance.

Future research should further explore the relationship between previous task demands and the relationship on present controller experience, including the exploration of sudden, and unexpected, transitions. Better predictions are needed to identify and prevent potential performance declines and associated performance-related incidents. Such predictions may be particularly relevant for adaptive automation technologies that support operator performance.

5 Conclusion

The affect of task demand transitions on workload and one, efficiency related performance measure was investigated within the context of an air traffic control task. Initial findings suggest that task demand variations affected participants' perceptions of workload, although the effect appeared to be influenced by the direction of the previous demand periods. This was also influenced by the level of automation available to the controller, with the controller experiencing less workload when controlling with automation in an Arrival manager condition in the L-H-L scenario.

Performance appeared to vary to some extent with taskload, in the direction expected, although findings were again disparate between scenarios. The most interesting findings suggest that the relationship between workload and performance was affected both by level of automation available to the controller, and direction of taskload. This finding has potential implications for the assessment of new automation and applying increased levels of automation in the control room. Previous research has infrequently considered transitions of task demand in an applied environment. Findings are consistent with the description of workload history effects [8], and that equivalent task demand periods can elicit different experiences for a human operator depending on what precedes the time of rating. Attentional resource and arousal theories appear to support interpretation of the results. Further research is required to enhance understanding of demand transition and history effects. Practical applications include guidance for operations room supervisors, and implications for predictions of performance in high and low demand periods, with important implications for identifying and preventing potential performance declines and associated performance-related incidents.

References

1. Di Nocera, F., Fabrozo, R., Terenzi, M., & Ferlazzo, F.: Procedural Errors in Air Traffic Control: Effects of Traffic Density, Expertise, and Automation. *Aviat. Space Envir. Md.*, 77, 639--643 (2006)
2. Mogford, R., Guttman, J., Morrow, S. & Kopardekar, P. The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature, DOT/FAA/CT-TN95/22, Washington, D.C., FAA (1995)

3. Loft, S., Sanderson, P., Neal, A. & Mooij, M.: Predicting Mental Workload in En Route Air Traffic Control: Critical review and broader implications, *Human Factors*, 49 (3) 376--399 (2007)
4. Djokic, J., Lorenz, B., & Fricke, H.: Air Traffic Control Complexity as Workload Driver. *Transport Res. C-Emer.*, , 930--936 (2010)
5. Wickens, C.D.: *Engineering Psychology and Human Performance*. Harper Collins, New York (1992)
6. Edwards, T., Sharples, S., Kirwan, B., Wilson, J.R.: Identifying Markers of Performance Decline in Air Traffic Controllers. In: Di Bucchianico, G., Vallicelli, A., Stanton, N.A., Landry, S.J. (eds.) *Human Factors in Transportation: Social and Technological Evolution Across Maritime, Road, Rail, and Aviation Domains* (In Press)
7. Sperandio, J.C.: Variation of Operator's Strategies and Regulating Effects on Workload. *Ergonomics*, 14, 571--577 (1971)
8. Cox-Fuenzalida, L.E.: Effect of Workload History on Task Performance. *Hum. Factors*, 49, 277--291 (2007)
9. Hancock, P.A., Williams, G., Manning, C.M.: Influence of Task Demand Characteristics on Workload and Performance. *Int. J. Aviat. Psychol.*, 5, 63--86 (1995)
10. Helton, W.S., Shaw, T., Warm, J. S., Matthews, G., Hancock, P.: Effects Of Warned and Unwarned Demand Transitions on Vigilance Performance and Stress. *Anxiety Stress Copin.*, 21, 173--184 (2008)
11. Moroney, B.W., Warm, J.S., Dember, W.N.: Effects of Demand Transitions on Vigilance Performance and Perceived Workload. In *Proceedings of the Human Factors And Ergonomics Society Annual Meeting 39*, 1375--1379. SAGE Publications (1995, October)
12. Edwards T., Gabets C., Mercer J., Bienert N. (2017) Task Demand Variation in Air Traffic Control: Implications for Workload, Fatigue, and Performance. In: Stanton N., Landry S., Di Bucchianico G., Vallicelli A. (Eds) *Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing*, vol 484. Springer, Cham
13. Homola, J., Morey, S., Cabrall, C., Martin, L., Mercer, J., Prevot, T.: Analysis of interactive conflict resolution tool usage in a mixed-equipage environment. *AIAA Guidance, Navigation and Communications Conference*. Boston, MA (2013)
14. Zeghal, K., Hoffman, E.: Delegation of separation assurance to aircraft: Towards a framework for analysing the different concepts and underlying principles. *ICAS 23rd Congress*. Harrogate, UK (2000)
15. Galster, S.M., Duley, J.A., Masalonis, A.J., Parasuraman, R.: Air Traffic Controller Performance and Workload Under Mature Free Flight: Conflict Detection and Resolution of Aircraft Self-Separation. *Int. J. Aviat. Psychol.*, 11, 71--93 (2001)
16. Tenney, Y.J., Spector, S.L.: Comparisons of HBR models with Human-in-the-Loop Performance in a Simplified Air Traffic Control Simulation with and without HLA Protocols: Task Simulation, Human Data and Results. In: *Proceedings of the 10th Conference on Computer Generated Forces & Behaviour Representation*, Norfolk, VA, (2001)
17. Erzberger, H.: The Automated Airspace Concept. 4th USA/ Europe Air Traffic Management R&D Seminar, Santa Fe, NM, December 3-7 (2001)
18. Farley, T., Erzberger, H.: Fast-time simulation evaluation of a conflict resolution algorithm under high air traffic demand. 7th USA/ Europe Air Traffic Management R&D Seminar, Barcelona, Spain, July 2-5 (2007)
19. Paielli, R., Erzberger, H.: Tactical conflict alerting aid for air traffic controllers. *AIAA Journal of Guidance Control and Dynamics*, 32, vol,1, pp 184-93 (2009)
20. Wickens, C.D., Mabor, A.S., & McGee, J.P.: *Flight to the Future: Human Factors in Air Traffic Control*. National Academy Press Washington D.C. (1997)