

PROGRAMME FOR
HARMONISED AIR TRAFFIC
MANAGEMENT RESEARCH
IN EUROCONTROL



EUROPEAN COMMISSION FOR THE SAFETY OF AIR NAVIGATION EUROCONTROL



DOC 99-70-01

VOLUME 2 OF 4

PART 2 OF 2

ANNEXES

CENA PD/3 FINAL REPORT

Annex A: Experimental Design and Methods



EUROCONTROL

EUROCONTROL
96 rue de la Fusée
B-1130 BRUXELLES

Prepared by: C.Chabrol, B.Donnette

Version: 2.0

Date: May 1999

Revision History

Date	Issue	Description of change
17/02/99	CENA first release v 1.0	
20/05/99	CENA second release v 2.0	Publication Version following PHARE DRG review.

LIST OF CONTENTS

1. INTRODUCTION.....	5
2. AIMS OF PD/3 TRIALS ANALYSIS	7
2.1 INTRODUCTION.....	7
2.2 FIRST LEVEL OF EVALUATION	8
2.2.1 Workload Comparison.....	8
2.2.2 Capacity comparison.....	8
2.2.3 Quality of service comparison.....	8
2.3 COMPLEMENTARY LEVELS OF EVALUATION.....	8
2.3.1 Operational level.....	8
2.3.2 HMI Level.....	8
3. TRIAL CONFIGURATION.....	9
3.1 INTRODUCTION.....	9
3.2 USE OF THE COMPUTER ASSISTANCE TOOLS AND DATA-LINK	9
3.3 TRAFFIC SAMPLES	9
3.4 CONTROLLERS	13
3.5 TRIAL TIMETABLE.....	14
4. MEASUREMENTS.....	17
4.1 INTRODUCTION.....	17
4.2 SUBJECTIVE MEASUREMENTS.....	17
4.2.1 ISA Scores.....	17
4.2.2 TLX Scores.....	18
4.2.3 Questionnaires	18
4.2.4 Observations	19
4.3 OBJECTIVE MEASUREMENTS.....	19
4.3.1 Workload.....	19
4.3.2 Quality of service.....	20
4.3.3 Capacity.....	20
5. ANALYSIS METHODOLOGY.....	21
5.1 INTRODUCTION.....	21
5.2 STATISTICAL TESTS USED	21
5.2.1 General description.....	21
5.2.2 Tests implementation	21
5.3 HYPOTHESES	22
5.3.1 Workload Hypotheses	22
5.3.2 Quality of Service Hypotheses	23

5.3.3 Capacity Hypotheses	24
5.4 MISSING ISA SCORES	24
5.5 CRITERIA FOR ACCEPTABLE RUNS	24
6. METHOD SUMMARY	27

LIST OF FIGURES

Figure 3-1 : Vertical plan of the PD/3 sectors and main traffic flows	11
---	----

LIST OF TABLES

Table 3-1 : Proposition of grown samples (number of flight plans / exercise)	12
Table 3-2 : PD/3 samples identification	13
Table 5-1 : Synthesis of runs	25

1. INTRODUCTION

This annex describes the aims of the CENA PHARE Demonstration 3 (PD/3) analysis, and discusses how the trial met those aims. The data and measurements collected during the trials follow the VAL recommendations as much as possible (see reference [5]). The analysis actually performed on them are described herein.

This page left intentionally blank

2. AIMS OF PD/3 TRIALS ANALYSIS

2.1 INTRODUCTION

The simulation concentrated on the air and ground systems which could be available in the 2005-2015 time-scale. The airspace simulated at CENA's site represented TMA departure (CDG), ETMA (TN/TB - ACC) and En-Route sectors (UN/XN - UAC). The operational objectives of CENA in PD/3 was to assess the advanced organisation through the analysis of the phases of flight from departure up to the integration by En-Route control.

The Air Traffic Controllers were assisted by the provision of a set of PHARE Advanced Tools (PATs) designed to aid the decision making process, to permit the timely exchange of data and to improve the aircraft role in the flight planning process.

The general objectives of the PD/3 trial can be summarised as follows:

- to determine the effect on controller workload and traffic throughput of the introduction of computer assistance tools from the PHARE Advanced Tools (PATs) programme and new Ground HMI (GHMI) ;
- to determine the effect on controller workload and traffic throughput of the increasing proportion of four dimensional flight management system (4-D FMS) and DL equipped aircraft.

The trials were designed to meet these objectives, with the analysis of results considering the PD/3 main criteria which were examined through three levels of evaluation.

The first level of evaluation represented these PD/3 main criteria which are further described below : controller workload, sector capacity and quality of service. These main topics were strongly linked together by many factors such as the 4D a/c DL equipment, the quality of sharing of tasks and the co-operation between PC and TC, the quality of co-ordination between the adjacent sectors, the set of advanced tools, and the GHMI¹ usability. As those main criteria had common and similar influence on each of these factors, they were investigated through specific evaluation themes:

- Assessment of the tools suitability and working methods (second level of evaluation) ;
- Assessment of the HMI usability (third level of evaluation).

Therefore, the first level of evaluation was supplied with both specific measures (objective² / subjective) and conclusions drawn from the second and third levels of evaluation. The second level of evaluation (operational level) was supplied with both specific measures (objective / subjective) and conclusions drawn from the third level of evaluation. The third level of evaluation (HMI level) was supplied with specific objective / subjective measures.

In association with the general objective of PD/3, the CENA Demonstration aimed at assessing the operational support provided by the Departure Manager (DM). This tool could be considered as a complement to the Arrival Manager tool developed within PD/2 and the NLR PD/3 Demonstration. This work was centred upon Paris Roissy-

¹ Presented in reference [1]

² Presented in the Annex [B] of the CENA Experimental Protocol for PD/3 (see reference [5])

CDG airport, and looked at the interactions and co-ordination issues between the Departure Manager and the surrounding en-route airspace.

2.2 FIRST LEVEL OF EVALUATION

2.2.1 Workload Comparison

One main objective of the PD/3 trial was to determine the effect of the introduction of computer assistance tools and DL on the controller's workload. The observation focused on the way the tools and DL could cause a reduction in controller workload by observing whether the controllers could cope with higher volumes of traffic. Section 5.1.1 describes the investigation into controller workload in more detail and Annex C1 reports the results.

2.2.2 Capacity comparison

In addition to controller workload, other measures can help to indicate whether the introduction of computer assistance tools and DL would have any effect on airspace capacity. Note that the Safety aspects were included in this part. Section 5.1.2 describes the investigation of such measures in more detail and Annex C1 reports the results.

2.2.3 Quality of service comparison

The objective is to show whether the introduction of the computer assistance tools and DL significantly improves the quality of service provided to ATM users. This analysis would help gain a degree of approval from the ATM users for the costs of eventual operational implementation of the advanced tools and DL. Section 5.1.3 describes the investigation into quality of service in more detail and Annex C1 reports the results.

2.3 COMPLEMENTARY LEVELS OF EVALUATION

2.3.1 Operational level

The controllers were assisted by the PHARE Advanced Tools (PATs). The new toolset windows including Co-operative Tools (CT) for ETMA and En-Route positions, Trajectory Editor and Problem Solver (TEPS) for all positions and DM for departure positions was carried out through associated operational procedures and specific working methods that are presented in detail in reference [3].

The sharing of tasks between PC and TC in accordance with the way they used the advanced tools they were provided with led us to this second level of assessment.

2.3.2 HMI Level

The third level of evaluation aimed at answering the following question: Do the HMI mechanisms correctly realise the operational procedures requirements?

For the advanced organisations, each controller working position was provided with dedicated tools and the interactions with the air and ground systems were managed through advanced Human-Machine Interfaces (see reference [1]). For the Baseline organisation, a reduced mode of this system was used with some assistance functions conveyed by the electronic environment. The computer assistance was strip-less and the air-ground communications were performed via R/T (no air-ground DL available).

The evaluation concerned the following points:

- GHMI quality of available dialogues ;
- GHMI quality of information displays.

Detailed results of these levels of Evaluation are reported in Annex C2.

3. TRIAL CONFIGURATION

3.1 INTRODUCTION

After a Pilot Phase led in March 1998, the trials were carried out over two periods of three weeks in May 1998 (Main Phase 1) and June 1998 (Main Phase 2), including a live aircraft in the beginning of June. Nine different controllers from varying nationalities participated each session of three weeks. PD/3 investigated the ATM in the time frames of 2005-2015 and focused on the management of the arrival and departure traffic in the Extended TMA and En-Route sectors.

3.2 USE OF THE COMPUTER ASSISTANCE TOOLS AND DATA-LINK

To accomplish the aims described above, the PD/3 trial comprised three organisations and different traffic volumes, as described below :

- a Baseline organisation which is close to today's operational with limited planning aids ;
- an Advanced organisation in which the PHARE Advanced Tools and a new GHMI are implemented to assist the controllers (A0) ;
- A30 and A70 which have the same functionality as A0, but in which are respectively introduced 30% and 70% of 4-D FMS and DL equipped aircraft.

In each organisation, three different traffic volumes were employed :

- 'Low' corresponding to today's traffic demand (June 96) ;
- 'Medium', June 96 x 1.5 ;
- 'High' June 96 x 2.25.

The analysis of the Baseline organisation against the Advanced organisation (A0) data investigated the differences due to the effect of introducing PATs and GHMI, whereas effects of increasing proportions of 4-D FMS / DL aircraft were shown by comparing A0 data against A30 and further against A70 data.

3.3 TRAFFIC SAMPLES

Several traffic samples were used in the runs, in order to vary the level of traffic demand and so to examine in further detail the responses of the controllers to the introduction of the computer assistance tools and DL.

The space model used in the PD/3 experiments was based on the existing one during summer 1996. The June 21st was the summer peak for French ACC(s) with 6405 realised Flight Plans overhead France. The traffic samples were built from data corresponding to this day.

The PD/3 sectors were of two sorts : the measured sectors and the feeder sectors.

The measured sectors comprised :

- the TMA departure sector which mainly concerns Ch de Gaulle Airport (CDG) with a specified volume in PARIS TMA. Only departure trajectories (SID) were used. Regarding the technical constraint of the Departure Manager (DM), a single configuration was chosen for CDG and its two runways; this configuration was facing west. As a result, the other airports in TMA were on the same configuration. In June 1996, the runway capacity of CDG was

approximately 80/85 movements increasing. The traffic samples for the departure sector were always built with two runways, given the fact that any additional runway would lead to a change of the space model for 2005 and 2015. All SID(s) and STAR(s) were considered ;

- the ETMA sector was represented by sectors TN and TB considered as combined. They extend from ground up to level 245 and have common boundaries in the northern part with Belgium and UK. These sectors are dedicated for departure flights. They integrate northbound departures from the TMA to their own overflying traffic. They manage departing and arrival flights for an international airport, Lille Lesquin (LFQQ) and take into account the descending traffic to Brussels. In 1996, the hourly capacity for the combined sectors was 30 a/c. The main exchanges are realised with Belgium, UK and French Centres;
- the Upper En-Route sectors were represented by volumes UN and XN also considered as combined. They approximately have the same boundaries as TN/TB immediately below. They extend from level 245 to 320 for UN and 320 up to unlimited for XN. These sectors are mainly dedicated to the overflying traffic. They also integrate northbound departure flights from the TMA into their own overflying traffic. In 1996, the hourly capacity for the combined sectors was 37 a/c.

The feeder sectors were the following ones :

- one ETMA sector (TE) which belongs to Paris ACC and dedicated to the arrival traffic for the TMA. In 1996, the hourly capacity for the TE sector was 28 a/c.
- two En-Route upper sectors (UZ/ZU and UR/UY), both considered as combined. In 1996, the hourly capacity for the UZ/ZU sector was 45 a/c, and 42 a/c for UR/UY sector.

- Figure 3-1 below gives an overview of the PD/3 sectors and main traffic flows:

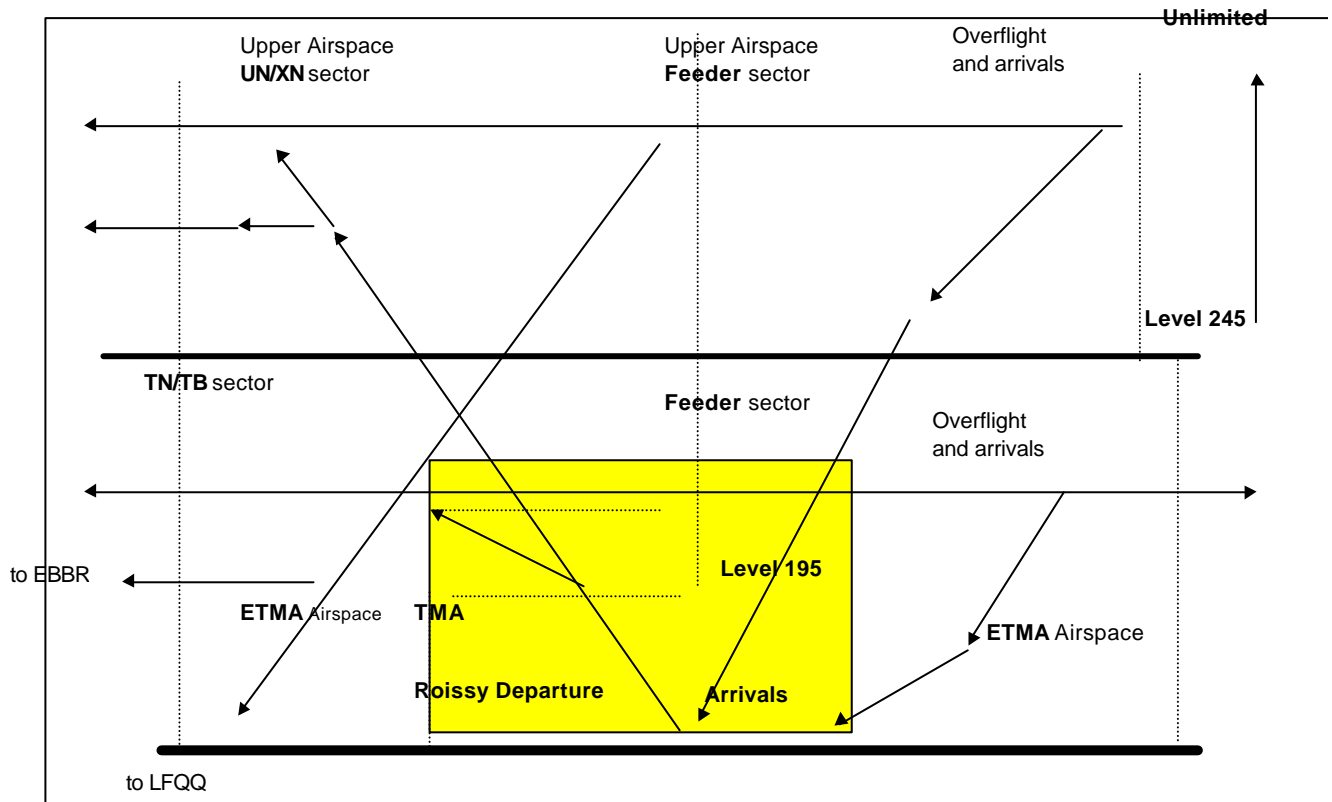


Figure 3-1: Vertical plan of the PD/3 sectors and main traffic flows

In the CENA PD/3 trial, the departure position was played with a new concept of tool, the Departure Manager. To be fully operational, the DM had to evaluate the runways availability. It had so to be informed of all incoming flights. The runway configuration being chosen, all SID(s), STAR(s) and Approaches controlled by Charles De Gaulle have been identified. Some other trajectories were added for Orly, Le Bourget, Villacoublay, Toussus le Noble because they are under CDG's responsibility.

In order to generate the samples, the methodology consisted in creating a basic sample, called reference sample, from which the others could be derived from applying specific recommendations (see reference [4]). From the low traffic sample sequences made for the Baseline scenario have been generated the corresponding sequences for the medium and heavy samples. Initially, before the PD/3 experiment, the sequences were adjusted in order to meet the increasing ratios : the low traffic samples corresponded to the traffic existing in June 1996 ; the medium traffic samples corresponded to this traffic $\times 1,5$; and the high traffic samples corresponded to the traffic in June $\times 2,25$. In order to prevent learning effects, data processed procedures have been set up allowing to modify the callsign information and the hours.

However, these ratios could not be maintained. For the Departure position, the obtained sample reached already the maximum hourly capacity with two tracks. In order to apply the growth ratios of traffic load, it would have been necessary to increase the number of tracks and to modify the space model, which was not planned. The decision made not to increase the traffic load on the Departure

position had consequences on the resulting traffic load on the ETMA and En-Route sectors, for the Medium and High samples.

During the Pilot Phase, the first resulting samples have been proposed to the controllers. Following controllers many requests and actions, the systems encountered some difficulties, the latter being mostly linked to the heavy load of flight plans. Another intervention on the samples consisted in:

- eliminating a number of flights, mainly those which had not to cross any measured sectors ;
- and, as this first solution was not sufficient, reducing the traffic load on the measured sectors (ETMA and En-Route).

The resulting ratios of traffic load were then less important than the expected ones. Thus they allowed to lead to an acceptable system load during the experiment.

Table 3-1 below presents the total number of flight plans which was proposed in each kind of sample (for the main phases):

	TMA. Ch de Gaulle Airp*	ETMA sector TN/TB	En route sector UN/XN
Low traffic 163 FPL	Arrivals = 50 a/c Departures = 54 a/c from/to all airports	41 a/c	56 a/c
Medium traffic 188 FPL	Arrivals = 52 a/c Departures = 54 a/c from/to all airports	53 a/c	77 a/c
Heavy traffic 220 FPL	Arrivals = 52 a/c Departures = 64 a/c from/to all airports	63 a/c	96 a/c

Table 3-1: Proposition of grown samples (number of flight plans / exercise)

*These values include departures towards the north of : LFPB (Le Bourget), LFPO (Orly), LFPN (Toussus), LFPV (Villacoublay).

This table has to be interpreted this way:

- each sample was expected to last 90 minutes, whereas measures were carried out only the last 60 minutes ;
- the basic sample Low traffic was already positioned in the high limit of the sectors capacities ;
- a same FPL might cross the three sectors ;
- there was virtually no increase in the arrival or departure traffics.

The induced capacities have been calculated and presented in annex C1 (see chapter 4 on Capacity).

The samples were categorised into 'low', 'medium' and 'high' traffic volumes, corresponding to the aircraft throughput. These traffic samples were classified as in Table 3-2 below.

RUN(S)			
	Low traffic 163 FPL	Medium traffic 188 FPL	High traffic 220 FPL
Baseline	Sample N° 450	Sample N° 452	Sample N° 456
Adv 0% D/L	Sample N° 451	Sample N° 453	Sample N° 457
Adv 30% D/L		Sample N° 454 / 454B	Sample N° 458 / 458B
Adv 70% D/L		Sample N° 455 / 455B	Sample N° 459 / 459B

TRAINING			
Adv 30% D/L	N° 462	162 FPL	
Adv 70% D/L	N° 461	162 FPL	
Adv 100% D/L	N° 460	54 FPL	No departure, no arrival.
Adv 70% D/L	N° 475	102 FPL	
Adv 70% D/L	N° 476	102 FPL	
Minimal level for training N° 480 with 50 FPL			
Class Very Low traffic, Advanced with 70% 4D, departures and arrivals.			

Table 3-2: PD/3 samples identification

This section is further described in Annex F.

3.4 CONTROLLERS

A total of twenty seven controllers of four nationalities participated in the PD/3 trials, as described in Annex [B] of the CENA Experimental Protocol for PD/3 (see reference [5]).

The experiment was based on repeated measurements of two teams of six controllers. All teams performed identical tasks after an adequate training period. Four different tasks, respectively organisations, each performed under low, medium and high traffic load resulted in twenty measured runs per teams. The controllers rotated between the tactical, planning and feeder positions for the En-Route and ETMA positions. The controllers rotated between the tactical and planning positions for the Departure working positions.

The controllers were trained using a full six days course, entailing a mixture of theoretical lessons and practical training on the full system. The day before the measured runs began, all controllers undertook two refresher runs.

3.5 TRIAL TIMETABLE

In each first week, Monday was used for refresher training, and during the other eight days a total of twenty simulation runs of 90 minutes duration each were carried out, scheduled according to the timetables below.

- Controller rotation indicators: 12 = Controller 1 as PC / Controller 2 as TC (the other controllers may play feeder positions for this run).
- Organisations: Baseline (B), Advanced 0% DL (A0), Advanced 30% DL (A30) and Advanced 70% DL (A70).
- Traffic volumes: Low (L), Medium (M) and High (H).

Tables below give an overview of the PD/3 CENA timetable set of runs:

- for the En-Route sector :

Week	Day	Morning	Afternoon	Afternoon
1	All	Training	Training	Training
Week	Day	Run 1	Run 2	Run 3
	1	Training	Training	Training
	2	12 - B - L	41 - A30 - M	<i>spare</i>
2	3	23 - A70 - H	12 - B - H	41 - A0 - M
	4	12 - A30 - H	34 - B - M	23 - A0 - L
	5	12 - A0 - H	34 - A70 - M	<i>spare</i>
	1	41 - A70 - M	23 - A0 - H	<i>spare</i>
	2	12 - A0 - L	41 - B - M	23 - A30 - H
3	3	34 - A0 - M	23 - B - H	12 - A70 - H
	4	34 - A30 - M	23 - B - L	<i>spare</i>
	5	General	Debriefing	<i>spare</i>

- for the ETMA sector :

Week	Day	Morning	Afternoon	Afternoon
1	All	Training	Training	Training
Week	Day	Run 1	Run 2	Run 3
	1	Training	Training	Training
	2	12 - B - L	31 - A30 - M	<i>spare</i>
2	3	31 - A70 - H	23 - B - H	12 - A0 - M
	4	23 - A30 - H	31 - B - M	12 - A0 - L
	5	31 - A0 - H	12 - A70 - M	<i>spare</i>
	1	31 - A70 - M	23 - A0 - H	<i>spare</i>
	2	23 - A0 - L	12 - B - M	31 - A30 - H
3	3	31 - A0 - M	31 - B - H	23 - A70 - H
	4	12 - A30 - M	23 - B - L	<i>spare</i>
	5	General	Debriefing	<i>spare</i>

- for the Departure sector :

Week	Day	Morning	Afternoon	Afternoon
1	All	Training	Training	Training
Week	Day	Run 1	Run 2	Run 3
	1	Training	Training	Training
2	2	12 - B - L	21 - A30 - M	<i>spare</i>
	3	12 - A70 - H	21 - B - H	12 - A0 - M
	4	21 - A30 - H	12 - B - M	21 - A0 - L
	5	12 - A0 - H	21 - A70 - M	<i>spare</i>
	1	12 - A70 - M	21 - A0 - H	<i>spare</i>
	2	12 - A0 - L	21 - B - M	12 - A30 - H
3	3	21 - A0 - M	12 - B - H	21 - A70 - H
	4	12 - A30 - M	21 - B - L	<i>spare</i>
	5	General	Debriefing	<i>spare</i>

This page left intentionally blank

4. MEASUREMENTS

4.1 INTRODUCTION

During the trials a large quantity of data was recorded, which may be partitioned into two distinct categories : subjective and objective measurements.

The data from subjective measurements was gathered from the following sources : controller questionnaires, controller debriefings, video recording of the runs, and notes of the specialist observers (ergonomist experts).

The objective measurements were automatically recorded by the ground system. Several categories were foreseen for each evaluation level :

- off line logging (pre-run logging) ;
- in line logging :
 - * in line immediate basic logging (i.e. low level data available in a straightforward lay out just after each run in order to enrich the post run debriefing) ;
 - * in line differed basic logging (i.e. low level data provided after the overall main phase) ;
 - * in line differed processing (i.e. upper level data such as Sum, Percentage, Average, ... directly extracted from the low level data).

The detailed tables of objective measurements are presented in Annex [B] of the CENA Experimental Protocol for PD/3 (see reference [5]).

4.2 SUBJECTIVE MEASUREMENTS

The subjective data representing the opinion of the controllers was gathered from a variety of sources during the trial. The three sources of quantitative, subjective measures in the PD/3 trials were ISA, TLX and the questionnaire responses. ISA and TLX were used to record the controllers' perceived workload. In addition to the questionnaires, the debriefs and controllers' comments were used to obtain controllers' opinion. Video recordings were made for subsequent analysis of the controllers' actions.

4.2.1 ISA Scores

Subjective workload estimates were collected during the course of the simulation runs. The Instantaneous Self Assessment (ISA) measurement is a method to assess workload in real time. The ISA workload measurement technique has been developed over a number of years by the NATS ATM Development Centre, based at Hurn. Prior to participating in the trial, the controllers were briefed on the ISA assessment technique.

The ISA panel consists of five numbered buttons each representing a defined level of workload : Under-utilised, Relaxed, Comfortable, High and Excessive.

Every two minutes during the simulations, the controllers had to give their estimates of workload by pushing the button corresponding to the level of workload experienced.

The ISA scores were recorded in a separate log file. Various measures centred on the mean ISA scores and recorded by each controller position (PC and TC) for each run have been investigated.

4.2.2 TLX Scores

The NASA-Task Load index (TLX) provided a summary workload estimate immediately after completion of each run. TLX was developed at NASA-AMES research centre in the USA, and has been used extensively for measuring pilots' subjective workload. TLX identifies six factors contributing to workload : Mental demand, Physical demand, Time pressure, Own performance, Effort expended and Frustration.

At the end of a run the TLX input dialogue popped up on each of the controller screens, asking them first to determine the relative importance of the six factors for this run by pairwise comparison of the factors, and then to rate their workload for each factor on a 20-point scale from 'low' to 'high'. From combination of weights and ratings, an overall score was calculated to estimate the controller's workload.

The reason for recording workload with TLX was to gain a more detailed view of the causes underlying the controllers' perception of workload. The results were compared with the ISA results to identify any discrepancies between the two measures. TLX was especially useful for investigating the causes of the observed ISA results, with respect to the individual workload factors as listed above.

4.2.3 Questionnaires

In order to assess the degree of controller acceptance of the computer assistance tools, GHMI and operational concept, a part of subjective data was collected through questionnaires. The final questionnaires took place before the final debriefing so as to avoid mutual influences that might result from discussions between controllers. The questionnaires contained categories of questions as follows :

First level of evaluation

- PD/3 operational concepts in accordance with the main topics (Quality of service, Sector capacity, Controller Workload) ;
- the operational aspects of the simulation, e.g. traffic handling, ATC procedures.

Second level of evaluation

- the acceptance of each specific tool and its functions ;
- confidence and trust in using the tools, GHMI and concepts to their full potential.

Third level of evaluation

- the overall GHMI aspects, e.g. displays, human-computer dialogues and interaction.

The simulation environment

- the conditions of training ;
- the simulation organisation (traffic samples, rotations, ...).

The final questionnaires were initially split up into sections devoted to:

- specific organisation separately ;
- themes (main topics, sub topics) ;
- level of evaluation ;
- controller roles ;
- organisations.

The questionnaires were designed to format the responses to the individual questionnaire items onto a uniform rating scale, making possible the quantitative analysis and significance testing of the controller responses. All items in the questionnaires were given as statements. The controller then indicated the extent to which he/she agreed or disagreed with the statement by selecting the appropriate level on a six-point scale. The levels ranged from “strongly disagree” to “strongly agree”. Additional written comments were also allowed.

Detailed results of the final questionnaires are reported in Annex D.

4.2.4 Observations

During the trial, the performance of the controllers was observed by specialists observers. The observers were provided with analysis charts allowing them to quickly collect all events that were of interest through an adapted lay-out and a clear events coding. The purpose of the observations was to subjectively assess the controller body language, the controller responses to particular events, and the occurrence of certain actions. Throughout the observations, any relevant comments were also noted and formally recorded as part of the post-trial analysis.

4.3 OBJECTIVE MEASUREMENTS

The objective measurements that were considered during the analysis of results are presented for the main topics below.

4.3.1 Workload

No objective measurement directly devoted to the Workload topic was available. Indeed, R/T recordings have been performed but they could not be processed within the time-frame assigned for the whole data analysis. Nevertheless, other measurements that were linked to the use of advanced tools (TEPS, APD, DM) could be considered as of interest since time and attention demanding when interacting with the HMI might directly contribute to the controller workload.

Objective measurements devoted to the use of TEPS (second level of evaluation) :

- number of trajectory edition (TEPS in EDIT MODE) ;
- number of trajectory consultation (TEPS in DISPLAY MODE) ;
- duration of trajectory edition.

Objective measurements devoted to the use of the APD (second level of evaluation) :

- percentage of PROSITs in alarm ;
- percentage of PROSITs manually removed.

Objective measurements devoted to the use of the DM (second level of evaluation) :

- number of manual re-sequencing ;
- number of DEP TEPS activation by TC ;
- percentage of choice of alternative trajectory by PC.

4.3.2 Quality of service

The quality of service provided to the ATM users is an extremely difficult parameter to measure, but one element of it is the extent to which the aircraft are allowed to follow their requested flight plans. The following objective measures were adopted as indicators of the quality of service offered to ATM users by each organisation :

- average time spent by aircraft in sector ;
- percentage of time aircraft were at Preferred Cruise Flight Level (PCFL) ;
- percentage of aircraft with a take-off delay;
- average of delays ;
- average number of ground constraint violations.

4.3.3 Capacity

The capacity of the sector was directly recorded as an objective measure, but has to be considered as an input data rather than as a data which could be interpreted as a result of the simulation. These data aimed at calibrating the « traffic volume » variable (Low, Medium and Heavy traffic), so as to measure (in a qualitative and quantitative way) the conditions of integration of these traffic volumes by the team of controllers according to the ORG (Baseline, A0, A30 and A70) and the sector (TMA, ETMA, En-Route).

The parameter used was the number of aircraft flying through the sector. It was initially defined by the traffic samples used during the trial.

As it was done in PD/1, objectives measures of safety were chosen as an indication of whether a capacity increase may be possible by implementation of the Advanced organisations. For the PD/3 analysis, the following objective measures were adopted as possible additional indicators of the airspace capacity change caused by each organisation:

- average number of Short Term Conflict Alerts (STCAs) activated per run ;
- average duration of those STCAs ;
- average number of minimum safe separation infringements per run ;
- average duration of those minimum safe separation infringements.

All the objective measurements that were initially forecast (in fact only a part could be used) are presented in Annex [B] of the CENA Experimental Protocol for PD/3 (see reference [5]).

5. ANALYSIS METHODOLOGY

5.1 INTRODUCTION

In the analysis plan, null and alternative hypotheses were developed for each detailed objective of the trial, and the appropriate subjective and objective data required to test the hypotheses was identified.

For the measures relevant to each detailed objective, descriptive statistics, usually in the form of a graph or histogram, were produced prior to statistical analysis when of interest.

The nature of the principal measures used, ISA and TLX, meant that non-parametric statistical tests were the main analysis tool. These tests were chosen in order to determine whether a statistically significant difference existed between any of the organisations in each of the measures.

Pairs of matched observations from the same controller and traffic sample were used to compare the organisations, wherever this was feasible.

Workload, sector capacity and quality of service to ATM users were investigated in order to meet the aims of PD/3.

5.2 STATISTICAL TESTS USED

5.2.1 General description

The effect of introducing the PATs and the operational procedures they support were examined by comparing the Baseline organisation with the Advanced 0% organisation. The statistical tests used were the Wilcoxon Matched-Pairs Signed-Ranks and the Binomial test (Questionnaires).

The effect of different proportions of 4-D FMS equipped aircraft were examined by comparing A0 vs. A30 vs. A70 organisations. The tests which were used were the Friedman Two-Way Analysis of Variance (Anova).

- The Wilcoxon and Friedman tests are standard statistical tests used for experimental designs with repeated measurements (matched pairs of observations) and for non-parametric data (i.e. without assumption about the data distribution).
- The Binomial Test was used for detecting significant differences in the questionnaire responses. It tests two-class frequency distributions (binomial distributions) formed by combining all responses on the left-hand ("disagree") part of the scale against all combined responses on the right-hand ("agree") part of the scale.

The level of statistical significance chosen for each test was set at $p < 5\%$.

All the statistical results were combined with the results gathered from the qualitative sources such as the observations and the controller comments, thus enabling a fair interpretation of the statistical results.

5.2.2 Tests implementation

Basically, two statistical, non-parametric tests were used so as to get an optimal reliability from a fairly reduced number of samples. These tests were the Wilcoxon Signed Rank (WSR) test and the Friedman test.

Wilcoxon test - The Wilcoxon signed rank test is used on a single family of samples, and it aims at testing the Null Hypothesis H_0 "the average over the samples is not different from 0" against H_1 "the average over the samples is non-zero".

The latter means that the distribution of the samples is likely to come from an average of the underlying statistic different from zero, and the former hypothesis means that the average on the samples does not significantly differs from 0 and is likely come from chance.

The test can be used more easily with better reliance by testing both H_0 "The average is ≤ 0 " against H_1 "The average is > 0 ", and then the opposite.

We applied the WSR test to matched pairs (X_i, Y_i) so as to actually use the test on the statistic $(Y_i - X_i)$. Thus we tested whether $X_i < Y_i$ or $X_i \geq Y_i$, and then if the test was in favour of the latter, whether $Y_i < X_i$ or $Y_i \geq X_i$. We had to make sure we actually had matched pairs results (i.e. with controllers kept the same) so that a statistic like $X_i - Y_i$ could make sense.

From a mathematical point of view, this test relies on the distribution of the W_n^+ statistic, which is the sum of the ranks the elements of strictly positive value, the elements being sorted in ascending order of their absolute value.

Friedman test - The Friedman test is a "mere" ANOVA (ANalysis Of VAriance) on the rank of several (i.e. more than 2) statistics, the rank being that of the considered statistic in a matched tuple. Say the tuples are disposed in rows, and the columns then each contains the statistics on a given tested element.

Here we used matched triples. The ANOVA test consists of a measure of the variance of the resulting sum of the ranks in every column, divided by the intrinsic variance of the column. The final result is scaled in order to reflect the number of rows and columns. The original ANOVA test suggests the intervention of the variance of the average on lines, but this variation is 0 for the ranks are considered.

For both test the question of possible equal measures arose. For both tests the easy solution of a randomly ordering of the equalities was chosen. For the WSR test it was not much of a problem since its high robustness and the possible difference of 1 induced to the W_n^+ statistic was in no way significant. The problem was very different for the Friedman test however since two differently ordered ranks could lead to a much greater difference in the result. Therefore a good pseudo-random choice had to be implemented as well. This sometimes led to different results in some cases where two successive computations on the same data did not always give rise to the same result. Though this proves the quality of the pseudo-random function, this made the double-checking of the results hard.

Binomial test - The Binomial Test gives the probability of an observed proportion of the two frequencies, under the hypothesis that there is no true difference between the two. This hypothesis is equivalent with saying that there is no significant ratio of answers on either side of the scale, and thus observed differences are caused by chance alone. The more different the two observed frequencies are, the smaller becomes this probability, and in case of a low probability of less than five percent ($p \leq 0.05$) a significant trend, either to agree or to disagree, is stated.

5.3 HYPOTHESES

Each of these investigations were accompanied by an appropriate null and alternative hypothesis. The null hypothesis which was used in the statistical tests was that there was no difference between compared situations.

5.3.1 Workload Hypotheses

In order to examine controller workload, subjective and objective measurements were conducted. The following null hypotheses (H0) were stated to statistically test the data, using data pooled across controller rotation and traffic volumes. H1 referred to the alternative hypothesis which would be accepted if the H0 were rejected:

Baseline Vs A0

- H0-1: There is no difference in term of perceived Workload between the Baseline and A0.
- H1-1: Perceived Workload is different as an effect of introducing PATs and GHMI.

Baseline Vs A70

- H0-2: There is no difference in term of perceived Workload between the Baseline and A70.
- H1-2: Perceived Workload is different as an effect of introducing PATs, GHMI and 4-D FMS/DL equipped aircraft.

A0, A30 and A70

- H0-3: There is no difference in term of perceived Workload between different proportions of 4-D FMS/DL equipped aircraft.
- H1-3: Perceived Workload is different, depending on the percentage of 4-D FMS/DL equipped aircraft.

The statistical tests were applied separately for low, medium and high traffic load for each of the workload measurements described in the section 4.3.1.

5.3.2 Quality of Service Hypotheses

The following set of hypotheses was tested, using data pooled across both controller rotation and traffic volumes :

Baseline Vs A0

- H0-1: There is no difference in term of Quality of service between the Baseline and A0.
- H1-1: Quality of service is different as an effect of introducing PATs and GHMI.

Baseline Vs A70

- H0-2: There is no difference in term of Quality of service between the Baseline and A70.
- H1-2: Quality of service is different as an effect of introducing PATs, GHMI and 4-D FMS/DL equipped aircraft.

A0 Vs A30 Vs A70

- H0-3: There is no difference in term of Quality of service between different proportions of 4-D FMS/DL equipped aircraft.
- H1-3: Quality of service is different, depending on the percentage of 4-D FMS/DL equipped aircraft.

The tests were carried out using the objective indicators of sector capacity as described in the section 4.3.2.

5.3.3 Capacity Hypotheses

The approach to the analysis of the measures was similar to that for the workload measures. First, the capacity measures of each organisation within the same traffic sample and controller were compared. The following null hypotheses (H0) were stated to statistically test the capacity data, using data pooled across controller rotation and traffic volumes. H1 referred to the alternative hypotheses which would be accepted if the H0 were rejected.

Baseline Vs A0

- H0-1: There is no difference in term of Sector capacity between the Baseline and A0.
- H1-1: Sector capacity is different as an effect of introducing PATs and GHMI.

Baseline Vs A70

- H0-2: There is no difference in term of Sector capacity between the Baseline and A70.
- H1-2: Sector capacity is different as an effect of introducing PATs, GHMI and 4-D FMS/DL equipped aircraft.

A0 Vs A30 Vs A70

- H0-3: There is no difference in term of Sector capacity between different proportions of 4-D FMS/DL equipped aircraft.
- H1-3: Sector capacity is different, depending on the percentage of 4-D FMS/DL equipped aircraft.

The tests were carried out using the objective indicators of sector capacity as described in the section 4.3.3.

5.4 MISSING ISA SCORES

Most of time, ATCO felt at ease with using this tool and quite willing to use evaluation tools ; however, it happened that some ATCO simply forgot to assess and input ISA score on the panel ; it is clear that this misuse of the tool had an effect on general results.

For this reason, some rules were established in order to process ISA data:

- a score of 5 (maximum) is set for all forgotten marks ;
- runs for which number of forgotten mark exceed 15 are excluded from calculation of average and determination of differences between organisations.

Only 17 samples were excluded on a total of 240.

5.5 CRITERIA FOR ACCEPTABLE RUNS

The criteria for acceptable run were:

- the duration for simulation time had to be of at least 45 minutes ;

- the cumulative duration of frozen periods within one run had to be lower than 20 minutes.

Table 5-1 below presents a synthesis of the runs which have been played during the 2 sessions A and B.

Phase	launched	aborted	successful	measured	average duration for measured runs
A	22	2	20	20	1H08'
B	24	2	22	20 (Results of 2 runs could not be included in the comparative study)	1H09'

Table 5-1 : Synthesis of runs

This page left intentionally blank

6. METHOD SUMMARY

After a Pilot Phase led in March 1998, the trials were carried out over two periods of three weeks in May 1998 (Main Phase 1) and June 1998 (Main Phase 2), including a live aircraft in the beginning of June. A total of twenty seven controllers of four nationalities participated in the PD/3 trials, nine different ones in each of the three weeks sessions. The simulation concentrated on the air and ground systems which could be available in the 2005-2015 time-scale. The airspace simulated at CENA's site represented TMA departure (CDG), ETMA (TN/TB - ACC) and En-Route sectors (UN/XN - UAC).

The PD/3 trial was defined in terms of three organisations corresponding to : a Baseline for comparison, the introduction of PHARE Advanced Tools and a new GHMI (A0) , and three traffic "mixes" (A0, A30 and A70), indicating the percentage of aircraft with 4-D FMS and DL equipment.

Different traffic samples were used in the simulation runs, in order to vary the levels of traffic demand and to examine in further detail the responses of the controllers to the introduction of the computer assistance tools and DL. The samples were categorised into 'low', 'medium' and 'high' traffic volumes, corresponding to the aircraft throughput per hour.

Null hypotheses were formulated in order to compare the organisations and traffic mixes with respect to controller workload, airspace capacity and quality of service provided to ATM users. Tools suitability and HMI usability were also investigated. By testing these hypotheses, the objectives of the PD/3 trials were met.

The null hypotheses were tested using both subjective and objective measures. These measures were derived from the data that was gathered during the runs from various sources, such as the ground simulation system itself, controller workload self-assessment, controller questionnaires, and controller debriefings.

The statistical tests used were the Wilcoxon Matched-Pairs Signed-Ranks and the Binomial test (Questionnaires). The Wilcoxon and Friedman tests are standard statistical tests used for experimental designs with repeated measurements (matched pairs of observations) and for non-parametric data (i.e. without assumption about the data distribution). The Binomial test was used for detecting significant differences in the questionnaire responses.

System software and equipment problems during the trials meant that from time to time some data was not collected for a particular run. Criteria for usable runs were created, and only data from runs meeting those criteria were considered in the final analysis.