PROGRAMME FOR HARMONISED AIR TRAFFIC MANAGEMENT RESEARCH IN EUROCONTROL

PHARE

EUROPEAN ORGANISATION FOR THE SAFETY OF AIR NAVIGATION, EUROCONTROL

DOC 96-70-24

PD/1 FINAL REPORT
Annex A

Experimental Design and Methods

PHARE/NATS/PD1-10.2/SSR;1.1

EUROCONTROL

EUROCONTROL
96 rue de la Fusée
B-1130 BRUXELLES

Prepared by: D Hudson
S Kay

Date: January 1997

REVISION HISTORY

May 1996	Draft version 0 issued to DRA for comment
June 1996	Draft version 1 following internal NATS review
July 1996	Draft version 2
September 1996	Draft version 3, with minor changes following review by PD3CG
October 1996	Version 1.0, with minor changes, approved by the PCC
January 1997	Version 1.1, issued as a final document in Volume 2, approved by the PHARE Management Board

LIST OF CONTENTS

1. INTRODUCTION 5

2. AIMS OF PD/1 TRIALS ANALYSIS 6
2.1 INTRODUCTION 6
2.2 WORKLOAD COMPARISON 6
2.3 CONTROLLER APPROVAL 6
2.4 CAPACITY COMPARISON 6
2.5 QUALITY OF SERVICE COMPARISON 7

3. TRIAL CONFIGURATION 8
3.1 INTRODUCTION 8
3.2 USE OF THE TOOLS AND DATALINK 8
3.3 TRAFFIC SAMPLES 8
3.4 CONTROLLERS 9
3.5 TRIAL TIMETABLE 10

4. MEASUREMENTS 11
4.1 INTRODUCTION 11
4.2 SUBJECTIVE MEASURES 11
4.2.1 ISA Scores 11
4.2.2 TLX Scores 12
4.2.3 Questionnaires 12
4.3 OBJECTIVE MEASURES 13

5. ANALYSIS METHODOLOGY 14
5.1 INTRODUCTION 14
5.1.1 Workload Investigation 14
5.1.2 Capacity Investigation 15
5.1.3 Quality Of Service Investigation 15
5.2 STATISTICAL TESTS USED 15
5.3 HYPOTHESES 16
5.3.1 Workload Hypotheses 16
5.3.2 Capacity Hypotheses 16
5.3.3 Quality Of Service Hypotheses 17
5.4 MISSING ISA SCORES 17
5.5 CRITERIA FOR ACCEPTABLE RUNS 17

6. METHOD SUMMARY 19

7. ABBREVIATIONS 20

LIST OF TABLES

Table 3-1 - Categorisation of traffic samples into traffic volumes 9
Table 3-2 - The timetable of runs 10
Table 4-1 - Timetable for questionnaires 13
Table 5-1 - Unacceptable runs 18

This page is intentionally blank

1. INTRODUCTION

This annex describes the aims of the PHARE Demonstration 1 (PD/1) analysis, and discusses how the trial met those aims. The data and measurements collected during the trials, and the analysis actually performed on them, are described herein.

2. AIMS OF PD/1 TRIALS ANALYSIS

2.1 INTRODUCTION

The simulation represented en-route airspace circa the year 2000, based on the UK New En-Route Centre (NERC) North Sea sectors 10 and 11. However, the ATC procedures simulated would not be implemented until after that date, say the year 2015.

The planner and tactical controllers' basic roles were not changed significantly from the NERC concept, but assistance was provided in terms of "computer assistance tools" and datalink. The introduction of these new elements supported the introduction of new control procedures which, whilst still based on current practices, provided conflict-free flight trajectory planning based upon "looking ahead" twenty minutes or so.

The objectives of the PD/1 trial can be summarised as:

- to determine the effect on controller workload and traffic throughput of the introduction of computer assistance tools from the PHARE Advanced Tools (PATs) programme;
- to determine the effect on controller workload and traffic throughput of the increasing proportion of four dimensional flight management system (4-D FMS) equipped aircraft with full duplex datalink;
- to gain a degree of controller approval for the computer assistance tools introduced.

The trials were designed to meet these objectives, with the analysis of results considering the four main criteria described below.

2.2 WORKLOAD COMPARISON

One main objective of the PD/1 trial was to determine the effect of the introduction of the computer assistance tools and datalink on the controllers' workload. If the tools and datalink cause a reduction in controller workload, then it is reasonable to conclude that controllers could cope with higher volumes of traffic. Section 5.1.1 describes the investigation into controller workload in more detail and Annex C (Section 2) reports the results.

2.3 CONTROLLER APPROVAL

Gaining a degree of controller approval for the computer assistance tools was an important feature of the trial, since tools which are accepted and used well could facilitate a workload reduction and should be developed further. This aim is addressed in more detail in Section 4.2.3 and the results reported in Annex D.

2.4 CAPACITY COMPARISON

In addition to controller workload, other measures can help to indicate whether the introduction of computer assistance tools and datalink would have any effect on airspace capacity. Section 5.1.2 describes the investigation of such measures in more detail and Annex C (Section 3) reports the results.

2.5 QUALITY OF SERVICE COMPARISON

If it can be shown that the introduction of the tools and datalink significantly improve the quality of service provided to airlines, it will help gain a degree of approval from the airlines for the costs of eventual operational implementation of the tools and datalink. Section 5.1.3 describes the investigation into quality of service in more detail and Annex C (Section 4) reports the results.

3. TRIAL CONFIGURATION

3.1 INTRODUCTION

The trials were carried out over a period of eight weeks, from the 23 October 1995 until 15 December 1995, using the NATS Research Facility (NRF) real-time simulator. Four different controllers from varying nationalities participated each week, although each sector's tactical/planner team was usually from the same nation. The environment simulated represented en-route airspace circa the year 2000, based on the UK New En-Route Centre (NERC) North Sea sectors 10 and 11.

3.2 USE OF THE TOOLS AND DATALINK

To accomplish the aims described above, the PD/1 trial comprised three "organisations" and three traffic "mixes", as described below:

ORG 0 : Organisation 0 was the reference baseline organisation. It was derived from the Operational Display and Input Devices (ODID) scenario, which operates without paper flight strips.

ORG 1 : Organisation 1 included the introduction of the computer assistance tools derived from the PHARE Advanced Tools (PATs) 2000 Toolset.

ORG 2 : Organisation 2 was similar to ORG 1 but with the introduction of datalink for communication with aircraft capable of 4-D navigation. A live aircraft, the DRA (Bedford) BAC 1-11, was also included in this Organisation.

Mix 0 : This traffic sample contained an aircraft population as forecast for the year 2000. None of the aircraft were equipped with 4-D flight management system (FMS).

Mix 30 : The Mix 0 traffic samples, but with 30% of the aircraft being equipped with 4-D FMS.

Mix 70 : The Mix 0 traffic samples, but with 70% of the aircraft being equipped with 4-D FMS.

3.3 TRAFFIC SAMPLES

Several traffic samples for each mix were simulated in the runs, in order to vary the level of traffic demand and so examine in further detail the responses of the controllers to the introduction of the tools and datalink. Within a given level of traffic demand, several samples were used to ensure the generality of the results and to prevent learning effects perturbing the results of the later simulation runs.

The traffic samples used in the experiments were based on samples previously used and validated for the NATS' Computer Assistance for En-Route (CAER) Trials. The samples consisted of projected 'busy day' traffic flows for the year 2000, and were generated by

'growing' a selected base day, Friday 17 July 1992, to the traffic demand forecast for the year 2000. Then, by increasing the numbers of aircraft per route by 20% and 40%, the traffic flow was further increased to provide higher volume traffic samples. Finally, the traffic demand was smoothed, so that major airports' hourly limits were not exceeded, and a 75 minute 'slice' taken from the day's traffic.

The labels for the traffic samples identify various details about the samples:

- the first number (13, 14 or 15) denotes respectively Easterly traffic flow, Easterly/Westerly mix, and Westerly traffic flow; these correspond to the expected pattern of traffic at different times of the day
- the second number (20 or 40) denotes respectively 20% or 40% more traffic per hour compared to the baseline sample; if absent there was no additional traffic growth
- the third number (6 or 7) denotes the random number seed used in the automatic generation of the traffic sample.

The samples were categorised into "low", "medium" and "high" traffic volumes, corresponding to the aircraft throughput per hour. These traffic samples were classified as in Table 3-1, below.

Traffic Sample	Aircraft Throughput per Hour		Classification
	Sector 10	Sector 11	
s_13_6	31	33	"Low"
s_13_7	30	34	"Low"
s_14_40_6	46	54-55	"Medium"
s_14_40_7	45-46	57-59	"Medium"
s_15_20_6		55-56	"Medium"
s_15_20_6	50-51		"High"
s_15_40_7	53-56	66-68	"High"

Table 3-1 - Categorisation of traffic samples into traffic volumes

The samples were originally categorised into different levels of traffic demand according to the percentage that they were grown by, so the "ungrown" samples were classified as "low" traffic, the samples grown to 20% were classified as "medium", and those grown to 40% were labelled "high". As can be seen from the table above, differences between the baseline traffic samples and a degree of randomness in the sample generation meant that the original split did not reflect the actual aircraft throughput achieved, and so the decision was made to analyse the samples in terms of the actual throughput. Accordingly, since the traffic sample s_15_20_6 did not logically fall into either the 'medium' or 'high' categories, it was necessary to partition the sample into different traffic volumes for the two different sectors.

3.4 CONTROLLERS

A total of thirty two controllers of seven nationalities participated in the PD/1 trials, as discussed in Annex B (Section 4.3).

Each individual controller performed the same role on the same sector in each run from that week. This made possible the comparison of ORGs using *matched pairs* of observations, that is, each controller was exposed to two ORGs at different times under otherwise similar or identical conditions.

The controllers were trained in the use of the tools on a specially designed one week course, entailing a mixture of classroom lessons and hands-on training on the full system. For the majority of the controllers the training took place one week before the trials actually took place. However one set of four controllers were trained prior to participation in the Pilot Phases, which took place in April and September 1995. The day before the measured runs began, all controllers were allowed to undertake three refresher runs.

3.5 TRIAL TIMETABLE

The number and combination of runs that could be conducted during the PD/1 trials was limited by time and controller availability. In each week, Monday was used for refresher training, and during the other four days a total of fourteen simulation runs were carried out, scheduled, for the first four weeks, according to the weekly timetable in Table 3-2. The timetable was rearranged for the second four weeks to minimise the effect of becoming more familiar with the system as the trials progressed.

	Run 1	Run 2	Run 3	Run 4
Monday	Refresher	Refresher	Refresher	
Tuesday	ORG 0, mix 0, s_13_6	ORG 0, mix 0, s_15_40_7	ORG 0, mix 0, s_15_20_6	ORG 0, mix 0, s_14_40_6
Wednesday	ORG 2, mix 30, s_15_40_7	ORG 2, mix 70, s_14_40_7	ORG 2, mix 70, s_13_6	ORG 2, mix 70, s_15_20_6
Thursday	ORG 2, mix 70, s_15_40_7	ORG 2, mix 30, s_14_40_7	ORG 1, mix 0, s_15_40_7	ORG 1, mix 0, s_13_7
Friday	ORG 1, mix 0, s_15_20_6	ORG 1, mix 0, s_14_40_6	Spare	Spare

Table 3-2 - The timetable of runs

The ORG 2 runs scheduled for Run 2 on Wednesday and Thursday involved the BAC 1-11 flying laboratory. The traffic volumes corresponding to each traffic sample can be found in Table 3-1.

4. MEASUREMENTS

4.1 INTRODUCTION

During the trials a large quantity of data was recorded, which may be partitioned into two distinct categories: subjective, such as the ISA and TLX workload measures; and objective, such as the number of R/T calls made.

Data were gathered from various sources including controller questionnaires, controller debriefings, the video recording of the runs, the R/T and telephone systems, and the notes of the observers. The main source of objective information was the PD/1 Ground System, which allowed the automatic recording of "log files" of event data such as aircraft positions, co-ordination offers and even mouse clicks. The raw data in the log files were then processed into useful measures, such as the average number of aircraft in sector.

4.2 SUBJECTIVE MEASURES

The three sources of quantitative, subjective measures in the PD/1 trials were ISA, TLX and the questionnaire responses.

4.2.1 ISA Scores

The instantaneous self-assessment (ISA) workload measurement technique has been developed over a number of years by the NATS ATC Evaluation Unit, now called the NATS ATM Development Centre, based at Hurn. It provides a measure of the controller's perception of his or her workload recorded at regular intervals during the run. Prior to participating in the trial, the controllers were briefed on the ISA assessment technique. Although technically ISA was an intrusive data collection method (i.e. one that requires controllers to interrupt their normal duties), experience gained during the CAER Trial work programme has shown that the controllers do not find that the use of ISA affects their work.

For the PD/1 Trial, DLR manufactured an "ISA box", which was designed to cue the controller to enter his ISA score every two minutes. As soon as the controllers noticed the prompt, they responded by pressing a button on the box to indicate how "loaded" they felt. The controller had a choice of five buttons with which to respond, relating to:

1 Under Utilised

2 Relaxed

3 Comfortable

4 High

5 Excessive

The instantaneous self-assessment (ISA) scores were recorded in a separate log file. Various measures summarising the ISA scores recorded by each controller for each run have been

investigated, including the mean[1] ISA score, the peak ISA score, the percentage of ISA scores ≥ 4, and the percentage of ISA scores = 5.

4.2.2 TLX Scores

At the end of each run, the controller's workload was also measured by administering the task load index (TLX) questionnaire. TLX was developed at NASA-AMES research centre in the USA, and has been used extensively for measuring pilots' subjective workload. As TLX was administered at the end of a trial it was a non-intrusive data collection method. The controller was asked to think back over the trial and to rate his or her overall workload on a 20 point scale of low to high for five separate factors:

- mental demands;
- physical demands;
- time pressure;
- effort expended;
- frustration experienced.

The five factors were also grouped into all ten possible pairs, e.g. effort expended versus mental demand, effort expended versus physical demand, etc., and the controllers were asked to circle the element of each pair that they considered to be the most important contributor to workload. These rankings were used to give each factor a weighting, which was multiplied by the individual score for that factor. Finally the weighted scores for each factor were summed to achieve an overall TLX rating of workload for that particular run.

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 approval for the computer assistance tools, subjective data were collected through questionnaires. The questionnaire used in the PD/1 trial was developed specifically for PD/1, and was based on experience gained from initial versions of the questionnaire issued during the pilot phases of PD/1. It addressed issues of:

- the operational aspects, e.g. traffic handling, ATC procedures;
- the overall HMI aspects, e.g. displays, human-computer dialogues and interaction;
- each specific tool and its functions;
- the training and simulation environment;
- the ATC operational concept in general.

[1] Strictly the ISA scores are ordinal data rather than interval scale data. The 'mean' ISA score therefore needs to be treated with caution.

The questionnaire was 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 questionnaire were given as statements; the controller then indicated the extent to which he or 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.

The experience gained from the questionnaires used during the pilot phases of PD/1 led to three separate questionnaires being used during the main trials:

- organisation questionnaire;
- final questionnaire for tactical controllers;
- final questionnaire for planner controllers.

The questionnaires were given to the controllers as per the schedule in Table 4-1 for the first four weeks. This was then re-arranged to suit the order of the runs for the second four weeks.

Questionnaire	When Given
ORG 0	After Run 4 on Tuesday
ORG 1	After Run 2 on Thursday
ORG 2	After Run 4 on Thursday
Final questionnaire	After Run 2 on Friday

Table 4-1 - Timetable for questionnaires

The reasons for the separate questionnaires are that the majority of items tended to capture more general PD/1 issues, irrespective of particular ORGs, and that many items were relevant to only one controller role. The ORG questionnaires were kept relatively short and relevant only to that ORG, thus keeping the repetitive part of the inquiry as limited as possible. The statistical analysis and results of the questionnaires are given in Annex D.

4.3 OBJECTIVE MEASURES

The following are the objective measures that were considered during the analysis of results:

- number of ATC instructions issued per aircraft;
- number and duration of R/T communications per tactical controller per run;
- number and duration of telephone communications per planner controller per run;
- number and duration of infringements of the minimum separation standards;
- number and duration of short term conflict alerts (STCAs);
- percentage of time spent per aircraft at the cruising flight level;
- average time spent by aircraft in sector.

5. ANALYSIS METHODOLOGY

5.1 INTRODUCTION

The nature of the principal measures used, ISA and TLX, meant that non-parametric statistical tests were the main analysis tool, although parametric tests were used where appropriate. Early analysis indicated that different controllers responded differently to the various ORGs and traffic demand. Thus, averaging results across all the controllers combined would have introduced significant extra variability into any comparisons. Instead, pairs of matched observations from the same controller and traffic sample were used to compare the ORGs, wherever this was feasible.

It should be noted that although data for the whole simulation run were recorded, for analysis purposes the first 15 minutes from each run were ignored. This allowed the airspace, which began empty, to reach the intended level of traffic.

Workload, sector capacity and quality of service to airlines were investigated in order to meet the aims of PD/1, as described in Section 2.

5.1.1 Workload Investigation

The mean ISA scores per run of the different ORGs were compared across both sectors together, both controller roles and across all three traffic volumes. Those results that were statistically significant at the 5% level of significance were investigated more deeply, by repeating the comparisons taking each controller role separately and each sector separately[2]. Finally the ORGs were compared within each individual sector, role, and level of traffic demand.

The alternatives to the mean ISA statistic, namely the peak ISA score per run, the percentage of ISA scores ≥ 4 per run, and the percentage of ISA scores $= 5$ per run, were also investigated.

The subjective TLX responses were also analysed for each controller role, looking at the overall score and at each TLX factor individually.

Various objective measures of workload were examined, for example the number of ATC instructions issued per run, in order to back-up - and investigate the reasons for - the results obtained from the subjective workload analysis. Objective measures of workload included:

- number of ATC instructions issued per aircraft;
- number and duration of R/T communications per tactical controller per run;
- number and duration of telephone communications per planner controller per run.

[2] This procedure of investigating further only if the overall results differed could hide some significant results. For instance, were sector 10 to give a positive difference but sector 11 a negative difference then the overall difference of the sectors combined could be minimal.

5.1.2 Capacity Investigation

The following objective measures were adopted as indicators of the airspace capacity change caused by each ORG:

- number and duration of infringements of minimum separation standards per run;
- number and duration of short term conflict alerts (STCAs) per run.

The approach taken to the analysis of the measures was similar to that for the workload measures. First, the capacity measures of each ORG within the same traffic sample and controller were compared. As was the case in the workload investigation, those results that proved to be significant were investigated further, so as to ascertain whether they were significant in general or just caused by some 'outstanding' result in one particular sector and traffic volume.

5.1.3 Quality Of Service Investigation

The quality of service provided to the airlines 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 airlines by each ORG:

- percentage of time spent at requested cruise flight level;
- the average time spent by aircraft in the sector.

The approach to analysing the quality of service measures was similar to that taken for analysing the workload measures: the ORGs were compared in the most general way, across all traffic volumes and sectors. Those differences between ORGs which proved to be statistically significant at the 5% level were investigated deeper, taking sectors then levels of traffic demand separately, in order to ensure that the observed differences did not just apply to one single traffic volume and sector which happened to be very extreme, but actually applied across all sectors and traffic volumes.

5.2 STATISTICAL TESTS USED

The statistical tests employed were normally non-parametric tests, although parametric tests were used where appropriate. The level of statistical significance chosen for each test was 5%. Although the more powerful parametric tests were considered for use, in general the data did not meet all the assumptions required for use of such tests. Therefore it was decided to opt for the broader and more reliable results given by non-parametric tests. The non-parametric tests that were used are:

- Wilcoxon Signed Ranks test;
- Wilcoxon-Mann-Whitney Ranks Sum test;
- Kruskal-Wallis One-Way Analysis Of Variance.

Wherever possible, the Wilcoxon Signed Ranks test was used to compare measures between ORGs, using matched pairs each coming from the same controller and traffic sample. In this way, the effects of variability between controllers was minimised.

One parametric test that was used was a four-way analysis of variance (ANOVA), carried out on mean ISA scores across controller roles, sectors, traffic volumes and ORGs. Only weeks 3, 7 and 8 contained sufficiently complete sets of runs (see Section 5.5) to create a set of data in factorial form, and for the purposes of this test it was assumed that ISA scores are equally spaced points on an interval scale and that the 'errors' (in the statistical sense) were normally distributed with constant variance.

5.3 HYPOTHESES

For each of the three investigations, i.e. comparisons between ORGs of workload, airspace capacity, and of quality of service, the 'null hypothesis' used in the statistical tests was that there was no difference.

5.3.1 Workload Hypotheses

The following set of null hypotheses was tested, using data pooled across both controller roles, both sectors and all three traffic volumes:

- the workload of ORG 0 is the same as that of ORG 1 (H_{0-1});
- the workload of ORG 0 is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-2});
- the workload of ORG 1 is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-3});
- the workload of ORG 2 with 30% of the aircraft 4-D FMS and datalink-equipped is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-4}).

These hypotheses were tested using the measures of workload described in Section 5.1.1.

In addition to examining the effect of the PATs and the introduction of 4-D FMS and datalink-equipped aircraft over the complete airspace studied in PD/1, their effect on the workload of the individual sectors and on the specific controller roles was examined. The following null hypotheses were tested for selected measures:

- the workload of the tactical controller is the same as that of the planner controller (H_{0-5});
- the workload in sector 10 is the same as that in sector 11 (H_{0-6}).

These two hypotheses were tested using the Wilcoxon-Mann-Whitney test. The Wilcoxon Signed Ranks test did not apply because the comparison was between controllers rather than matched pairs of the response of each controller to differing scenarios.

5.3.2 Capacity Hypotheses

The following set of hypotheses was tested, using data pooled across both controller roles, both sectors and all three traffic volumes:

- the capacity of ORG 0 is the same as that of ORG 1 (H_{0-1});
- the capacity of ORG 0 is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-2});

- the capacity of ORG 1 is the same as that of ORG 2 with 70% aircraft 4-D FMS and datalink-equipped (H_{0-3});
- the capacity of ORG 2 with 30% of the aircraft 4-D FMS and datalink-equipped is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-4}).

The tests were carried out using the objective indicators of sector capacity as listed in Section 5.1.2.

5.3.3 Quality Of Service Hypotheses

The following set of hypotheses was tested, using data pooled across both controller roles, both sectors and all three traffic volumes:

- the quality of service of ORG 0 is the same as that of ORG 1 (H_{0-1});
- the quality of service of ORG 0 is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-2});
- the quality of service of ORG 1 is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-3});
- the quality of service of ORG 2 with 30% of the aircraft 4-D FMS and datalink-equipped is the same as that of ORG 2 with 70% of the aircraft 4-D FMS and datalink-equipped (H_{0-4}).

The tests were carried out using the measures of quality of service to airlines as described in Section 5.1.3.

5.4 MISSING ISA SCORES

Either due to problems with the ISA boxes for recording the ISA scores, or due to some controllers not noticing the ISA prompt for some reason, there are examples of missing ISA scores. In the analysis of the ISA measures of workload reported in Annex C, these gaps were ignored. To examine the robustness of this approach, it was decided to fill in the gaps by linear interpolation on several runs having few gaps, and then check whether this made any difference to the ISA workload conclusions.

5.5 CRITERIA FOR ACCEPTABLE RUNS

During the trials, there were problems from time to time with the equipment and system software which meant that either no data were collected for a particular run, a run was aborted early, or a run did not take place at all. Out of the scheduled 112 runs, 98 yielded usable data. The table below gives details of the runs that did not, for some reason, yield usable data:

Reason	Week	Traffic Sample	ORG
Spoiled runs	4	s_14_40_7	2 (70%)
	4	s_15_20_6	2 (70%)
	7	s_13_7	1
Datalink problems	1	s_14_40_7	2 (70%)
Short run (less than 40 minutes)	4	s_14_40_6	0
Runs not done due to system problems	1	s_14_40_6	0
	2	s_13_7	1
	3	s_14_40_6	0
	4	s_13_6	2 (70%)
	5	s_13_6	0
	6	s_13_6	0
	6	s_13_7	1
	6	s_15_20_6	2 (70%)
	8	s_14_40_6	0

Table 5-1 - Unacceptable runs

6. METHOD SUMMARY

The trials were carried out over a period of eight weeks, from the 23 October 1995 until 15 December 1995. A total of thirty two controllers from seven nations participated in the PD/1 trials, four different ones in each of the eight weeks. During each week, each controller assumed a single role, i.e. either planner or tactical, in a single sector. The environment simulated represented en-route airspace circa the year 2000, based on the UK New En-Route Centre (NERC) North Sea sectors 10 and 11.

The PD/1 trial was defined in terms of three "organisations", (ORG 0, ORG 1 and ORG 2), corresponding to: a baseline for comparison; the introduction of computer assistance tools; and the introduction of those tools plus datalink, and three traffic "mixes", (Mix 0, Mix 30 and Mix 70), indicating the percentage of aircraft with datalink capabilities.

Several traffic samples were used in the simulation runs, in order to vary the levels of traffic demand and so examine in further detail the responses of the controllers to the introduction of the tools and datalink. The samples were categorised into "low", "medium" and "high" traffic volumes, corresponding to the aircraft throughput per hour. Within a given level of traffic demand, several samples were used to ensure the generality of the results and to prevent learning effects.

Null hypotheses were formulated in order to compare the ORGs and mixes with respect to controller workload, airspace capacity and quality of service provided to airlines. By testing these hypotheses the objectives of the PD/1 trials were met.

The null hypotheses were tested using both objective and subjective 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 questionnaires, controller debriefings, and the R/T and telephone systems.

The statistical tests employed were non-parametric tests in general, although parametric tests were used where appropriate. Wherever possible, the Wilcoxon Signed Ranks test was used to compare measures between ORGs, using matched pairs each coming from the same controller and traffic sample. In this way, the effects of any variability between controllers was minimised.

System software and equipment problems during the trials meant that from time to time either no data was collected for a particular run, a run was aborted early, or a run did not take place at all. Criteria for usable runs were created, and only data from runs meeting those criteria were considered in the final analysis.

7. ABBREVIATIONS

ANOVA	analysis of variance
ATC	air traffic control
ATM	air traffic management
CAER	NATS' Computer Assistance for En-Route programme
DLR	Deutsche Forschungsanstalt fur Luft- und Raumfahrt
DRA	Defence Research Agency
FMS	flight management system
ISA	instantaneous self assessment
NATS	National Air Traffic Services Ltd (UK)
NERC	UK New En-Route Centre
ODID	Operational Display and Input Devices
ORG	organisation
PATs	PHARE advanced tools
PD/1	PHARE demonstration 1
PHARE	Programme for Harmonised ATM Research in Eurocontrol
R/T	radiotelephony
STCA	short term conflict alert
TLX	NASA - task load index