Human Factors and Aerospace Safety 3(1), 1-23 **FINAL DRAFT**

POSITION PAPER

Control workload, airspace capacity and future systems

Peter Brooker
Cranfield University, UK

Everything is what it is, and not another thing – Joseph Butler

Introduction

Orville never complained to Wilbur about the workload at Kitty Hawk. ‘Workload’ is a fairly recently invented word. It seems to have originated in the 1940s as ‘work load’, allied to teaching load and case load. After a few years with a hyphen, it took its present form. In the dictionary, work means ‘physical or mental effort directed towards achieving something’, and workload now means ‘the amount of work assigned to an individual for completion within a certain time’.

In air traffic control (ATC), controller workload – or controller mental workload – is an extremely important topic. There have been many research studies, reports and reviews on workload (as it will be referred to here). Indeed, the joke is that researchers will produce ‘reviews of reviews’ (Stein, 1998). The present document necessarily has something of that flavour, and does review many of the ‘breakthrough’ research results, but there is a concentration on some specific questions about workload.

The aim here is to explore how the understanding of workload feeds into the measurement of ATC airspace sector capacity, and how predictive workload and sector capacity techniques need to be available to estimate the traffic handling capacity of future ATC systems. These future systems could be ones in which controllers use computer assistance tools, or where some control tasks are delegated to pilots, or where some control tasks are automated, or where all control tasks are transferred to pilots… The central point is that control workload is a core determinant of future ATC system capacity, whatever the system might

p.brooker@cranfield.ac.uk

be, *and* that the benefits (and costs) of new control arrangements will need to be well understood before investments will take place.

A subtext in this exploration is an examination of the role that applied psychology should play to be most useful. What kinds of problems should applied psychologists attack? In the present context, much of the work on sector capacity has in fact been carried out by systems engineers and operational researchers. This paper is written by an operational researcher who recognises the great value that applied psychology can offer – but not all researchers, operational staff or aviation managers share such a positive view. Applied psychologists have often found their role to be that of marginal improvers of operational systems (or worse, carrying remedial work on defective systems) rather than being a trusted part of the research or design teams producing new ATC systems.

An aviation psychology pioneer working on workload precursors serves to illustrate the roles that applied psychologists could play. Three pieces of work by Kenneth Craik serve to substantiate an assertion that he was an originator of workload research, with results that are still relevant today. Craik developed the modern formulation of the concept of a mental model (Craik, 1943). He argued that human beings translate external events into internal models and then reason by manipulating these symbolic representations; so a mental model is in essence a dynamic representation or simulation of the world. Craik's work on flight simulators is recognised by the UK's Royal Aeronautical Society (Rolfe, 2002) – his objective was the measurement of the onset of fatigue.

Craik also put forward the important 'single-channel hypothesis'. This holds that an individual cannot normally carry out two distinct tasks (i.e. not ones with close mappings between stimulus and response) completely independently when each of them requires a choice of response (Craik, 1947 and 1948). When this is attempted, substantial delays occur in one or both tasks, even for 'trivial' tasks. Welford (1967) was the first to assert that the brain is subject to a single-channel bottleneck arising in the selection of responses. More recent work (Ruthruff et al, 2001) shows that this bottleneck is structural, i.e. it is a basic limitation of the human cognitive/neural architecture.

Operational design and project staff – customers in modern jargon – valued Craik's work and guidance, because he tackled big practical problems and supplied normative answers. It was successful *applied* psychology, 'useful' in the best sense of the word, that was underpinned by a sound philosophical approach.

The following section lists some useful papers in the published literature. The sections after that are labelled Control Workload, Airspace Capacity, and Future Systems respectively, but these titles just indicate the main topics in each section – the text is essentially continuous. The final section is 'Good Predictive Models', and attempts to set out some lessons about the kinds of problems that applied psychologists should be addressing.

The literature

The research and technical literature on workload and airspace capacity is enormous. Perhaps the earliest published work is that by Arad and his colleagues (Arad, 1964); and possibly the earliest critical review article is by Ratcliffe (1969) – who managed to anticipate many of the principal research themes. The concentration here is on techniques used in the UK by National Air traffic Services (NATS), with which the author has been involved, but some of the general references used also need to be acknowledged.

There are several good recent reviews of workload assessment. Stein's 1998 paper has already been noted. Chapter 11 of the book by Wickens and Hollands (2000), titled 'Attention, Time-Sharing, and Workload', analyses the fundamental issues. Kirwan et al (1998) is written very much from the point of view of the practical applied psychologist.

There are rather fewer references on the topic of sector capacity in the open literature: government research establishments and ATC service providers tend to produce internal reports, so researchers in different countries have often been unaware of existing work. More recently, industrial relations, commercial confidentiality, and the use of contractors are increasingly important. One useful survey of sector capacity techniques is that from the INTEGRA project (Eurocontrol, 2000).

A related relevant report is Eurocontrol (1997), a report on the development of a Cognitive Model for ATC (contracted out to the "Institute of Evaluation Research"). It is intended to provide a basic understanding of the cognitive components and processes in ATC – modelled as an information processing activity governed by rules, plans and the controller's acquired knowledge. Hendy et al (1997) is not primarily a review paper, but presents 'top-down' examination and critique of information processing and time pressure in relation to workload.

A further Eurocontrol report (2002) examines Human Performance Metrics in ATC. It provides an extensive bibliography and compares the characteristics of different approaches to workload, and also to:

- Situation awareness
- System monitoring and error detection
- Teamwork
- Trust
- Usability and user acceptance
- Human error

The need to consider these other metrics serves as a caution that workload is not the only human performance factor that must be investigated when developing future systems.

Control workload

One way of dividing the workload literature into categories is to ask if the author(s) propose a definition of workload. Definitions are sometimes suggestive of calculation methods. One appealing definition is from Stein (1998):

"…the amount of effort, both physical and psychological, expended in response to system demands (taskload) and also in accordance with the operator's internal standard of performance."

This has the advantage of introducing the concepts of taskload (= system demands) and internal (*sic*) standard of performance. Taskload and workload are analogous to physical stress and strain, the latter being the consequence of the former. A second definition is from Kirwan et al (1998), quoting one of the authors (Megaw):

"…the more difficult the task is, then the more complex the mental operations are, the more mental processing power and capacity is used, and the more human physiological variables (e.g. heart rate) are affected, and the more the subject 'feels' a higher degree of workload."

This introduces elements such as mental processing, physiological variables and the subject's feelings (although why did Megaw use quotes?). Thus, workload is a multi-dimensional concept encompassing both the difficulty of tasks and the effort – physical and mental – that has to be brought to bear, plus a personal dimension.

The two definitions overlap to a degree but are noticeably different – and seasoned professionals produced them both. Is an agreed definition of workload ever possible? Workload is a construct or concept. Its definition is surely largely a narrative instructing the reader about its relationships to a neighbouring family of concepts. An analogy is a philosopher saying that it is only possible to understand a concept such as 'responsibility' by considering the causes underlying particular actions and then their consequences, including the blame for particular decisions.

Figure 1 shows some of the activities and concepts associated with workload: the figure maps onto to both of the definitions above. Given this complexity, how could there be an objective 'scientific' definition of workload? To be scientific, one would have to be able to prove that it had been achieved. Where is the 'gold standard' – an absolute measuring scale – against which it could be compared?

Workload cannot be an 'objectively scientific' concept because it includes subjective elements – 'internal performance standards' and 'feels' are subjective words used in the two definitions, and indeed in most people's understanding about what workload represents. Thus, workload cannot entirely be represented by brain or physiological functions: it necessarily involves consciousness and mental states that are perceived and assessed subjectively.

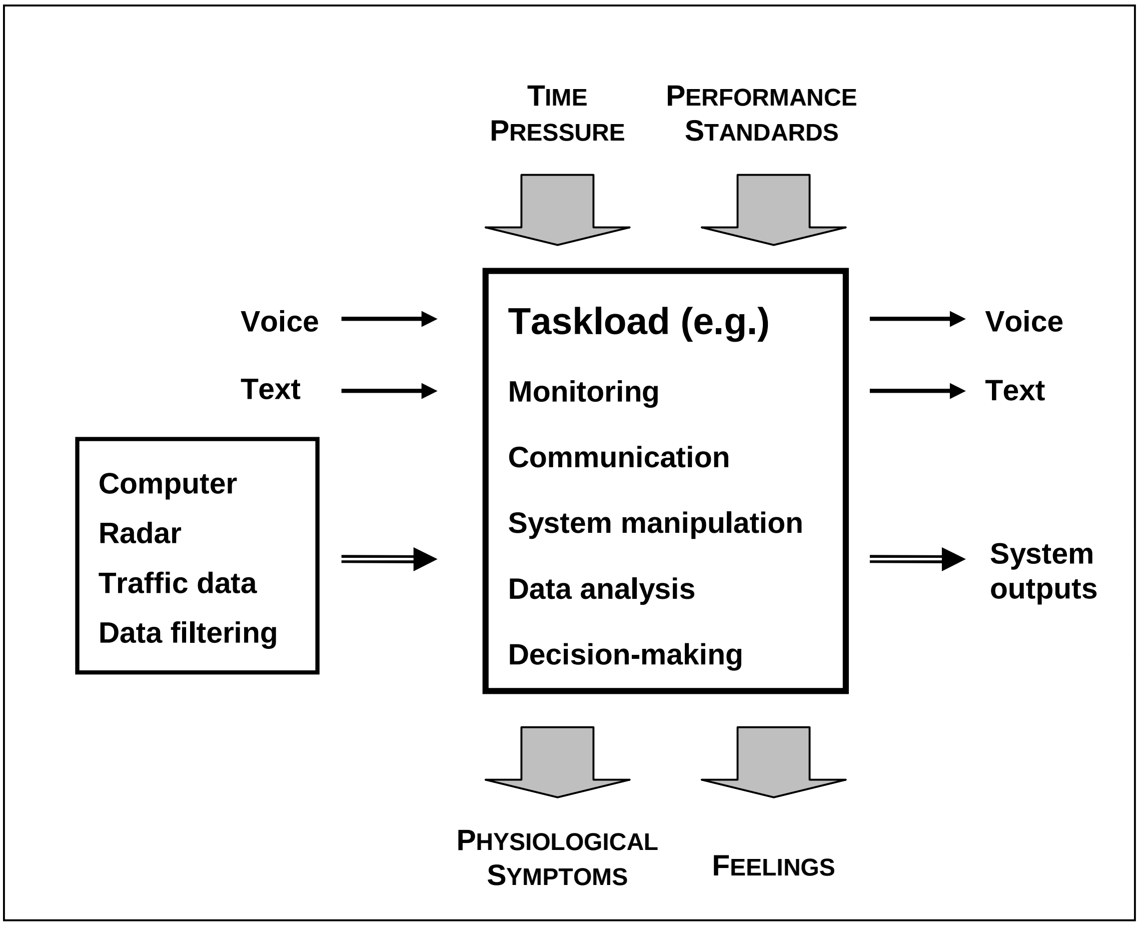

Figure 1 Some aspects of workload

Although the word psychology is derived from the Greek for ‘mind’, applied psychologists tend to avoid analysis of the mind or consciousness; e.g. Wickens and Hollands (2000) book does not include the words ‘mind’, ‘feelings’, ‘subjective’ or ‘consciousness’ in its index. In psychology, definitions of ‘cognition’ note human thought processes and their components – perception, memory, decision-making, etc. An individual tends to be modelled in information processing terms, with physiological components, such as stress, changing the individual’s performance. Psychology is viewed as a science, so the inclination is to use mathematically based techniques, e.g. information theory, and experimentally orientated methods, e.g. physiological indicators, rather than delve into the realms of philosophy. Unfortunately, philosophical methods do not test hypotheses of approximations to the truth – and hence are seldom of practical use to a team trying to design a better ATC system.

The fact that scientific methods operate by trying to eliminate personal subjective prejudices is often somehow equated with the idea that all subjective elements are ‘bad’ and can/should be eliminated. Nagel (1999) and Searle (1999) provide some arguments to demonstrate that this is a confusion. For present purposes, the assertion is that statements about workload are, in the last resort, subjective ones, albeit overlaying objective data on taskload, time pressure, etc. If there is to be a ‘gold standard’ for workload comparisons, then it has to be the product of subjective assessments. Can such a standard be constructed with ‘reasonable accuracy’? To try to answer this question, it is necessary to examine the training, management and culture of controllers.

Controllers are selected to have good intelligence and stable personalities – depressive and introspective tendencies would obviously not be helpful characteristics – and extensively trained over several years. There are usually two phases: classroom and simulator training over about 18 months to earn a ‘rating’, and then around the same amount of time of OJT (On-the-Job-Training) to earn ‘validations’ to operate on one or two real airspace sectors. Classroom training takes place in only a few colleges in any one country, e.g. the UK has one college for ‘airspace’ ATC. College instructors need to have been operational controllers; in many cases, they would have trained originally at the same college. The time involved for OJT varies between individuals; they haveachieved a specified performance. Much of OJT involves sitting with experienced controllers and being mentored about the right techniques to use to move traffic safely and expeditiously in the sector. Controllers work on shifts, and people on the same shift pattern tend to work together for long periods of their career.

Selection and training therefore exhibits considerable consistency of instruction, correction and reinforcement, and this is further conditioned through ‘controller culture’. Controllers are trained in standardised ways to make correct analytical judgements, and to recognise and tackle typical problems, e.g. climbing an aircraft through another’s flight level, in a particular way; i.e. sets of heuristics will tend to be adopted by trainees. A standard language – a very restricted subset of English – is used for communication between controllers and pilots. Data entry to computer or paper record uses *pro formas* rather than free formats. Trainee validation on specific sectors – with largely stable routeings and traffic patterns – means that the same common kinds of ATC problems have to be encountered and resolved many times. Controller operational performance is monitored throughout an individual’s career. The controller community discusses thoughtfully what is ‘best practice’ through its professional journals (e.g. Transmit, published by the UK Guild of Air Traffic Control Officers), and studies and debates incident and Airprox reports. There is thus a considerable commonality of experience.

It is therefore argued that experienced controllers are able to assess ‘externally’ the workload that an operational sector controller is experiencing – ‘over-the-shoulder’ rating. Workload experience may be mental ‘private property’ to the controller, but such observers understand the thought processes and pressures for that person. They can comprehend the issues raised by the data flows, the information on the radar screen and in communication messages. They know what the experience being undergone is like for the sector controller. These are reasonable inductive arguments, but are in no sense a formal 'proof' that:

Average external controller assessments ⇔ workload.

But how could a proof be constructed for any other technique that one might propose for workload? Subjective assessments bring into play all the elements of figure 1, so they have 'face validity' (an irritating phrase to non-psychologists, as it gives a cryptic hint about some much better technique known only to the cognoscenti). If a new technique did not match controller assessments then why would one believe that would be an adequate workload measure? On what rational basis could a novel technique be shown to be 'better'? How, indeed, could external subjective assessments be falsified – the 'acid test' of a scientific hypothesis (Popper, 1959)?

An example of external controller assessment is the DORA method, developed in the UK in the early 1970s (Smith and Stamp, 1973). The expert observer assesses the workload by noting mentally such things as the activities of the controller on the R/T and marking strips, the verbal liaison with colleagues, the nature of the radar picture, etc. An important element is that of ensuring consistency in ratings between the different controller assessors. (NB: with modern data processing and video facilities, the controller's performance can be replayed and discussed by assessors.) Every two minutes the assessor rates the controller workload, these ratings later being matched against the levels of traffic on the controller's frequency at the same time. The ranking scale is given in table 1 (NB: various labels were used).

Table 1 Controller workload ranking scale.

Workload Category	Interpretation
I	Fully loaded – 'could not handle another aircraft'
II	Very busy, but with some spare capacity
III	Busy, but without special difficulties
IV	Workload below III

When the two-minute workload ratings are matched against traffic flow the resulting picture of workload versus hourly traffic produces figure 2 – the actual workload over the period would need to weight the responses here by the number of aircraft under control. Note that this figure is constructed for particular equipment and particular procedures for this particular sector. As expected, the proportions of time spent in the heavier workload categories increase with the traffic flow, but there is not a sharp cut-off at any sort of critical hourly throughput

– so there is no ‘inherent’ sector capacity figure where ‘one more aircraft is too many’. What this means in terms of sector capacity is discussed in the next section.

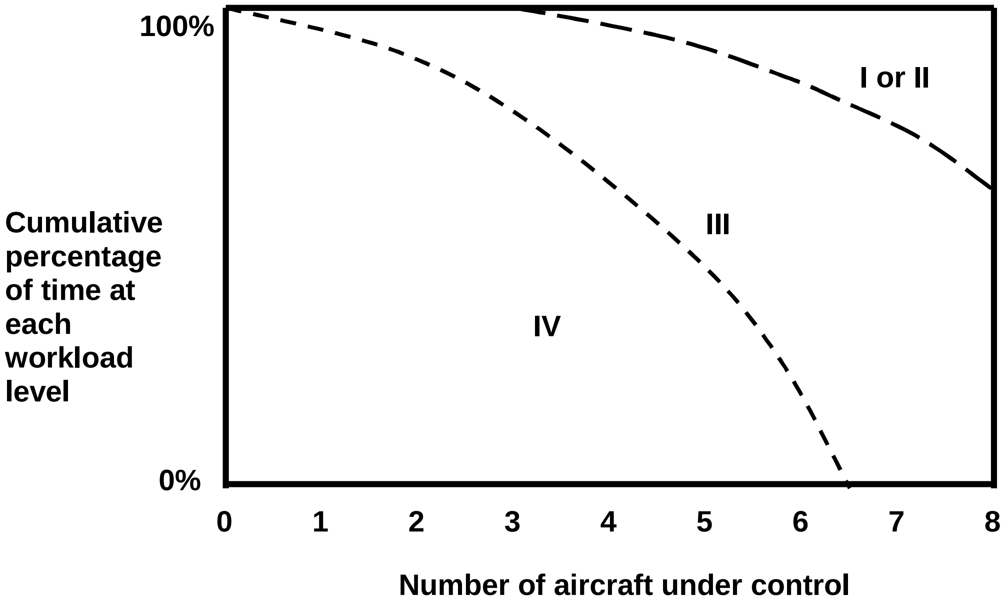

Figure 2 Illustrative relationship between workload and number of aircraft controlled (adapted from figures and measurements in Smith and Stamp, 1973)

An important factor, which can affect workload dramatically, is the pattern of flow through the hour. The highest workload levels (I and II) most generally arise from short-term peaks in traffic, when taskload is naturally highest. If the flow of aircraft is regulated in some way, e.g. by controlling airport departure times, thus preventing strong traffic troughs and peaks, then the frequency of these high taskload periods can be markedly reduced. Very tight and effective flow regulation produces a shift upward in the curves of figure 2.

Other methods for assessing workload

Dozens, perhaps hundreds, of ways have been put forward for measuring workload. Many of these calculation methods are properly called ‘metrics’ – proxies for workload – as a person’s performance of numerical tests is a metric for intelligence but not a complete picture of his or her abilities. Taskload-based and predictive metrics are examined in the next section. The rest of the workload measurement techniques can be categorised as physiological, operator subjective, or performance: Stein (1998), Kirwan et al. (1998) and Eurocontrol (2002) give a very full account – the superficial comments here are just intended to illustrate the possible kinds of metric.

Physiological metrics

Physiological measures assume that changes in workload cause measurable differences in certain physiological processes, generally involuntarily. Some physiological indicators are available even in the absence of overt behaviour. They include Galvanic skin response, heart rate and similar from the electrocardiograph readings, and ocular (eye) responses.

Operator-subjective metrics

Operator-subjective workload assessment methods use the operator’s self-reported effort in carrying out some task(s). They are cheap and easy to use and analyse, and acceptable to operational staff. However, there can be significant individual biases – why should individuals’ self-perceptions be consistent? As discussed further in the next section, there are very wide variations in people’s judgement about their workload, memory limitations, and their mental models for different tasks. However, these metrics often have an important ‘marketing’ role in helping controllers faced with new equipment and/or procedures to ‘accept’ their operational introduction. The Instantaneous Self Assessment (ISA) technique, developed by the UK NATS, is a ‘1 to 5 rating’ method that is now used by the Eurocontrol Agency: it allows for online registration of controller ratings at intervals down to two minutes or so. Other examples (see Eurocontrol, 2002 for references) include the Subjective Workload Assessment Technique (SWAT), the NASA Task Load Index (TLX), and the Air Traffic Workload Input Technique (ATWIT). The key question of course is the extent to which these metrics are benchmarked – and then calibrated – against ‘external subjective’ workload.

Performance metrics

Performance metrics estimate workload through direct measurement of task performance. There are essentially two types of methods: *primary (or direct) task* and *secondary task* measures, but both rely on measuring the influence of increasing task load on the performance of a particular task(s). The primary task technique typically involves varying some primary task parameter (e.g., tracking complexity) that will affect task demands to the point that performance falls below some criterion, thereby providing a measure of residual capacity at resource allocations below criterion performance. Thus, primary task measures can directly relate workload to system performance. Performance metrics have shortcomings, e.g. performance is dependent on strategies (because priorities will shift between tasks) and can often be intrusive and artificial. Secondary task performance adds in a secondary task and then the analyst examines how its performance deteriorates with increased traffic through the sector, e.g. simple digit reading, repetition of words presented to controller. One variant is the inclusion of realistic ‘embedded tasks’ – tasks that provide proxy measures of workload, while appearing ‘normal’ to the controller.

Sector capacity

Sector capacity is more than the application of workload measurement to airspace sectors. A Sector’s capacity can be defined as that sustainable flow of traffic generating maximum acceptable (sic) controller workload. But ATC is a complex ‘socio-technical system’: people and machines are linked together through structures and processes. Controllers are employed by an organisation that has to operate within an evolving industrial relations framework. The introduction of system changes has to be negotiated rather than imposed. One of the common words that controllers use about themselves is ‘professionalism’: this is often seen as being at odds with anything reminiscent of F. W. Taylor’s ‘scientific management’.

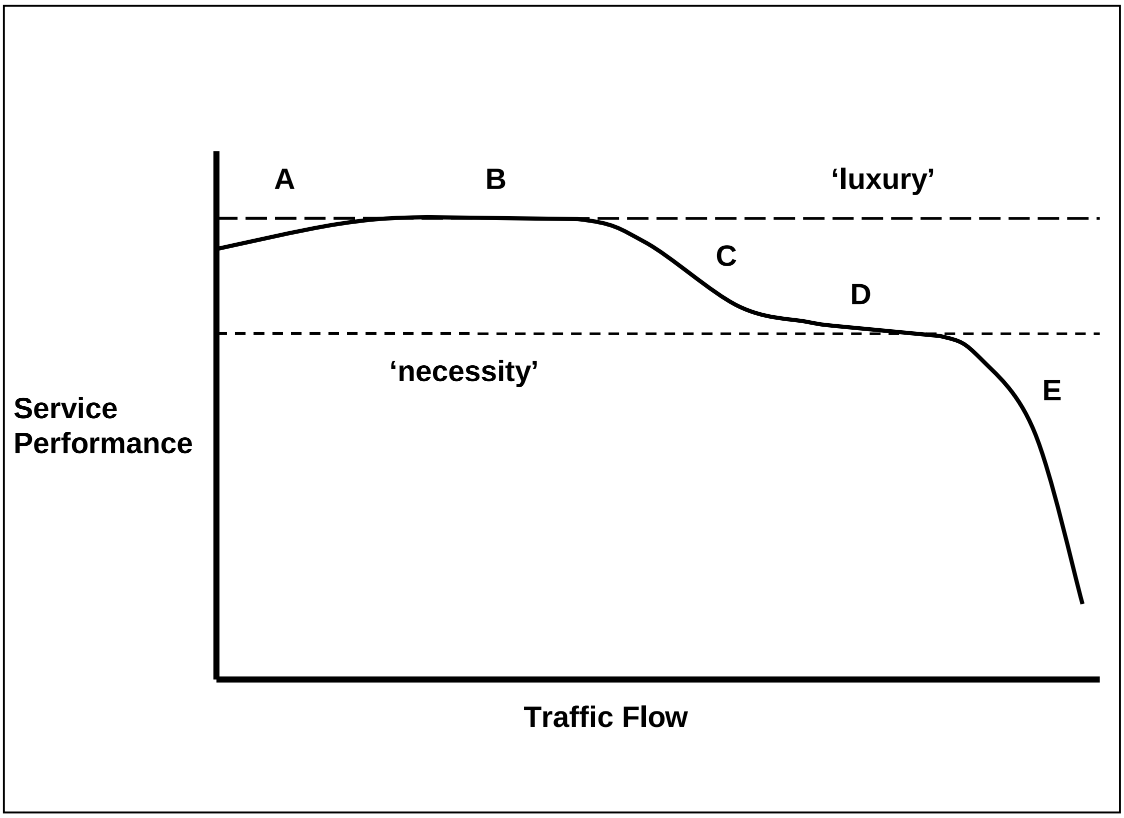

Figure 3 Postulated average service performance by controller on sector for different traffic flows

ATC delivers a *service performance* product to its customers – safe and expeditious flight through airspace. By expeditious is meant the 'economic quality' of the flightpath, with a perfect quality flightpath presumably being one adopted if there were no other aircraft anywhere in the vicinity. How does performance vary with workload?

Consider a single controller handling a specific sector at different levels of traffic flow (measured as flights per hour) – the words 'single' and 'specific' are important. The traffic routeings and patterns through the sector are assumed constant, i.e. increased traffic flow produces 'more of the same'. Figure 3 illustrates what average (sic) service performance by that controller on that sector for the different levels of traffic flow. Performance here is an (undefined) appropriate combination of safety and expedition: a hazardous incident occurring at an unacceptable frequency over time would obviously count as an extremely low performance component and an economical flightpath as a high performance component.

The general shape of the curve in figure 3 can be supported by results in the research literature. In particular, Sperandio (1978), noted that controllers handle increasing traffic by adopting successively more economical strategies in operating methods to defer the onset of 'overload' conditions. At very low flow rates (A in diagram) there would be a concern that the controller 'underload' would be reflected in boredom and 'coping behaviours', which might result in increased rates of hazardous error (Hopkin, 1988, quoted in Stein, 1998). When there are a few aircraft on the controller's frequency he/she can provide a 'luxury' service to the aircraft, and provide the best flight path for each of them (e.g. provide expeditious routeing and better climb/descent profiles (B)). For higher flows, more stereotyped flight paths have to become the norm, although speeds and tracks can still be tailored on some occasions (C). When the controller has to handle many aircraft, the only feasible control method is to concentrate on keeping the flow of aircraft moving through the sector safely (D): individual operating characteristics are now low priority and ATC instructions have to concentrate on 'necessary' elements. For very high flow rates, operational safety/economy errors would be frequent, and hence general performance would be very poor (E).

The most important point, echoing the previous section's comments about workload, is that there is no evidence to suggest that, except at very high flow rates, there are anything but steady changes in the curve in figure 3, i.e. no reason to suppose that any dramatic changes occur at some 'crucial' traffic flow(s).

The natural variations between individuals also have to be addressed by ATC planners and managers. There are well-known differences between the performances of individual controllers on identical tasks. An individual's performance with the same traffic can also vary considerably: on some days, the controller might just be a better performer; there might be different kinds of background distractions in the control room; slight changes in communications might affect how the traffic pattern develops. Tattersall (1998) explores these and other kinds of differences: relevant factors include age, experience, gender, personality, cognitive style, and time-sharing ability. Likewise, Stein (1998) identified variations in the mental models used by controllers.

The statistical distribution of controller service performance on a particular sector for a particular flow rate probably looks something like figure 4 – there would be a family of curves for different flow rates. Some individuals at some times perform very well, most produce good service performance, and, on a few occasions, the performance by some controllers is poor. This is no more than a vaguely Gaussian distribution plus some assumed 'censoring' at the low performance end: chronically poor performers would not have been validated on the sector; others would be removed from their posts if they showed deterioration; those showing short-term degradation of performance would strive to achieve safety; etc.

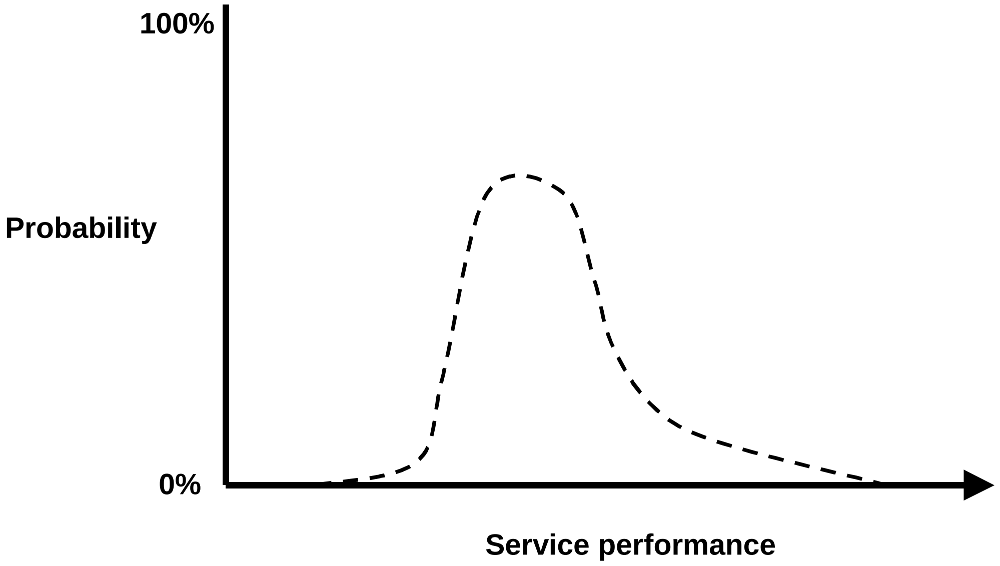

Figure 4 Speculative probability distribution of controller service performance for a particular sector and flow rate

So, given the curves in figures 2, 3 and 4, the word 'acceptable' in the definition of sector capacity, and the comments about variations between controllers, how is sector capacity to be determined? For a given ATC system, there are two crucial points:

- The setting of 'acceptable' workload is an industrial relations issue.
- The most important 'control tool' available to ATC planners is traffic flow.

Thus, in practice:

'Acceptable' workload corresponds to a planned (sic) service performance at about the region C in figure 2. Were it to be set at a point where all the aircraft under control would be receiving the 'luxury' service' (i.e. at B) then the hourly throughput would be probably be too low in terms of airlines' commercial needs.

A workload level at D would be too risky – the odd ‘unplannable’ peak in traffic could put the workload into the unacceptable E region.

This acceptable workload would correspond to an individual level of service performance at about the kink in the lower end of the performance distribution in figure 4. This would represent a ‘standard validated competent controller’, as judged by the managers and negotiated with operational controllers en bloc.

In practice, ATC planners are interested only in those workload levels that are ‘just acceptable’. ATC has to control the levels of workload by controlling the levels of taskload. They need to be able to estimate how taskload changes with traffic flows, which *inter alia* will generally produce changes in spatial and time patterns for conflicts. *They therefore want a taskload measure that consistently provides the best match to ‘just acceptable’ workload.* The desired measure must accurately predict when workload would be at the planned performance level, immaterial of the nature and magnitude of workload components/traffic flows/patterns.

Sector capacity might thus correspond to a set of conditions such as ‘no more than 50% of the controller’s time at ratings I and II’ – ‘a good day’s work’ for a controller with a low risk of fatigue.

The significant point here is that it is taskload which is the ‘instrumental variable’. Taskload measures are traditionally important sources of workload-related data. They can be divided into those that concentrate on some particular aspect of the job, e.g. number of aircraft under control, radio communication bandwidth and duration, and number of flight transitions, and those involving a summation of (most of?) the tasks carried out – i.e. where the processing of information and execution of tasks are of central importance.

A good example of the former is some current work by the FAA (e.g. Manning et al, 2002), which recently examined the addition of communication events – transmissions between pilot/controller and controller/controller – to activity variable taskload factors. The declared aim of the work is ‘to develop objective taskload measures that could replace subjective workload measures’: ATWIT is used as the (active controller’s) subjective workload rating. One notable feature of the work is the depth of statistical analysis employed – mainly principal component analysis and multiple regression. The software used – POWER (Performance and Objective Workload Evaluation Research) includes 20+ variables of aircraft and controller activity. One of the problems about these kinds of performance metrics is the extent to which their predictions can be extrapolated – is the correlation with workload maintained?

Taskload summation metrics have in fact been developed from the very early years – Arad (1964) and Ratcliffe (1969). Significant steps forward were made by Schmidt (1976), who stated the main assumptions very succinctly:

‘Work load (sic)…is related to the frequency of occurrence of events which require decisions to be made and actions to be taken by the control team, and to the time required to accomplish the tasks associated with these events…With proper calibration, the model may be used to assess the impact on work load and sector capacity of future automation features.'

A well-used taskload and simple information processing/action model for workload is DORATASK. This is an analytical model supported by a fast-time simulation model (Phillips (1995) is a general review; Richmond (1989) gives clear examples of the calculation logic). The taskload is calculated by summing the time taken by a controller to carry out all the necessary tasks, both observable and non-observable, associated with the flow of traffic. Sector capacity is then set by total taskload plus a parameter indicating the proportion of time necessary for controller recuperation, i.e. this parameter ensures a match to acceptable workload.

Observable tasks are 'routine' and 'conflict resolution' tasks. A routine task is one that a controller must carry out for all aircraft regardless of whether he/she has any other aircraft under his control, for example the issuing of standard RT instructions. Conflict resolution tasks are additional tasks that must be performed if any aircraft are in potential conflict. Non-observable tasks are planning tasks carried out by a controller, and mental tasks involved in conflict prediction and detection. The time to carry out observable tasks may be measured directly.

Planning work is not directly observable. DORATASK therefore includes algorithms to estimate workload representing the amount of time a controller spends on planning tasks. Richmond (1989) explains the thinking and gives some examples. In the case of terminal areas, there are two non-observable tasks, initial processing and radar monitoring. Radar monitoring is modelled by the number of visual checks of the radar screen, the time per radar check, and the combination of aircraft pairs that have to be checked. The two non-observable tasks are linear and quadratic in the number of aircraft, each being multiplied by an unknown parameter (compare Arad, 1964!). These parameters have to be estimated by benchmarking against sectors of known capacity.

Future systems

There is considerable interest and debate about future ATC systems, e.g. see Brooker (2002) and Eurocontrol (1998). Workload is a vital element in such studies, where the goals are to handle increased traffic safely and cost-effectively. There are many different possibilities and combinations: controllers may get automation tools to help in their tasks, some control tasks may be passed to pilots, control tasks may be automated in some way, airspace sectors can be completely restructured, etc. Possible roles for the controller and pilot are explored (from very different viewpoints) in Kirwan (2001) and Brooker (2003). New systems must be designed to ensure the required capacity. It is not sufficient to rely upon the results of expensive real-time simulations of possible candidate systems.

A simple diagram may help to illustrate the workload issues. The basis is a geometrical construction used in applications from metallurgy to relativity.

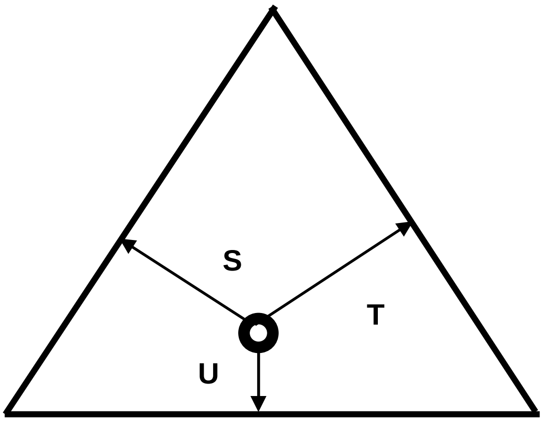

Figure 5 Perpendiculars in an equilateral triangle

Figure 5 shows an equilateral triangle, i.e. with the sides the same length and the angles between them at 60 degrees. From a point O inside the triangle, three lines are drawn so that they are perpendicular to the triangles' sides. It can be proved that the sum of the lengths S, T and U, of these lines is constant, no matter where O is located within the triangle. This enables the possible values of an equation:

S + T + U = Constant,

to be explored within the triangle.

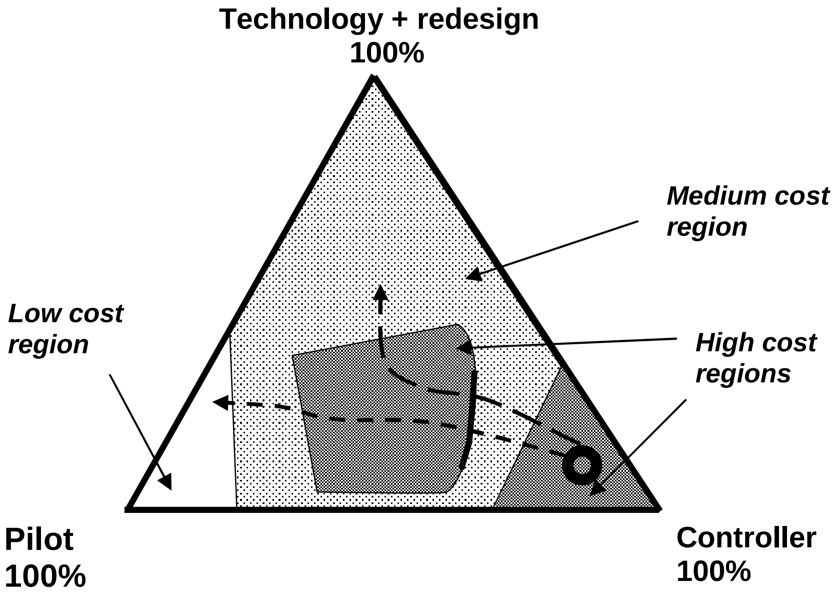

Figure 6 Possible migration paths in the STU triangle with (notional) isocost contours for safe optimised systems (see text)

Figure 6 uses figure 5 for control workload in a particular volume of airspace (which might be one or several sectors). The three variables are:

S = Controller workload
T = Pilot workload
U = Technology/redesign workload equivalent

The second and third of these need some explanation. The second is the control workload component of pilot's total workload. The third is essentially a residual after pilot and controller workload have been counted. It consists of the 'workload equivalent' of those tasks that have been allocated to 'the computer', in the widest sense. However, it also includes the equivalent for those tasks that have been eliminated, e.g. if all control functions were to pass to the pilot then there would obviously be no controller/pilot communication tasks.

The present system is shown by the O label in figure 6 – almost entirely controller workload but some tasks have had some degree of computer assistance, e.g. Short Term Conflict Alert systems supplement controller scanning of the screen for conflicts. Two possible migration paths are shown. The lower one moves from the present system to one in which pilots carry out most of the control functions and there is increased automation. The upper one moves to a system in which there is much more 'automation' and in which pilots and controllers share tasks in some way. The first would be termed full delegation and the latter shared responsibility. Which precise tasks might pass from one actor to another is not the question here: it is assumed that the system represented by each point is the *best* that can be achieved with such an STU combination, i.e. it is safe and with minimum full running cost (i.e. operating cost plus capital investment, equipment maintenance, etc).

This cost dimension could be displayed by creating a three dimensional diagram. For present purposes, it is sufficient to add in some (purely illustrative) 'isocost' contours in the triangle: the darker the shading the higher the cost. The lower path finishes up with a lower cost endpoint than the upper one. But this now shows the difficulties involved with both the migration paths. The first part of the migrations look appealing because there are improvements in the cost function for comparatively small changes in operational concept. Then both paths move into an extensive higher cost region, in which there are no apparent benefits from marginal changes. The thickened boundary for the high cost region shows what is effectively a barrier for the next phase of changes – what decision-maker would want to incur higher costs? The answer is a decision-maker who could be convinced by workload research results, which would predict accurately that the end of the migration path delivers substantial cost effectiveness improvements. This is the challenge for workload research on future systems: considerable faith would be needed to act on the advice of a psychologist to take such a course –

when all you can see is an expanse of desert, you have to have a great deal of trust in a guide who says that beyond the desert is the promised land.

Figure 6 is a considerable simplification. There is a time/traffic dimension – increased total workload would correspond to a larger triangle. Moreover, the nature of the isocost contours will change for higher traffic levels. Some STU combinations, e.g. continuing with existing control concept, could become infeasible, so their costs would be very much higher (NB: they would have to include the economic opportunity cost of the flights that could not operate). There may well be areas within the triangle that are not feasible in safety terms, i.e. such STU combinations would not achieve the necessary safety targets, and so would effectively have infinite cost.

Real-time simulations are only suitable for evaluating systems that are at a very late design stage. Designers need workload measures early in the development cycle, for prediction of new concepts of operation, interfaces and tools, to enable early concentration on the most promising potential systems. Workload models for future systems have to be developed from taskload models plus some new ideas about how task timings will change with different data flows, computer assistance/automation. It is essential that taskload be modelled in such a way that it correlates well with acceptable workload over the whole region of interest in the triangle. Research in this area tends to concentrate on information processing, multiple specific resources and time pressures.

Different points in the STU triangle correspond to changes in the nature of workload tasks. This leads to the need for task analysis – breaking down new tasks into constituent ‘well-understood’ basic-task elements, which may overlap or interfere with each other to produce higher control loading.

The ‘secret’ is to measure the mental effort that is expended on basic-task mental resources. Mental resources have a variety of limits. Different types of resource are available to deal with different types of mental processing (e.g. visual versus auditory). A mental ‘channel’ is defined as a distinct information processing capability in the brain; different channels represent loadings on different processing centres.

Two examples are TLAP (Timelines Analysis and Prediction) and VACP (Visual, Auditory Cognitive Psychomotor) – see Kirwan et al (1998) for descriptions. Wickens Multiple Resource Pool theory – W/INDEX – is similar in broad principle to these and has been used in the ATC area. Wickens’ (see Wickens and Hollands, 2000, for references) Model of Multiple Resources assumes that the workload experienced by a controller performing an action can be split between a number of different ‘channels’. These channels are representative of different functions within the brain, such as talking, thinking, listening, moving. For each action, a weighting is assigned to each channel representing the effort required of that channel by that action. This allows the resulting workload to be predicted.

W/INDEX uses six channels – visual, auditory, spatial cognition, verbal cognition, manual response and voice response. Tasks load the channels and produce interference between them, e.g. it is difficult to listen and speak at the same time, so a task requiring simultaneous verbal response and auditory monitoring is dual processing and hence weights the channel loading. The demand for each resource are summed and weighted, and then summed to produce task workload. ATC concurrent tasks are likely to interfere. It should be noted that Wickens' work was based on pilot workload; there do not appear to be any reported validations of his methods for ATC systems.

PUMA (Day et al, 1993, Kilner et al, 1998, Householder and Owens, 1995, Kirwan, 1998) also uses W/INDEX. The PUMA toolset is used to explore different future ATC concepts and system designs. However, validity is not assured since it is based on W/INDEX. Observational task analysis must capture not only visual and communication tasks but also cognitive tasks – future ATC systems are likely to have very different cognitive tasks, as evidenced by the large changes in the STU diagram. PUMA hinges on the weighting values used in the conflict matrix, which quantify the extent of channel interference. Householder and Owens (1995) note that 'the values of the weightings are obtained from the subjective judgement of human factors experts in the ATC environment' – but how can these people make a subjective judgement on something that they are not directly experiencing? Validation of the weightings through objective experiments involving controllers is therefore a prerequisite. Moreover, validation against (e.g.) DORATASK for some of the different kinds of future systems is essential.

There have been few research studies that attempt to model workload in the STU triangle, i.e. for a wide variety of future concepts. A notable one is by Hudgell and Gingell (2001), arising out of the INTEGRA project. This models the components of the ATC system purely through an examination of information processing load (IPL) – which is stated to be a 'surrogate measure' for capacity in each part of the system. The IPL is calculated for each 'actor' – controllers, pilots, computer tools, communications – in the system.

Hudgell and Gingell identify seven 'causes' of information processing, including 'flight arrival' (into the list of flights relevant to the actor), monitoring and resolution planning. The amount of processing p for actor i on cause α is time dependent. Flight arrival is simple: p is equal to a weighting factor times the number of flights 'arriving in the lists' in each unit time. The weighting factor λ is initially taken as 1 for the actors processing that piece of information (each cause has a different λ).

More complex causes produce much more complex expressions. Resolution planning is the task of planning a resolution for each forecast interaction (i.e. when trajectories breach pre-set separation criteria). An additional weighting factor w is required, to express the difficulty of resolution – the number of constraints present in the interaction. The amount of processing is taken as equal to the product of the λ and w weighting factors times the forecast number of interactions in each unit time.

The IPL for each actor is found by adding the seven IPL contributions. For a current controller – who undertakes all the control tasks – the IPL thus has seven terms, hence seven λ weighting factors plus two additional 'difficulty' weightings for resolution planning and monitoring causes. The authors state:

'The experimental analyst needs to assess the "cut-off" for each actor – the maximum information the actor can process in unit time…For pilots and controllers the cut-off can be assessed by calibration simulations in which the actor is known…to be at capacity…It is not a simple cut-off…the actor can accept an overload of information for a short period…'

The authors stress that the IPL methodology provides a framework (sic) for capacity assessment – with the detailed assumptions about weightings being supplied by the user. But a framework is just part of what is required – a random person with a violin is usually some way short of producing a credible musical performance. Given the kinds of analytical problems noted in the earlier text here, this is obviously an important – and very large – problem: there are potentially many weightings to validate in what is recognised as a model with non-linear components. The extent of the calibration measurements and statistical analyses required to estimate the weighting factors with any confidence is extremely large. The second problem is that IPL appears to take little account of the channel interference aspects explored by Wickens in W/INDEX and used in PUMA. The third problem is that IPL needs to be validated and benchmarked against established 'real world' methods such as DORATASK – asserting that something is a surrogate measure does not make it a reliable or unbiased one.

Hendy et al's (1997) work is relevant to the Hudgell and Gingell approach. (NB: this paper was not picked up in the INTEGRA literature search.) Hendy et al. note that 'while many models of workload exist, few appear to be well founded in theory or to provide a satisfactory basis for a quantitative representation of operator load'. They argue that a workload construct requires three components: a time-based factor (time pressure), a factor due to the intensity of demand for attentional resources (the amount of information to be processed = task difficulty/task complexity) and a catch-all factor attributed to the operator's psychological/physiological state (anxiety, arousal, motivation, fatigue, etc). Hendy et al construct and test a model in an ATC simulation explicitly to investigate the relationship between a time-based factor and an intensity-based factor in such an environment. They model the load on the human information processing system from the ratio of the time necessary to process the required information, to the time allowable for making a decision. This ratio – 'time pressure' – is found to determine both subjective estimates of workload as well as operator performance.

Good predictive models

Customers for applied psychology products are generally project or design staff working to tight timescales and budgets. They will ask: "What will your techniques and models accomplish for me?" Their focus is on usefulness and practical benefit: it is very unlikely that they would value debates on cultural relativism. What kinds of models of ATC (socio-technical) systems should applied psychologists therefore aim to produce?

Social scientists know very well that their models are not 'absolute', as evidenced by the kinds of definitions used, e.g. (Chorley and Haggett, 1968):

> *Model*: 'A simplified structuring of reality which presents supposedly significant relationships in a generalised form…(models) are valuable in obscuring incidental detail and in allowing fundamental aspects of reality to appear.'

In other words, a good model is one that illuminates what are the 'key factors' at any point in the topic's development – which presumably provides help to the customer in seeing the way forward.

The previous text has demonstrated the importance of control workload: it is a crucial element in developing future ATC systems. These must produce – in essence – reduced total workload costs per aircraft. Figure 6 demonstrates that a 'marginal analysis'; of workload effects may not be sufficient – or is a local optimum an acceptable objective? Taskload models based on information processing and the summation of the effort required for individual tasks appear to be the way forward, but they need to incorporate channel interference and time pressure aspects. Predictive taskload models, particularly when they direct attention to areas of the 'STU diagram' in figure 6 that are distant from present operational practices, need validation against the subjective workload benchmark. However, this kind of activity requires major resources.

As evidenced by recent work on Human Performance Metrics (Eurocontrol, 2002), these aspects of future ATC systems fall methodologically into the 'very tough' category. This is just one of the challenges for applied psychology in producing useful models to help their ATC designer and planner customers. Human factors research outputs are usually well thought through and intellectually persuasive, but they are not always immediately applicable, as some kind of 'tool kit' or as an integral part of systems design processes. The debate about 'Human-centred automation' (Billings, 1997) serves as an example. The phrase can be construed in several ways – Sheridan (2000) offers ten alternative meanings, some of which would certainly rather more useful to system designers than others. To an outsider, rather a lot of the applied psychology literature seems to consist of demonstrations that someone else's work or hypotheses are flawed.

The prime objective for applied psychologists is surely to produce fruitful and illuminating models that stand generic validation against the experimental data. This is obviously easier to say than to accomplish. Nevertheless, it is crucial. Systems design customers want guidance beyond the level of 'you can make a small extrapolation but we will need to validate the results through simulations'. The studies reported in the previous section did not solve all the problems posed, but they are bold and potentially fruitful steps forward. However, without further progress, applied psychologists might well see other disciples – operational research and systems engineering – occupy vacant professional territory.

An analogy can be made with the history of chemistry. The atomic theory and knowledge of the relative weights of elements enabled some progress to be made in the early 19th century, but the development of the concept of valency – the number of other atoms with which an atom can combine – was a major step forward by Frankland in 1852. In fact, valency only really made complete sense to chemists in the early 20th century, when electron orbitals in atoms begin to be understood. The limits and approximations of the valency concept then became clear, but from the start valency was an extremely fruitful hypothesis, both for theory and for practical implementation. Control workload does not yet have a theory to match something like valency. Perhaps information processing, time pressure and channel models are edging towards a new phase?

References

Arad, B.A. (1964). The Control Load and Sector Design. *Journal of Air Traffic Control 12 (60),* 12-31.

Billings, C.E. (1997). *Aviation Automation: the search for a Human-Centred Approach.* New Jersey, NJ: Lawrence Erlbaum Associates.

Brooker, P. (2002). Future Air Traffic Management – Passing the Key Tests. *The Aeronautical Journal, 106 (1058),* 211-215.

Brooker, P. (2003). Future Air Traffic Management: Strategy and Control Philosophy. *The Aeronautical Journal* – to appear.

Chorley, R. J. and Haggett, P. (Eds). (1968). *Socio-Economic Models in Geography*. London: Methuen and Co. Ltd.

Craik, K. (1943). *The Nature of Explanation.* Cambridge: Cambridge University Press.

Craik, K. (1947). Theory of the human operator in control systems: The operator as an engineering system. *British Journal of Psychology, Part I 38(2),* 56-61.

Craik, K. (1948). Theory of the human operator in control systems: Man as an element in a control system. *British Journal of Psychology, Part II 38(3),* 142-148.

Day, P.O., Hook, M. K., Warren, C. and Kelly, C. J. (1993). The modelling of air traffic controller workload. *In, Proceedings on Workload Assessment and Aviation Safety*. London: Royal Aeronautical Society.

Eurocontrol (1998). Air Traffic Management Strategy for 2000+ Volume 2. Eurocontrol, Brussels.

Eurocontrol (2000). *INTEGRA Capacity Metrics: Literature Survey*. Eurocontrol, Brussels. http://www.eurocontrol.int/care/integra/documents/capacity_report1.pdf

Eurocontrol (2002). *Eurocontrol CARE/ASAS Activity 2: Human Performance Metrics.* Report Reference: CARE/ASAS/NLR/02--034. Eurocontrol, Brussels. http://www.eurocontrol.int/care/asas/documentation/care-asas-a2-02-034.pdf

Eurocontrol. (1997). *Model of the Cognitive Aspects of Air Traffic Control.* Report reference HUM.ET1.ST01.1000-REP-02, Eurocontrol, Brussels. http://www.eurocontrol.int/humanfactors/docs/HF7-HUM.ET1.ST01.1000-REP-02.pdf

Hendy, K.C., Liao, K., and Milgram, P. (1997). Combining Time and Intensity Effects in Assessing Operator Information-Processing Load. *Human Factors, 39(1),* 30-47.

Householder, P. and Owens, S. (1995). *An initial assessment of the effects of a number of computer assistance tools upon controller workload: summary report.* CS Report 9503. London: Civil Aviation Authority.

Hudgell, A.J. and Gingell, R.M. (2001). *Assessing the Capacity of Novel ATM Systems.* 4th USA/Europe Air Traffic Management R&D Seminar. http://atm2001.eurocontrol.fr/finalpapers/pap171.pdf

Kilner, A., Hook, M., Fearnside, P., and Nicholson P. (1998). Developing a predictive model of controller workload in air traffic management. In, Hanson, E. Lovesey, E.J. and Robertson, S.L. (Eds.). *Contemporary Ergonomics 1998.* London: Taylor and Francis.

Kirwan, B. (2001). The role of the controller in the accelerating industry of air traffic management. *Safety Science, 37(2-3),* 151-185.

Kirwan, B.I., Kilner, A.R. and Megaw, E.D. (1998). Mental workload measurement Techniques: A Review. R & D Report 9874, National Air Traffic Services Ltd, London.

Manning, C.A., Mills, S.H., Fox, C.M., Pfleiderer, E.M., Mogilka, H.J. (2002). *Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload*. DOT/FAA/AM-02/4. Office of Civil Aerospace Medical FAA, Washington, USA.

Nagel, T. (1999). *Other Minds: Critical Essays 1969-1994.* Oxford: Oxford University Press.

Phillips, R.M. (1995). *A Guide to the DORATASK procedure for sector capacity estimation*. CS Report 9506. London: Civil Aviation Authority.

Popper, K.R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Ratcliffe, S. (1969). *Mathematical Models for the Prediction of Air Traffic Controller Workloads.* RRE Memorandum No. 2532. Malvern, UK: Royal Radar Establishment, Ministry of Technology.

Richmond, G.C. (1989). *The DORATASK Methodology of Sector Capacity Assessment: an Interim Description of its Adaptation to Terminal Control (TMA) Sectors*. DORA Report 8916. London: Civil Aviation Authority.

Rolfe, J. (2002). History of Flight Simulation (the Cambridge Cockpit). London: The Royal Aeronautical Society. http://www.raes.org.uk/fl-sim/FSG%20Cambridge%20Cockpit.htm

Ruthruff, E., Pashler, H.E. and Klaassen, A. (2001). Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement. *Psychonomic Bulletin and Review 8(1),* 73-80.

Schmidt, D.K. (1976). On modelling ATC workload and sector capacity. *Journal of Aircraft 13(7),* 531-537.

Searle, J. (1999). *Mind, Language and Society.* London: Weidenfeld and Nicolson.

Sheridan, T.B. (2000). Function allocation: algorithm, alchemy or apostasy? *International Journal of Human-Computer Studies. 52 (2),* 203-216.

Smith, A.D.N. and Stamp, R.G. (1973). *A method for estimating the capacity of air traffic sectors - An interim report.* CAA DORA Research Paper 7301. London: CAA.

Sperandio, J.-C. (1978). The Regulation of Working Methods as a Function of Work-load among Air Traffic Controllers. *Ergonomics, 21(3),* 195-202.

Stein, E.J. (1998). *Human Operator Workload in Air Traffic Control.* In, *Human Factors in Air Traffic Control*. San Diego, CA: Academic Press.

Tattersall, A.J. (1998). Individual Differences in Performance. In, *Human Factors in Air Traffic Control*. San Diego, CA: Academic Press.

Welford, A.T. (1967). Single-channel operation in the brain. *Acta Psychologica, 27,* 5-22.

Wickens, C. D. and Hollands, J. (2000). *Engineering Psychology and Human Performance.* New York, NY: Addison Wesley.

Acknowledgements

I would like to acknowledge support from the Civil Aviation Authority's Safety Regulation Group (SRG). Thanks are due to Dr Don Harris for discussions and his comments on an earlier draft.