

Comparison of EACAC and AGIE metrics using CARE/ASAS #2 framework

CEAMCA Project

Action Plan 5

DOCUMENT REVIEW

Version	Date	Description of evolution	Modifications
0.1	08/04/2002	Initial report structure and objectives according to PMP	All
0.2	07/05/2002	Framework for the review of AGIE and EACAC experiments (step 1)	Section 2, 3
0.3	26/06/2002	Update following comments from Kick-off meeting held on the 14 th of may 2002, Analysis of AGIE (step 2) and EACAC (step 3a) experiments	Sections 1 and 2 Sections 3 and 4
0.4	02/07/2002	Update following comments from progress meeting held on the 26 th of June 2002, Preliminary analysis of INTEGRA metrics on EACAC (step 3b)	All (revised structure of the report)
0.5	14/08/02	Detailed review of INTEGRA metrics (step 3b), Comparative analysis of AGIE and EACAC experiments and metrics, and relationship with INTEGRA metrics (step 4)	All
0.6	06/09/02	Consolidation and amendments (following internal review) of the comparative analysis between AGIE and EACAC experiments and metrics	All
1.0	20/09/02	Update following comments from final meeting held on the 12 th of September 2002	All (mainly, new section 3 and 8)

Authorised by : Thierry Arino on 20-09-2002

CONSORTIUM

Representatives	Organisations
Thierry Arino, Béatrice Raynaud	SOFREA VIA (Project leader)
Nathalie De Beler	EUROCONTROL Experimental Centre (EEC)
Christian Aveneau	Centre d'Etudes de la Navigation Aérienne (CENA)

Executive summary

Study overview

The CEAMCA study has consisted in the comparison of the AGIE and EACAC'2000 experiments and metrics. In addition, the study has also investigated the INTEGRA metrics applied on EACAC. This comparison was based on initial work performed within CARE/ASAS #2, as well as inputs from the MAEVA Validation Guideline Handbook and the VALSUP Guidelines.

The study comes within the scope of the Eurocontrol/FAA Action Plan 5 in charge of developing a validation and verification strategy. The main objective was to help the Agency in progressing with the Action Plan 5 activities related to metrics comparison by producing material on a concrete subject.

The CEAMCA report is expected to help in getting a common understanding of US and European metrics, their context of use as well as their relevance. To support such understanding, the study provides detailed comparison of the experiments and draws the main lessons learnt on various elements of comparison, which respectively deals with:

- The validation approaches, the objectives and the technique used; and
- The metrics framework, their characteristics and their relevance.

The work has been carried out in a co-operative fashion using a phased approach in two major steps: analysis and synthesis. The activity has been realised by a consortium of three organisations: EEC (Eurocontrol Experimental Centre), CENA and Sofréavia (project leader).

First level of comparison

The comparison between AGIE and EACAC provides a good illustration of the relationship that exists between concept development strategy, stage in life cycle and validation approach and objectives. In this respect, both experiments adopted different concept development strategies, which impacted their respective validation approach.

The comparison between AGIE and EACAC validation objectives also provides a good opportunity to look at the dilemma that often exists in early stages of development between “external” validation of expected benefits of the concept and the “internal” validation of the new ATM system implied by the concept.

The adequacy of human-in-the loop experiments in early stages of concept development and validation is also discussed and some guidelines are developed. These guidelines support the definition and evaluation of metrics used during experiments in relationship with explicit validation objectives.

Second level of comparison

A detailed comparison of the metrics identified through the review of AGIE, EACAC as well as the INTEGRA study applied on EACAC, is made at different levels including:

- The characteristics of the metrics, including their different attributes, their means of calculation and decision criteria to conclude on whether or not the objective is met;
- The metrics scope and perspective with distinction between those related to ATM system outcomes (or outputs) and those related to specific ATM system components; and finally
- The different metrics related to common performance areas (either “internal” performance factors related to ATM system components or “external” ATM performance factors).

The purpose is to establish the relationship between the metrics used in both experiments to assess various aspects, to discuss the main differences and similarities between the various metrics, as well as to provide judgement about their relevance.

The range of performance areas addressed through the comparison of AGIE, EACAC and INTEGRA metrics span human performance (and acceptance), as well as efficiency and safety of the new ATM system.

Conclusions and future work

The comparison between AGIE and EACAC experiments allowed highlighting some general guidelines for experiment design in line with MAEVA and VALSUP, together with elements for consideration when defining the validation objectives and metrics for human-in-the-loop experiments.

These considerations constitute a first step towards the definition of a common framework for metrics definition and evaluation. Due to the limited scope of the study and resource constraints, further work would be required to consolidate the set of relevant performance areas and metrics measurable through experiments.

Table of Content

GLOSSARY	X
ACRONYM LIST.....	XI
REFERENCE LIST	XIII
1. INTRODUCTION	1
1.1. SCOPE AND OBJECTIVES	1
1.1.1. Scope of the study	1
1.1.2. Inputs of the study	1
1.1.3. Purpose of the study	1
1.2. BACKGROUND AND CONTEXT	2
1.2.1. FAA/EUROCONTROL Action Plan 5	2
1.2.2. CARE-ASAS Activity 2.....	2
1.2.3. CARE-INTEGRA	4
1.3. DOCUMENT OVERVIEW	5
2. APPROACH OVERVIEW	6
3. FRAMEWORK FOR THE EXPERIMENTS REVIEW	8
3.1. OUTLINE OF THE CONCEPT UNDER VALIDATION	8
3.1.1. Concept studied and its rational	8
3.1.2. Stage in the life cycle	9
3.2. OUTLINE OF THE VALIDATION APPROACH.....	10
3.3. EXPERIMENTATION OF THE CONCEPT.....	12
3.3.1. Experiment characteristics	12
3.3.2. Indicators, Metrics and measurements	14
4. COMPARISON OF OPERATIONAL CONCEPTS UNDER ASSESSMENT WITHIN AGIE AND EACAC.....	17
4.1. OUTLINE OF CONCEPTS UNDER VALIDATION	17
4.2. COMPARATIVE ANALYSIS OF CONCEPTS UNDER VALIDATION	18
4.2.1. Maturity of the concepts.....	18
4.2.2. Concept development strategies.....	18
4.2.3. Level of implication of the concepts on the ATM system	19
5. COMPARISON OF AGIE AND EACAC VALIDATION APPROACHES	20
5.1. OUTLINE OF AGIE AND EACAC VALIDATION APPROACHES	20
5.2. COMPARATIVE ANALYSIS OF BOTH VALIDATION APPROACHES	21
5.2.1. Validation strategies of both experiments.....	21
5.2.2. Human-in-the-loop experiments.....	22

6. COMPARISON OF AGIE AND EACAC’2000 EXPERIMENTS.. 23

6.1. OUTLINE OF AGIE AND EACAC EXPERIMENTS SET-UP AND PERFORMANCE	23
6.2. COMPARISON BETWEEN EXPERIMENT OBJECTIVES AND HYPOTHESES	25
6.2.1. Outline of AGIE and EACAC experiments objectives and hypotheses	25
6.2.2. Comparative analysis of both experiments objectives and hypotheses	27
6.2.2.1. Hypotheses and objectives focus on human factors	28
6.2.2.2. Explicit or implicit relation between hypotheses and objectives	28

7. COMPARISON BETWEEN AGIE, EACAC AND INTEGRA METRICS..... 29

7.1. COMPARISON BETWEEN METRICS CHARACTERISTICS	29
7.1.1. Outline of AGIE, EACAC and INTEGRA metrics characteristics	29
7.1.2. Comparative analysis of experiments metrics and measurements	31
7.1.2.1. Balance between subjective and objective metrics within AGIE and EACAC	31
7.1.2.2. Data collection and analysis methods within AGIE and EACAC	31
7.1.2.3. INTEGRA metrics on EACAC’2000	32
7.2. COMPARISON BETWEEN METRICS SCOPE AND PERSPECTIVE	33
7.2.1. Outline of AGIE, EACAC and INTEGRA metrics	33
7.2.2. Comparative analysis of metrics scope and perspective	35
7.2.2.1. Balance between “external” and “internal” performance metrics	35
7.2.2.2. Metrics scope and perspectives	36
7.3. COMPARISON BETWEEN AGIE, EACAC AND INTEGRA METRICS RELATED TO HUMAN PERFORMANCE	38
7.3.1. Comparative analysis of AGIE and EACAC controller workload metrics	38
7.3.2. Comparative analysis of AGIE and EACAC controller situational awareness metrics	41
7.3.3. Comparative analysis of AGIE and EACAC controller activity metrics	41
7.3.4. Comparative analysis of EACAC and INTEGRA capacity metrics	43
7.4. COMPARISON BETWEEN AGIE AND EACAC EFFICIENCY METRICS	45
7.4.1. Analysis of AGIE control efficiency metrics	45
7.4.2. Comparative analysis of flight efficiency metrics	46
7.5. COMPARISON BETWEEN AGIE, EACAC AND INTEGRA SAFETY METRICS	47
7.5.1. Comparative analysis of AGIE and EACAC safety metrics	47
7.5.2. Comparative analysis of EACAC and INTEGRA safety metrics	48

8. CONSOLIDATION OF THE AGIE AND EACAC COMPARISON

8.1. VALIDATION APPROACHES, OBJECTIVES AND TECHNIQUE	50
8.1.1. Relationship between concept development and validation	50
8.1.1.1. Outline of the comparison	50
8.1.1.2. Consolidation and lessons learnt	51
8.1.2. Balance between “Internal” and “external” validation	51
8.1.2.1. Outline of the comparison	52

8.1.2.2. Consolidation and lessons learnt	52
8.1.3. Human-in-the loop experiments in the validation process	53
8.1.3.1. Outline of the comparison	53
8.1.3.2. Consolidation and lessons learnt	53
8.1.4. Human-in-the loop experiment design	55
8.1.4.1. Outline of the comparison	55
8.1.4.2. Consolidation and lessons learnt	55
8.1.4.3. Guidelines for experiment design	57
8.2. METRICS FRAMEWORK, CHARACTERISTICS AND RELEVANCE.....	59
8.2.1. Balance between “internal” and “external” performance metrics	59
8.2.1.1. Outline of the comparison	59
8.2.1.2. Consolidation and lessons learnt	61
8.2.2. Metrics (and validation objectives) related to human performance	63
8.2.2.1. Outline of the comparison	63
8.2.2.2. Consolidation and lessons learnt	64
8.2.3. Metrics (and validation objectives) related to efficiency	65
8.2.3.1. Outline of the comparison	65
8.2.3.2. Consolidation and lessons learnt	65
8.2.4. Metrics (and validation objectives) related to safety	66
8.2.4.1. Outline of the comparison	66
8.2.4.2. Consolidation and lessons learnt	66

9. CONCLUSIONS..... 67

10. FUTURE WORK 68

ANNEX A Analysis of the Air-Ground Integration Experiment....A-I

A.1 Outline of the concept under validation	A-I
A.1.1 Concept studied and its rational	A-I
A.1.2 Stage in the life cycle	A-II
A.2 Outline of the validation approach	A-III
A.2.1 Stage in validation (possibly per validation objectives).....	A-III
A.2.2 Validation objective(s)	A-III
A.2.3 Validation point of view	A-III
A.2.4 Object(s) under validation	A-III
A.2.5 Validation activity (or technique as named in MAEVA)	A-III
A.3 Experimentation of the concept.....	A-IV
A.3.1 Experiment characteristics.....	A-IV
A.3.2 Indicators, Metrics and measurements	A-IX

ANNEX B Analysis of the EACAC’2000 experimentsB-I

B.1 Outline of the concept under validation	B-I
B.1.1 Concept studied and its rational	B-I
B.1.2 Stage in the life cycle	B-II
B.2 Outline of the validation approach	B-II
B.2.1 Stage in validation (possibly per validation objectives).....	B-II
B.2.2 Validation objective(s)	B-II
B.2.3 Validation point of view	B-III
B.2.4 Object(s) under validation	B-III
B.2.5 Validation activity (or technique as named in MAEVA)	B-III

B.3	Experimentation of the concept.....	B-III
B.3.1	Experiment characteristics.....	B-III
B.3.2	Indicators, Metrics and measurements	B-VII

ANNEX C Analysis of INTEGRA metrics on EACAC’2000 experiments.....C-I

C.1	Scope of the study	C-I
C.2	INTEGRA metrics applied in EACAC	C-I
C.2.1	Metrics related to capacity.....	C-II
C.2.2	Metrics related to safety	C-IV

List of figures

Figure 1: ATM system Performance Indicators versus Sub-System Performance Metrics (CARE-ASAS)	3
Figure 2: Operational benefits and expectations from operational concept	9
Figure 3: Outline of validation activities (or techniques in MAEVA).....	12
Figure 4: Relationship between metrics and validation objectives (MAEVA).....	13
Figure 5: Elements of the ATM system under experimentation	15
Figure 6: Relationship between concept development and validation	51
Figure 7: Development and validation of a new operational concept.....	52
Figure 8: Role of human-in-the-loop experiment in validation.....	54
Figure 9: Model of the ATM system related to a new operational concept	56
Figure 10: Model of human performance in relationship with a new concept.....	57
Figure 11: Number of AGIE and EACAC metrics per validation objectives	60
Figure 12: Scope of metrics measurable within human-in-the-loop experiments	61
Figure 13: Illustration of the relationship between “external” and “internal” performance criteria	62
Figure 14: EACAC complementary levels of controller’s activity analysis	B-XI

List of Tables

Table 1: Relationship between development and validation phases (VALSUP)	10
Table 2: Outline of the validation approach for experiment under review.....	12
Table 3: Outline of the experiment under review.....	13
Table 4: Description of each objective of the experiment under review.....	14
Table 5: Description of each metric from the experiment under review.....	16
Table 6: Outline of concepts under validation within AGIE and EACAC	17
Table 7: Outline of validation approaches within AGIE and EACAC.....	20
Table 8: Outline of AGIE and EACAC experiments characteristics	24
Table 9: Outline of AGIE and EACAC experiment objectives and hypotheses	26
Table 10: Outline of AGIE and EACAC metrics and measurements	30
Table 11: Outline of AGIE, EACAC and INTEGRA metrics	34
Table 12: Outline of AGIE and EACAC controller workload metrics	39
Table 13: Outline of AGIE and EACAC controller activity metrics	42
Table 14: Relationship between EACAC and INTEGRA capacity metrics.....	44
Table 15: Outline of AGIE control efficiency metrics.....	45
Table 16: Outline of AGIE and EACAC flight efficiency metrics	46
Table 17: Outline of AGIE and EACAC safety metrics	47
Table 18: AGIE Experimental Condition Characteristics Summary.....	A-II
Table 19: Outline of the Air-Ground Integration Experiment.....	A-VII
Table 20: Description of each experiment objective of AGIE	A-IX
Table 21: EACAC Operational Procedures Summary	B-I
Table 22: Outline of the EACAC experiment	B-VI
Table 23: Description of each objective of the EACAC'2000 experiments	B-VII

Glossary¹

Validation aim	Clear and unambiguous definition of what is to be achieved through the conduct of a validation exercise. In the context of ATM validation, to provide information that demonstrates the feasibility of an ATM operational concept and that the concept provides a solution to the specific ATM problem it has been designed to address.
High-level validation objective	Expression of the validation aim in terms of measurable factors . Measurable factors may be performance criteria (e.g. safety, capacity, and economics), as well as human and social factors such as acceptability, confidence, and skills required.
Low-level validation objective	Decomposition of high-level objective into lowest level objectives that are measurable and related to elementary ATM items. This is when feasible metrics and indicators can be identified. In this document, high-level objectives are derived in hypotheses and then, metrics are identified.
External validation	Validation of the operational concept regarding its expected benefits and constraints (i.e., performance of the operational concept).
Internal validation	Validation of the design of the ATM system associated with the operational concept (e.g. appropriate information, procedures and decision support tools)
Indicator	An indicator is a metric that is only indirectly related to the objective(s) of interest. [5]
Metric	A metric is a parameter that can either be measured directly, or be calculated from several measurements, and that expresses a significant quality of a system. [5]

¹ Unless otherwise associated with a reference, the following definitions have been developed for the study purposes.

Acronym list

ADS-B	Automatic Dependent Surveillance - Broadcast
AGIE	Air-Ground Integration Experiment
ANOVA	ANalysis Of VAriance
AOC	Airline Operations Center
API	Aircraft Proximity Index
ASAS	Airborne Separation Assistance System
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
ATS	Air Traffic Service
ATSP	Air Traffic Service Provider
CARE	Co-operative Actions of Research and development in EUROCONTROL
CDTI	Cockpit Display of Traffic Information
CEAMCA	Comparison of EACAC and AGIE Metrics using CARE/ASAS #2 framework
CENA	Centre d'études de la Navigation Aérienne
CWP	Controller Working Position
DST	Decision Support Tool
EACAC	Evolutionary Air-ground Co-operative ATM Concepts
EACAC'2000	EACAC study in the year 2000
EEC	EUROCONTROL Experimental Centre
EMERALD	EMERging RTD Activities of relevance for ATM concept Definition
EO	Expert Observer
FAA	Federal Aviation Administration
FD	Flight Deck
HSD	Honestly Significance Difference
IAF	Initial Approach Fix
INTEGRA	Advanced ATM Tool Integration Project
ISA	Instantaneous Self-Assessment
IPL	Information Processing Load
MAEVA	A Master ATM European Validation plan
N/A	Not Applicable
NASA	National Aeronautics and Space Administration

NASA ARC	NASA Ames Research Center
PC	Planning Controller
RTD	Research and Technical Development
SD	Standard Deviation
SEM	Standard Error of Measurement
Sofréavia	Société Française d'Etudes et Réalisations d'Equipements Aéronautiques
TC	Tactical controller
TLX	Task Load Index
URET	User Request and Evaluation Tool
VALSUP	VALidation SUPport
WAK	Workload Assessment Keypad
WJHTC	FAA William J. Hughes Technical Center

Reference list

- [1] 'Air-Ground Integration Experiment', DOT/FAA/CT-TN02/06, January 2002
- [2] 'EACAC 2000 Real-Time Experiments', EEC Report version 5.4, October 2001
- [3] 'INTEGRA Metrics in EACAC', Final Report – EATMP edition 0.A, February 2001
- [4] 'Towards a validation framework for ASAS applications', CARE-ASAS Activity 2, version 1.0, June 2001
- [5] 'Validation Guideline Handbook', MAEVA project, MVA/ISD/WP1/13DI, version 1.1, November 2001
- [6] 'Validation guideline', VALSUP project, version 1.2, April 2002
- [7] 'Research and Technical Development (RTD) Plan for ASAS Concept Development', EMERALD/WP5/SOF/014/3.0 – WP5.5 Report, March 1998
- [8] 'Using simulation to evaluate the safety of proposed ATC operations and Procedures', Paul, L. (1990), DOT/FAA/CT-TN90/22, Atlantic City, NJ
- [9] 'Operational Concept Validation Strategy Document', FAA/EUROCONTROL MoC on R&D: Action Plan 5, Version 1.0, November 2001
- [10] 'INTEGRA Metrics and Methodologies, Eurocontrol, Version 1.1, 13 September 1999
- [11] 'Detailed Specification of ATM System Capacity Metric Inputs Processing and Outputs', Eurocontrol, Version 1.0, August 2000
- [12] 'Detailed Specification of ATM System Safety Metric Inputs Processing and Outputs', Eurocontrol, Version 0.B, November 2000
- [13] 'Detailed Specification of ATM System Efficiency Metric Inputs Processing and Outputs', Eurocontrol, Version 3.1, February 2001
- [14] 'Detailed Specification of Environmental Impact Metric Inputs Processing and Outputs', Eurocontrol, Version 0.A, September 2000
- [15] 'INTEGRA Metrics in EACAC Validation Report', Eurocontrol, Version 0.A, February 2001

1. INTRODUCTION

1.1. Scope and objectives

1.1.1. Scope of the study

The study has consisted in the comparison of the EACAC (Evolutionary Air-ground Co-operative ATM Concepts) 2000 experiments and the AGIE (Air-Ground Integration Experiment) using the CARE/ASAS #2 framework, i.e., the CEAMCA Project.

CEAMCA stands for **C**omparison of **E**ACAC and **A**GIE **M**etrics using **C**ARE/**A**SAS #2 framework.

The EACAC'2000 human-in-the-loop experiments [2] and the AGIE experiment [1] were focused on human factor assessment. In both cases, the objective was to assess the operational acceptance of the concept, investigating the impact on practices and on safety. In EACAC'2000, only one operational concept (i.e., limited delegation with initiative remaining under controllers' responsibility) is investigated, while, in AGIE, three different levels of shared responsibility are compared to a baseline situation. In addition to the metrics used in the EACAC experiments, the study has also investigated the INTEGRA experience, which aimed at applying the INTEGRA metrics in EACAC [3].

The work has been carried out in a co-operative fashion using a phased approach in two major steps (analysis and synthesis). The activity has been realised by a consortium of three organisations: EEC (Eurocontrol Experimental Centre), CENA and Sofréavia (project leader).

1.1.2. Inputs of the study

The study used, as its foundation, the CARE/ASAS #2 report ("Towards a validation framework for ASAS applications") [4] and the ASAS metrics taxonomy that was established by the CARE/ASAS #2 collaborative Activity. In addition, the Project benefited from the CARE/INTEGRA outputs.

The MAEVA and VALSUP projects have also been considered since they both provide guidance for validation. MAEVA provides guidance [5] for validation process, based on a top-down approach starting from the identification of the validation objectives. VALSUP is dedicated to the programme and project managers. It addresses the EATMP programmes and projects that aim at specifying the new European ATM concepts. The VALSUP guideline [6] is intended to support the managers to understand what validation should do (and not do) and assist them in the formation of a validation plan.

1.1.3. Purpose of the study

The CEAMCA study comes within the scope of the Eurocontrol/FAA Action Plan 5 in charge of developing a validation and verification strategy. The main objective was to help the Agency in progressing with the Action Plan 5 activities related to metrics comparison by producing material on a concrete subject (i.e., the comparison of metrics identified in Projects conducted by Eurocontrol & FAA and related to ASAS as a case-study).

The purpose of the CEAMCA study was in particular to:

- Analyse and compare the approaches used in both AGIE/EACAC experiments;
- Discuss their main differences and/or similarities;
- Analyse and compare the measurements, metrics and forms used in both experiments;
and
- Identify the possible relationships between those US and European metrics.

The CEAMCA report is expected to help in getting a common understanding of US and European metrics, their context of use as well as their relevance.

1.2. Background and context

1.2.1. FAA/EUROCONTROL Action Plan 5

The Eurocontrol/FAA Action Plan 5 has been tasked to determine a strategy for validating and verifying the performance, reliability, and safety of ATM systems and its possible relations to certification. The strategy should allow for the validation and verification during the phases of research and development, as well as implementation of airborne and ground-based ATM sub-systems in relation to the current and future operational context.

In this context, an Operational Concept Validation Strategy document [9] has been jointly developed, which addresses this request and establishes a common understanding of the context, purpose, and scope of validation.

To further development such common understanding, there is also an action to present a detailed framework for comparison of metrics.

To begin addressing this action, the CEAMCA Project has compared the metrics used within the FAA/NASA Air-Ground Integration Experiment with those from the EACAC'2000 experiments.

1.2.2. CARE-ASAS Activity 2

As for any new ATM concept, the evolution to a mature ASAS environment is likely to be via a phased implementation. Along the way, compatibility must be assured between current and future systems and procedures. Before ASAS can be realistically implemented, questions remain to be answered in several areas.

In particular, a cornerstone in the initial ASAS development is to allow for comparability of ASAS metrics among different R&D organisations. It is also necessary to link (possibly low-level) metrics performed during ASAS experiments to more general ATM metrics (e.g., safety, capacity and efficiency).

The CARE-ASAS #2 Activity, which is part of the CARE (Co-operative Actions of Research and Development) programme in EUROCONTROL, deals with the identification of a validation framework for ASAS applications:

- To allow for comparability and consolidation of studies related to ASAS;
- To assess the performance and acceptability of the material to be developed in, and results of, the various ASAS studies.

Within CARE-ASAS Activity 2, the scope of the ASAS studies of interest encompasses all the R&D activities aiming at evaluating potential benefits and constraints of ASAS applications from various perspectives: technical, operational or economic, at different stages of development of the ASAS concept.

Initial CARE/ASAS #2 work [4] already includes some comparison and consolidation of metrics used in past ASAS projects in Europe. This report also tries to address the relationship between ASAS metrics and more general ATM metrics proposed in the CARE/INTEGRA Activity, which aims at providing a framework for the quantitative assessment of new ATM systems.

As illustrated in the figure below, it clearly appeared that a gap had to be filled between the work already carried on high-level ATM performance indicators and the low level measures that were done in past European ASAS experiments.

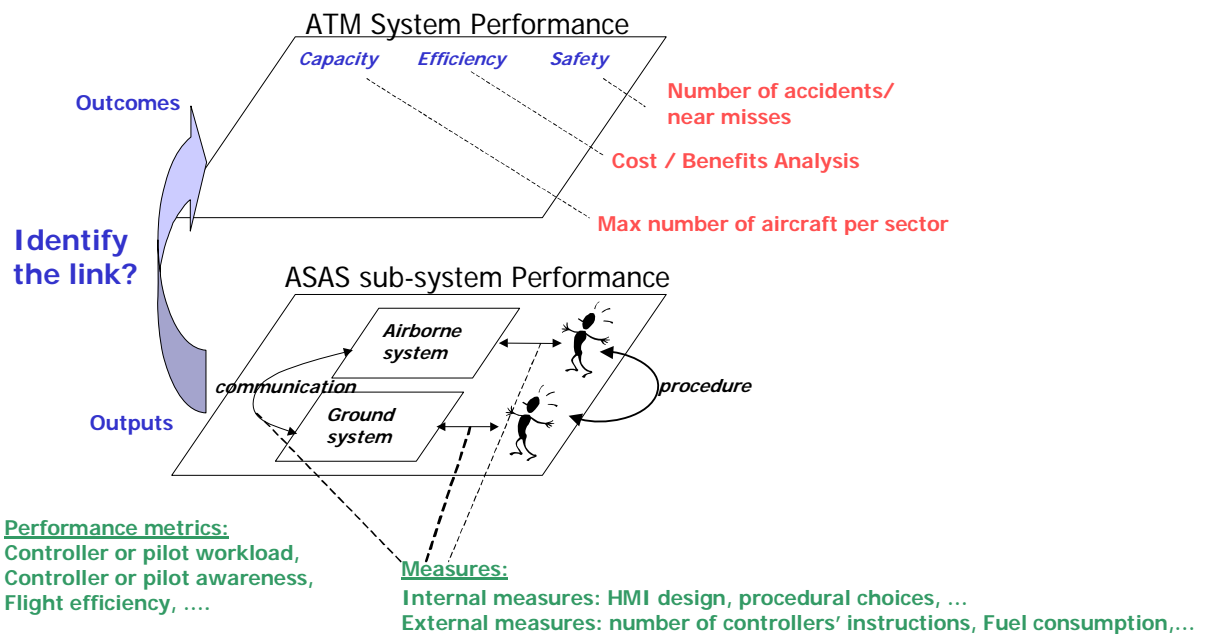


Figure 1: ATM system Performance Indicators versus Sub-System Performance Metrics (CARE-ASAS)

In this perspective, taxonomy was proposed that distinguishes between the measures, the performance metrics related to ASAS applications and the high-level system performance metrics related to the overall performance of ATM system. The necessity to connect the ASAS R&D validation strategy, the overall ATM validation strategy and Human Factors validation strategy was also underlined.

1.2.3. CARE-INTEGRA

INTEGRA is a CARE Action concerning Advanced ATM tool integration. The raison d'être of the INTEGRA Project is to provide a quantified assessment of proposed automated ATM tools and associated procedures within an ATM System.

The benefits of such tools are to be quantified in agreed terms using agreed methodologies against the criteria of: Safety, Capacity, Economy and Environmental Impact [10].

The INTEGRA Project was planned with a phased approach:

- The Initiation Phase was to focus on identification of tools and the associated metrics; platforms and architectures; potential Project partners and Projects with which INTEGRA may have synergy.
- The Definition Phase will be the detailed planning phase of the Project elaborating areas such as the testing scenarios and detailing costs and work-packages. At the end of the Definition Phase formal calls for participation will be issued to all member states, the European Commission and all potential partners. The intention being to put formal agreements in place prior to the Execution Phase.
- The Execution Phase will carry out the trials as planned.

In order to avoid duplication of efforts, the project focused during the initiation phase on the development of the metrics and postponed the definition of tool-sets until the execution phase.

The basic objectives of the INTEGRA metrics were as follows:

- To develop algorithms which would permit quantitative measurements to be derived in four key ATM areas;
- To provide a set of metrics that could be used for real-time and fast-time simulations of novel ATM Systems.
- To provide metrics that were to be non-intrusive to the simulations on which they were to be used.
- To provide metrics that could be used to enable comparisons between different organisations within a simulation and also between different simulations, either run at the same or different establishments

Furthermore, the metrics were to be abstract and independent of airspace design, operating concept and control procedures.

The metrics specification was completed in March 2001 [11], [12], [13], [14]. In addition INTEGRA software was developed to implement the algorithms. In the case of the Capacity and Safety metrics the software was configured and an initial validation exercise was performed using data from the EEC Bretigny EACAC ASAS experiments [15].

Further initial validation activities are underway, which aimed at applying the INTEGRA metrics to already performed simulations or to planned exercises. Depending on the results, the Definition Phase will begin quickly or the metrics will undergo more or less significant changes and revalidation runs.

1.3.Document overview

This **first section** describes the purpose of the document and presents the scope and background of the CEAMCA study.

Section 2 describes the **approach used to perform the comparison** of metrics between AGIE and EACAC experiments, and provides an overview of the successive steps performed during the study

Section 3 describes the general **framework used for the review of metrics** used in of both AGIE and EACAC experiments. The framework does not only support the description of metrics, but also their context of use in the validation process.

Section 4 is focused on the **comparison of the operational concept** under assessment within the AGIE and EACAC'2000 experiments. In particular, differences in terms of concept maturity, complexity and impact on the ATM system are outlined.

Section 5 is focused on the **comparison of the validation approach** used within the AGIE experiment and the EACAC'2000 experiments. In particular, it discusses the similarities of these two real-time simulations with human in the loop, as well as the different validation objectives of both experiments performed at different development stages of the concept under assessment.

Section 6 is focused on the **comparison between AGIE and EACAC'2000 experiments** set-up and performance, as well as the experiments objectives and hypotheses.

Section 7 performs a **comparison of the metrics** used within AGIE and EACAC'2000 experiments, as well as the INTEGRA metrics applied to the EACAC experiments. The metrics comparison is made at different levels: the metrics characteristics, the hierarchy of the metrics, and finally the different metrics related to common performance areas.

Section 8 consolidates the comparison between AGIE and EACAC experiments, and provides a synthetic description of the **relationship between the various metrics** analysed. It also draws main lessons learnt from the comparison of AGIE and EACAC experiments and metrics.

Finally, **sections 9 and 10** provide conclusions and suggestions for future work related to metrics comparison within the framework of Action Plan 5.

The annexes provide more detailed information related to:

- **ANNEX A** the analysis of the metrics used in the AGIE experiment,
- **ANNEX B** the analysis of the EACAC'2000 experiments,
- **ANNEX C** the analysis of the INTEGRA metrics applied to the EACAC'2000 experiments.

2. APPROACH OVERVIEW

The CEAMCA study has consisted in analysing both the AGIE and the EACAC'2000 experiments and performing an initial comparison of the metrics used during the experiments. The study has been conducted through the following steps:

- Step 1: Framework for the experiment review;
- Step 2: AGIE analysis;
- Step 3: EACAC analysis;
 - Step 3a: EACAC metrics;
 - Step 3b: INTEGRA metrics in EACAC;
- Step 4: Comparison and relationships between metrics;
- Step 5: Consolidation of the comparison; and
- Step 6: Conclusions and future work.

The purpose of step 1 was to provide a general framework for an effective review of the validation approach, objectives and metrics used in both experiments. Detailed description of the proposed framework is provided in section 3.

This framework was derived from the CARE-ASAS #2 framework for review of European ASAS study [4], with new inputs from the MAEVA Validation Guideline Handbook [5] and the VALSUP Guidelines [6].

Steps 2 & 3, which have been performed in parallel, were dedicated to the review of both the AGIE and the EACAC experiments. For the EACAC experiments, an additional task (Step 3b) consisted in investigating the INTEGRA experience aiming at applying the INTEGRA metrics in EACAC. Each review included:

- Identification of the **context** of the experiment;
- Identification of the **scope of the assessment** (e.g., operational performances, cost/benefit, safety, human factors, system performances) and validation point of view (e.g., airspace users, ATS providers, flight crew, air traffic controllers); and
- Identification of **metrics and indicators**² used during the experiments with the purpose of those metrics, and possibly, their description as aggregates of more low-level measures.

The detailed analysis of the AGIE experiment, the EACAC'2000 experiments and the INTEGRA study in EACAC are respectively provided in ANNEX A, ANNEX B and ANNEX C.

² In the following of the document, the word "metric" covers both metric and indicator notions.

The purpose of Step 4 was to **compare** the AGIE and the EACAC metrics. This comparison consists in identifying differences and similarities between the approaches and the metrics used to assess various aspects (e.g., ATM outcome metrics, aircraft operation metrics, human factor metrics, system metrics).

The different aspects of the AGIE and EACAC comparison are developed in sections 4, 5, 6 and 7.

The comparative analysis aimed at establishing the relationships between the AGIE and the EACAC experiments metrics, using an approach similar to that performed within CARE/ASAS #2 (when comparing European past ASAS studies). The purpose was also to provide judgement about the relevance of the metrics used, and possibly to identify forgotten metrics.

The Step 5 consisted in achieving a **consolidation of the comparison** performed in Step 4. This consolidation contained in section 8 provides a synthetic description of the relationships between US and Europe metrics and context of use. Particular attention is drawn to the lessons learnt during the performance of the Project.

Finally, the Step 6 consists in developing the **conclusions** of the report. At this stage, the Project is seen as the basis for further discussions about the relevance of the CARE/ASAS taxonomy for ASAS metrics and its possible extension to a framework for metric comparisons within AP5. Therefore, possible areas for future work are also highlighted.

3. FRAMEWORK FOR THE EXPERIMENTS REVIEW

This section describes the general framework used for the review of metrics used in AGIE and EACAC experiments, as well as the INTEGRA study on EACAC.

The framework does not only support the description of metrics, but also their context of use. In particular, it was considered necessary to also compare:

- The operational concept under validation,
- The validation approach and objectives,
- The experiments scope and objectives.

Indeed, these were anticipated to have an impact on the metrics used, and were also considered as essential elements to take into account during the comparison of metrics.

The proposed framework is based on CARE-ASAS #2 framework for review of past European ASAS study [4], with new inputs from the MAEVA Validation Guideline Handbook [5] and the VALSUP Guidelines [6].

In particular, the vocabulary used in this framework is compliant with MAEVA Handbook. And, the relation between validation process and design process, as described in VALSUP, has been considered.

3.1. Outline of the concept under validation

3.1.1. Concept studied and its rationale

- **Concept of operations**

Brief description of the operational purpose and (anticipated) impact on the ATM system.

- **Operating environment characteristics**

Brief description of the ATM environment in which the operational concept is expected to apply including the followings:

- Airspace and traffic characteristics,
- Main CNS assumptions.

- **Expected benefits and constraints (“high-level” objectives (and aims) of validation)**

Brief description of operational improvements expected from the operational concept, in relationship with expectations from (one or many bodies from) the ATM community.

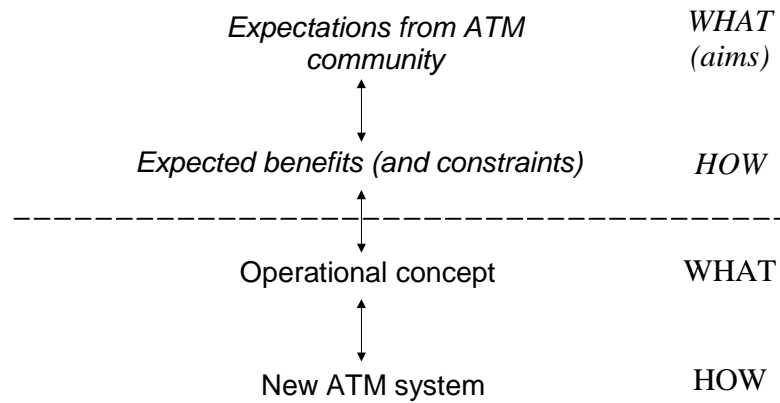


Figure 2: Operational benefits and expectations from operational concept

3.1.2. Stage in the life cycle

The CARE-ASAS #2 framework [4] included considerations from the EMERALD RTD (Research and Technical Development) Plan [7].

The RTD plan follows a model usually applied to the development of complex systems, going from the user requirements to the validation by users but with a recursive approach:

- User Requirement or Concept Phase
- User Requirement Analysis or Feasibility Phase
- Functional Requirement or Acceptability Phase
- ASAS Development or Prototyping Phase
- Pre operational (Experimentation) phase
- Implementation Phase

The implementation phase is not part of the RTD plan, but it is mentioned to highlight what is really necessary to achieve before starting implementation.

The three first phases of the EMERALD RTD correspond to the stage 1 defined in FAA/EUROCONTROL MoC on R&D: Action Plan 5: Operational Concept Validation Strategy Document [9], where 3 stages have been defined for the development of a concept until its implementation:

Stage 1: development of the operational concept specification

Stage 2: system procurement

Stage 3: pre-operational and operational phase.

3.2. Outline of the validation approach

- **Stage in validation (possibly per validation objectives)**

From a research and development perspective, the VALSUP document [6] distinguishes between three validation phases, which correspond to different development phases with different validation objectives:

V1: Basic principles of a new concept are agreed

V2: Initial proof of concept (e.g. through model or early prototype)

V3: Full specifications of concept, (e.g. pre-operational demonstration)

Development phase	User Requirement or Concept Phase	User Requirement Analysis or Feasibility Phase	Functional Requirement or Acceptability Phase
Purposes	<ul style="list-style-type: none"> - Operating Environment description: ATM segment(s) concerned (scope), geographical application of the concept - Functional requirements (services provided by segments) - Actors: Roles and responsibilities (man/machine) - Gap with current situation - Operational scenarios 	<ul style="list-style-type: none"> - ATM System description - Functional specifications, - Technical issues - Actors: tasks allocated to man /machine; procedures; general working method 	<ul style="list-style-type: none"> - Detailed working method, procedures - Human machine inter-action - Technical System and HMI specifications - System Performance required - Training programmes - Transition issues - Selection of technical options
Validation phase	V1: Basic principles of a new concept are agreed.	V2: Initial proof of concept (e.g. through model or early prototype)	V3: Full specifications of concept, (e.g. pre-operational demonstration)

Table 1: Relationship between development and validation phases (VALSUP)

- **Validation objective(s)**

A validation objective is the formulation of the validation aim in terms of measurable factors [5]. Within the 5th FP, one or more of the following eight high-level ATM2000+ strategic objectives provide the basis for identifying the measurable factors in ATM validation exercises:

- Safety
- Economics
- Capacity
- Environment.
- National security and defence requirements
- Uniformity
- Quality
- Human involvement and commitment

Based on the experience gained in the CARE-ASAS #2 review of past ASAS experiments [4], it is essential to distinguish between “internal” validation or “external” validation of the new ATM system associated with the operational concept under development, depending on the validation point of view and the objects under validation.

- **Validation point of view**

As the main objective of validation is to assess the ability to operate against a predefined set of requirements, it is essential to clearly identify the ATM body whose requirements are under assessment. It is proposed to distinguish between the following points of view of validation:

- ATM service users,
- ATM service providers, like ATC units or airports,
- ATM support providers, including ground automation systems providers, aircraft manufacturers or communication service providers,
- ATM players, like regulatory authorities,
- ATM operators, like air traffic controllers or flight crew or others.

- **Object(s) under validation**

Depending on the development stage of the concept under validation, the final operating environment may not necessarily be established, and the validation may be concentrated on the assessment of some elements of the ATM system, instead of the overall ATM system: These elements could be classified as follows:

- ATM environment: like airspace structure or separation minima,
- ATM operations and traffic flows,
- ATM rules and procedures: including nominal or contingency procedures, conditions of applicability/termination, phraseology,
- Aircraft or ATM systems: including communications, functions or HMI,
- ATM system transition phase.

- **Validation activity (or technique as named in MAEVA)**

As the accurateness and operational significance of the metrics may depend on the level of modelling of the object under validation and its environment, the kind of validation activity should be identified, and could be classified as follows:

- Judgmental technique,
- Analytic/statistical modelling,
- Fast-time simulation,
- Small/large scale real-time simulation,
- Operational trials,
- Real-life operations evaluation,
- ...

	Safety	Economics	Capacity	Environment	National security and defence requirements	Uniformity	Quality	Human involvement and commitment
Literature study	✓	✓	✓	✓	✓	✓	✓	✓
Judgemental technique	✓	✓	✓	✓	✓	✓	✓	✓
Fast-time Technique	✓	✓	✓	✓	✓			
<i>Analytical Modelling</i>	✓	✓	✓	✓	✓			
<i>Fast-time simulation</i>	✓	✓	✓	✓	✓			
Real-time Technique	✓	✓	✓	✓	✓	✓		✓
<i>Real-time simulation</i>	✓	✓	✓	✓	✓	✓		✓
<i>Field Test</i>	✓					✓		
<i>Shadow Mode</i>	✓	✓	✓	✓	✓	✓		✓
<i>Operational Trial</i>	✓	✓	✓	✓	✓	✓		✓

Figure 3: Outline of validation activities (or techniques in MAEVA)

The following table summarises the various validation characteristics that need to be identified during the review of the experiments.

Validation objectives	Validation phase	Validation point of view	Object under validation	Validation activity (or technique)

Table 2: Outline of the validation approach for experiment under review

3.3. Experimentation of the concept

Brief description of the experiment, including assumptions, limitations and objectives, in order to later analyse the different metrics, indicators and measurements performed during the experiment for the purpose of validating the operational concept under assessment.

3.3.1. Experiment characteristics

- **Scope of the experiment**

Assumptions/limitations (e.g. simplified ATM environment, focus on some ATM sub-system) of the experimented ATM system and its environment:

- Airspace organisation,
- Traffic scenarios,
- Procedures,
- Tools.

- **Experimental conditions**

Identification of the main characteristics of the experiment set-up and performance including:

- Sessions (including training, briefing/debriefing and simulation sessions),
- Participants,
- Data collection.

- **Experiment objectives (“low level” objectives of validation)**

Outline of the validation objectives for the experiment (either related to “external” validation or “internal” validation of the ATM system) with:

- Objectives (textual) description,
- Hypotheses to be validated (or invalidated) through the experiment,
- Related expected benefits (or constraints), or
- Related ATM system element under validation.

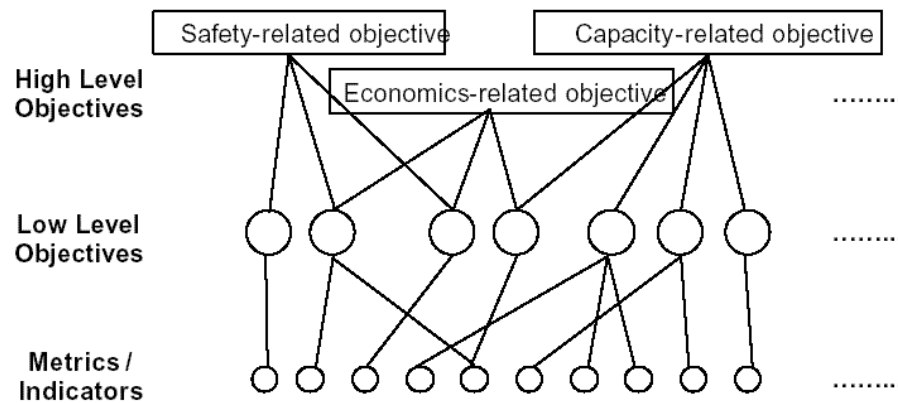


Figure 4: Relationship between metrics and validation objectives (MAEVA)

The following table summarises the various experiment characteristics that need to be identified during the review of the experiments.

Scope of experiment	Experimental conditions	Experiment objectives

Table 3: Outline of the experiment under review

Only, if explicitly stated for the experiment under review or if required to perform the comparison between experiments, detailed description of the experiment objectives with related “external” or “internal” validation objectives has also to be provided using the following table.

Experiment objective		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation

Table 4: Description of each objective of the experiment under review

3.3.2. Indicators, Metrics and measurements

“A **metric** is a parameter that can either be measured directly, or be calculated from several **measurements**, and that expresses a significant quality of a system. A ‘system’, in the context of ATM, may also include the human operators. An **indicator** is a metric that is only indirectly related to the objective(s) of interest. Consequently a change in the value of an indicator does not necessarily lead to a definite conclusion that the related objective has been achieved.” [5]

- **Metrics classification (if possible)**

Metrics can either be classified according to their related validation objectives or to the element of the ATM system on which they fall on.

Considering the experiments under review, the following elements of the ATM system have been identified:

- Controller Working Position (CWP), which refers to the display, interactive device and all the working tools available for the controllers’ team.
- Flight Deck (FD), which refers to the display, interactive device and all the working tools available for the flight crew.
- Decision Support Tools (DST), which provide processed and displayed information assisting the human actors in their tasks (e.g., URET or CDTI).
- Procedures, which purpose is to define the task allocation, responsibility sharing and how to perform these tasks, including the management of non-nominal situations (e.g., level of responsibility transfer, how to do in case of a necessary cancellation of delegation).
- Human actors (i.e., controllers/pilots), who play a major role in the ATM system through their use of the procedures and decision support tools.
- Communications, which include the exchanges of information between pilots and controllers.
- Air traffic operations, which correspond to the evolution of aircraft in the ATM system.

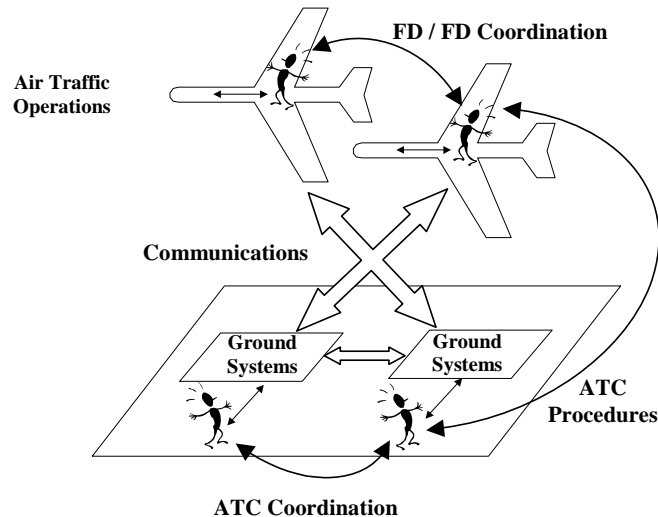


Figure 5: Elements of the ATM system under experimentation

- **Metrics description**

Description of each metrics derived from the experiment including:

- Its type, name and attributes,
- Related experiment objective(s) under assessment, and the decision criteria to decide whether or not the objective is met;
- Means of calculation including the possible aggregation of “lower level” metrics/measurements (if any).

- **Metrics type**

It is proposed to distinguish between:

- Absolute metric to be compared with a threshold / Relative metric (or indicator, as named in MAEVA) to conclude on trends

- **Metrics attributes**

Cf. MAEVA Validation Guideline Handbook (experiment data types) [5].

- Objective (Measured without asking for an opinion) / Subjective (Opinion requested and response based on subjective viewpoint of the data provider)
- Qualitative (Text based descriptions or opinions) / Quantitative (Numerically expressed values)
- Intrusive (Participant is necessarily aware of the data collection method during the operational work and likely to be affected by the measurement system) / Non-intrusive (Participant may be unaware of the data collection method used, as it does not impact on the work)
- Binary (Only 2 possible values) / Not binary (More than 2 possible values)

The following table summarises the various metrics characteristics that need to be identified during the review of the experiments.

Metric name			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective/subjective Qualitative/quantitative Intrusive (or not) Binary (or not)	Cf. Experiment objectives description	Absolute metric to be compared with a threshold / Relative metric to conclude on trends	Cf. data collection and analysis method

Table 5: Description of each metric from the experiment under review

- **Data collection and analysis method(s)**

Description of the source of the data used to compute each metric, as well as the data processing method.

4. COMPARISON OF OPERATIONAL CONCEPTS UNDER ASSESSMENT WITHIN AGIE AND EACAC

This section performs a comparison of the operational concept under assessment within the AGIE experiment (cf. ANNEX A) and the EACAC'2000 experiments (cf. ANNEX B).

4.1. Outline of concepts under validation

The following table presents the main characteristics of the concepts under validation within both experiments.

	AGIE experiment	EACAC'2000 experiments
Concept of operations	Shared-separation authority between air traffic controllers and flight crews (four conditions) Major changes in current roles and working methods of controllers and flight crews	Limited delegation of separation tasks from controllers to flight crew Minimum change in current roles and working methods of controllers and flight crews
Operating environment characteristics	High altitude En route airspace	Both extended terminal areas and en-route airspace in Europe core area
Stage in the life cycle	Concept Phase (or User Requirement)	In between , Concept Phase (or User Requirement) and Feasibility Phase (or User Requirement Analysis)

Table 6: Outline of concepts under validation within AGIE and EACAC

AGIE experiment

The AGIE experiment was a large-scale investigation of the concept of “shared-separation” authority in en-route airspace, including both ground-side and air-side evaluation.

Four operational conditions were simulated, the first condition corresponding to the current operational situation. The three other conditions were related to new concepts of operations suggesting a different shift of aircraft separation responsibility from air traffic controllers to flight crews.

Prior to AGIE, the concept of “shared-separation” had started to be evaluated separately at NASA and the FAA. These studies have been done on air issues, ground issues and supported tools individually.

EACAC'2000 experiments

The concept studied with EACAC consisted in “limited delegation” of separation tasks (i.e., implementation and monitoring tasks) from controllers to flight crews, both in extended terminal areas and en-route airspace. The operational concept was not new, and had already been refined following previous EACAC experiment (June 1999).

4.2. Comparative analysis of concepts under validation

Although AGIE and EACAC are both investigating the delegation of separation tasks to the flight deck, the operational concepts under assessment are distinct in terms of:

- Maturity (i.e., stage in the life cycle),
- Concept development strategy (e.g. AGIE four conditions), and
- Level of implications of the concept on the ATM system (i.e., impact on the current situation).

4.2.1. Maturity of the concepts

Both experiments are dedicated to the design of the concept (i.e., Concept Phase). However, the “shared-separation” concept studied within AGIE was a bit less mature than the concept of “limited delegation” studied within EACAC.

Since the concept studied within EACAC already benefited from initial feedback from controllers and flight crews obtained through previous experiment, the year 2000 experiments were investigating not only the concept viability, but to a lesser extent its feasibility in terms of procedures and tools.

On the other hand, the AGIE experiment was much more dedicated to the definition of the concept of “shared-separation” itself through the investigation of three operational conditions.

4.2.2. Concept development strategies

The underlying strategy for the concept development looks different within each of the experiments:

- AGIE strategy may be expressed as a **confrontation between several conditions** in order to identify the most promising concept of operations– maybe composed of some of the three conditions under evaluation.
- EACAC strategy may be expressed as an iterative and incremental building of the concept based on an idea or assumption. The approach is to progressively define, enrich or change the operational concept based on the experiment results. There is a **target concept studied against the baseline**.

The objective here is not to discuss on the efficiency of one or the other strategy, but to stress the characteristics that could explain the differences observed in validation approaches (cf. section 5). For example, one reason of the difference of strategies might be the slight difference in concept maturity.

Whatever the reason, the AGIE experiment was more complex than the EACAC experiment due to the investigation of various operational conditions at the same time.

4.2.3. Level of implication of the concepts on the ATM system

The concepts studied within both experiments also appear to have different impact on the ATM system. The concept studied within AGIE implies major changes in the responsibility sharing between flight crews and controllers. On the other hand, the concept studied within EACAC advocates minimum change in their current roles and working methods.

As a consequence, getting used to the new working methods might have been more difficult for the AGIE participants than for the participants in EACAC. From the controllers' point of view, the adequacy between the new procedures and decision support tools was less obvious for AGIE than the EACAC experiment.

In this respect, the AGIE experiment was more complex than the EACAC experiment because of the "gap" between the ATM system implied by the concept(s) under assessment and the current ATM operational situation.

5. COMPARISON OF AGIE AND EACAC VALIDATION APPROACHES

This section performs a comparison of the AGIE experiment (CF. ANNEX A) and the EACAC'2000 experiments (cf. ANNEX B) with respect to the validation approach used.

5.1. Outline of AGIE and EACAC validation approaches

The following table presents the main characteristics of the validation scope and objectives of both experiments.

	AGIE experiment	EACAC'2000 experiments
Stage in validation	Agreement about basic principles of the new concept (V1)	Initial proof of the concept (V2)
Validation objective(s)	Human involvement and commitment Safety Economics: flexibility and efficiency of the Airspace System	Human involvement and commitment Safety Capacity Economics: flight efficiency
Validation point of view	Air Traffic Controllers, and Flight crews.	Air Traffic Controllers (mainly)
Object(s) under validation	Use of shared separation operations under nominal conditions Information requirements	Use of the delegation procedures under nominal conditions Controller working position
Validation activity	Real-time simulation	Small-scale real-time simulation

Table 7: Outline of validation approaches within AGIE and EACAC

AGIE experiment

Within the early stage of the life cycle, AGIE simulators were of a rather high level of realism, aimed at **confronting three operational conditions of shared separation** in realistic technical environment. The purpose (i.e., validation objectives) was to collect data from controllers and pilots on shared separation procedures, information requirement, workload and situation awareness. Therefore, the focus was on “human involvement and commitment” of the four conditions.

The objective was **not to validate a concept against expected benefits, but rather to get information that would improve the design of the concept**. For example, one objective was to provide recommendations about procedures should the concept of “shared-separation” be implemented.

EACAC'2000 experiments

Being small-scale real-time simulations performed in early stages of validation, the EACAC'2000 experiments were focused on "human involvement and commitment" of the new procedures. The objective was to make initial evaluation of their effects on safety, capacity (through the impact on controller workload, strategy and activity) and flight efficiency.

Therefore, the purposes of the experiments were **both to validate the operational concept against some potential benefits, and to further develop it** using feedback from human participants (mainly, air traffic controllers).

INTEGRA study on EACAC'2000

As the first of a series of tests on various experiments, the INTEGRA metrics and methodologies were applied on the EACAC'2000 experiments. At that time, only the INTEGRA capacity and safety metrics were specified, and recently implemented.

There was an attempt to conclude on the effectiveness of the INTEGRA metrics under development, as well as on the expected benefits from the operational concept under assessment in the EACAC'2000 experiments.

However, there was no attempt at that time to compare the results obtained by the INTEGRA metrics to the results obtained by the EACAC metrics.

5.2. Comparative analysis of both validation approaches

The scope of validation performed by AGIE and EACAC experiments has a lot of common features, together with some differences in the approach due to:

- Their common nature, i.e., human-in-the loop experiments performed in early stages of concept development (i.e., Stage 1 according to the FAA/EUROCONTROL Action Plan 5: Operational Concept Validation Strategy Document).
- The slight difference of development stage (and maturity) of the concepts under assessment, and above all the different strategies for the concept development.

5.2.1. Validation strategies of both experiments

Both experiments are focused on the validation of the design of the operational concept, rather than the actual evaluation of expected benefits and constraints of the operational concept.

However, the validation approaches within AGIE and EACAC are quite different (cf. 4.2.2 Concept development strategies): comparison against a baseline scenario in EACAC versus confrontation of several conditions in AGIE.

This major distinction between both experiments may result from their slight difference in the development stage of the concept under validation (cf. section 4.2), but this could not be the only explanation.

The rather different “histories” of the concept development might also have been an important influencing factor in the experiments design:

- AGIE sprang from several studies focusing on different operational conditions (with more or less changes in human actors tasks and responsibilities) and with different perspectives (i.e., controllers, pilots).
- EACAC sprang from reflection leading to a set of assumptions and principles applied to the concept (e.g. make it as simple as possible, avoid too many changes for the human actors).

5.2.2. Human-in-the-loop experiments

Both experiments were focused on the same perspective (mainly, human point of view) to support the design of the concept elements, such as appropriate information, procedures and decision support tools.

Focusing on human factor aspects, it is relevant to perform real-time simulation. In the validation literature, it is often said that real-time simulation should be performed while the operational concept is defined in detail (later stage of the validation life cycle)³.

Defining human role, working method, procedures can be supported by real-time simulation as a means to explore possible solutions. As such, real-time simulation is a means supporting the concept design in the early stage of development (and so of the validation process).

³ See for example MAEVA

6. COMPARISON OF AGIE AND EACAC'2000 EXPERIMENTS

This section performs a comparison of the AGIE experiment (CF. ANNEX A) and the EACAC'2000 experiments (cf. ANNEX B) with respect to:

- The experiment set-up and performance, and
- The experiment objectives and hypotheses.

6.1. Outline of AGIE and EACAC experiments set-up and performance

The following table provides an overview of the main characteristics (scope, experimental conditions and recognised limitations) of both experiments.

The purpose is to highlight the main similarities and differences in experiment set-up and performance between AGIE and EACAC.

	AGIE experiment	EACAC'2000 experiment
Scope of the experiment	<p>Two adjacent en-route sectors (high altitudes);</p> <p>Traffic samples derived from two traffic recordings, close to moderate to high-density traffic;</p> <p>Full (CDTI) equipped traffic;</p> <p>Controller positions: no paper strips, URET.</p>	<p>Two distinct airspace organisation (extended TMA and en-route);</p> <p>Traffic samples derived from two traffic recordings; traffic close to high-density traffic;</p> <p>Full (CDTI) equipped traffic;</p> <p>Controller positions: paper strips (in June), and strip-less CWP with marking facilities (in November).</p>
Experimental conditions	<p>Four sessions (of one week each) between November 1999 –and February 2000; each session performed concurrently at the WJHTC and NASA ARC;</p> <p>Experienced controllers (12) and experienced pilots (6);</p> <p>Four groups participated for 3 days with 1 day and a half kept for training;</p> <p>Three new operational conditions to be compared against the current one: Current Operation (CO); CO + CDTI; Shared Separation Level 1; Shared Separation Level 2.</p>	<p>Two sessions (of two weeks each) in June and November 2000;</p> <p>Experienced air traffic controllers (2x6) from different European countries, and test and airline pilots (5) in November only;</p> <p>Set of exercises (training, qualitative and measured) each week of experiment;</p> <p>Each exercise was simulated twice: with and without delegation.</p>

	AGIE experiment	EACAC'2000 experiment
Experiment assumptions (and constraints)	<p>Limited number of participants, four unique data collection runs based on two traffic scenarios.</p> <p>Due to relatively short flight segments (20 minutes), most air-side efficiency measures were not possible to analyse.</p> <p>Some of the aircraft entered the simulation too close to the sector boundary.</p> <p>The controllers were able to distinguish the pilot participants from the pseudo-pilots, for most of the simulation.</p>	<p>All the traffic was equipped to receive delegations, thus offering maximum opportunities to use it.</p> <p>Due to the short training, the exercises without delegation were performed before the exercises with.</p> <p>Due to technical problems and lack of controllers' training to the new CWP (in November), the participants did not manage to get familiar enough with the delegation procedures.</p>

Table 8: Outline of AGIE and EACAC experiments characteristics

Since each experiment had its own constraints and assumptions, no detailed comparison is made hereafter. Instead, each experiment set-up and performance is briefly discussed.

AGIE experiment design and performance

Having identified early the limits of the simulation is a positive point that avoids abusive interpretation or generalisation of the results obtained.

According to the duration of the simulation and the number of participants, it seems that **comparing three operational conditions was rather ambitious**. In addition, the results of the study identified that non-harmonisation of the conflict alert look-ahead times may be an important operational limitation for the concept.

One and a half days kept for training seems limited to make participants get used to the new responsibility sharing and new tools.

From a technical perspective, the study was the first successful experiment in which both the flight deck and air traffic control were simulated with high fidelity. Actually, big amount of efforts were invested in the experiment to set-up a large-scale infrastructure that would support future experiments.

But, from the concept design and validation perspectives, the experiment, which involved limited number of runs per operational condition, did not perform optimal use of the technical work required on platforms, as well as the work required to set-up the experimentation of a great number of variables (i.e., three operational conditions).

EACAC'2000 experiments design and performance

Main assumptions and limitations of the EACAC experiments had been identified and sometimes taken into account in the judgement-based analysis of the experiment results.

In particular, the favourable context in which delegation procedures were experimented (e.g. all equipped aircraft and baseline scenario simulated first) was clearly stated in the EACAC report.

Similarly the limited training of the EACAC participants, particularly to the new simulated environment in November 2000, was well recognised. Nevertheless, despite these **discrepancies between the June and November experimental conditions**, the results obtained during both sessions were most of the time combined, except for the safety-related measures. The rationale for keeping some results, but not all, from the November experiment is not obvious.

6.2. Comparison between experiment objectives and hypotheses

6.2.1. Outline of AGIE and EACAC experiments objectives and hypotheses

The following table provides an overview of the main objectives and hypotheses of both the AGIE and EACAC experiments.

Distinction is made between objectives that support, although not directly, the validation of potential benefits and constraints of the operational concept (i.e., “external” validation) and those that directly support design of the new ATM system associated with the concept (i.e., “internal” validation).

	AGIE experiment	EACAC’2000 experiment
Experiment objectives (related to validation)	<p>To identify operational issues (e.g., communications and procedures) that affect shared-separation operations</p> <p>To evaluate impact on controller and pilot workload and situation awareness</p> <p>To evaluate impact on flexibility and efficiency of the Airspace System</p>	<p>To assess users (i.e., controllers and pilots) acceptance</p> <p>To assess the impact on controllers’ activity</p> <p>To assess the impact on flight efficiency</p> <p>To assess the impact on safety (not technical aspects of safety, but rather on the human contribution to safety, as well as the achievement of safe separations of aircraft)</p>
Experiment objectives (related to design)	<p>To provide recommendations for the information requirements and procedures</p>	<p>To assess the usability of the new controllers’ working position interface</p> <p>To improve the procedures and the phraseology</p>

	AGIE experiment	EACAC'2000 experiment
Hypotheses related to experiment objectives (related to validation)	<p>The three operational conditions tested have an impact on shared separation operations regarding:</p> <ul style="list-style-type: none"> - The perceived amount of time available to assure safe a/c separation and required co-ordination - The perceived level of safety - The frequency and duration of URET alerts - The minimum separation distance and cancellation of free flight - Controller/pilot manoeuvre strategies for conflict resolution, - The workload and situational awareness (respectively for PC and TC and captain and first officer). - The potential confusion in separation responsibility (flight crew/controllers) 	<p>The controllers' availability has a potential (positive) impact on capacity</p> <p>The users' acceptance of delegation has an impact on effective use of procedures</p> <p>Perceived level of safety as an impact on controllers' contribution to safety</p> <p>The use of delegation procedures have an impact on controllers' activity through:</p> <ul style="list-style-type: none"> - The controller's workload - The controllers strategies (in terms of tasks performed) - The controllers' activities (in space and time), including controller-pilot communications <p>Increase in controller's availability is expected through:</p> <ul style="list-style-type: none"> - Reduction of the manoeuvring instructions - Less time critical controllers' instructions (considered as time consuming since they require active monitoring) - Smoothing over time of controllers' instructions

Table 9: Outline of AGIE and EACAC experiment objectives and hypotheses

AGIE experiment objectives and hypotheses

As said previously, the AGIE experiment objectives reflect the stage of the design life cycle. The purpose was to explore operational options in order to guide the design (i.e., to define roles, procedures, rules of application and necessary tools). Therefore, the experiment objectives are **more design objectives than validation objectives**.

These objectives were assessed through **judgement-based interpretation of results** for the ground side, the air side, and the integration of both.

What is lacking (at least, in the AGIE final report) is the rationale and assumptions behind each condition investigated. **What are the expectations of each condition?** What are the hypotheses about their impact on the situational awareness, the workload, the perceived safety or the strategy of both controllers and pilots?

Moreover, some important factors affected the human performance and behaviour have not been object of study. For example, it would have been productive to study the impact of the conditions on factors such as confidence (in others confidence and self-confidence); motivation (factor of performance); teamwork; “learnability”. The reason is that these factors have an influence on the objectives such as safety, efficiency and capacity.

Furthermore, there are no explicit assumptions made on the relationship between the validation objectives related to human factors and the ATM performance objectives. For example, what is the likely relationship between workload and safety or efficiency, and between situational awareness and safety or efficiency?

EACAC'2000 experiment objectives and hypotheses

There was an attempt to conclude on some trends related to the expected benefits from the operational concept under assessment, and more precisely on its impact on the way the ATM system (including the human) would perform. This was done through **judgement-based interpretation of results** obtained during the baseline operational scenario and the new operational scenario (with delegation).

In order to conclude on some validation objectives, **some hypotheses** were formulated about the relationship that may exist between elements of the ATM system, like for instance between controller's workload, strategies and activities. It is not obvious whether all these hypotheses were explicitly stated prior to the experiments or were consolidated from the simulation results.

Furthermore, these hypotheses remained quite simple, and did not try to cope with the complexity of the ATM system and its components. For instance, controller's workload was only taken into account to assess the impact on controller's activity, and not his acceptance of the operational concept. And, although the relationship that may exist between the controllers' activity and the capacity of the overall airspace was evoked, there was no attempt to conclude on possible capacity increase.

More generally, there was almost **no attempt to correlate the metrics** related to factors that may interact with each other, except in case of some unexpected results. For example, the controllers' poor familiarity with the simulated environment was used to explain their ratings of their perceived workload during some experiment.

6.2.2. Comparative analysis of both experiments objectives and hypotheses

The validation objectives and hypotheses taken by both experiments have a lot of similarities in terms of:

- Their focus on human factors criteria with more or less level of details depending on the experiment,
- The weak relationship between the experiment objectives and clearly identified hypotheses about the impact of the concept on the ATM system.

6.2.2.1. Hypotheses and objectives focus on human factors

Both experiments were focused on the **assessment of human-related performance areas**, including human workload, human contribution to safety and controller/pilot strategies in performing their tasks.

The notion of “**situational awareness**” is not explicitly mentioned in EACAC, whereas the notion of “**acceptability**” is not formulated explicitly in AGIE. What was interesting in AGIE is the integration of both sides of the system: controllers versus flight crews, which made possible the investigation of the concept onto the overall system.

Nevertheless, like most of the experiments in the ATM area, AGIE and EACAC experiments suffer from the lack of model linking validation objectives such as workload, confidence (i.e., human factors criteria) with ATM performance objectives (e.g. safety, efficiency, capacity).

Although for both experiments, the underlying “high-level” validation question is close to: “does the concept **seem promising** in terms of safety and efficiency?”

6.2.2.2. Explicit or implicit relation between hypotheses and objectives

The difference between both experiments is that EACAC have defined some hypotheses whereas AGIE did not. For example in EACAC, “The controllers’ availability has a potential (positive) impact on capacity.” And, increase in controller’s availability is expected to derive from positive changes in controllers’ activity and workload.

It reflects a general feeling that in EACAC, there was a deeper analysis about the impact of delegation of controllers’ activity. This is compliant with the difference in the strategies of development of the concept (cf. section 5.2).

Nevertheless, both experiments suffer from a lack of hypotheses about the impact of the concept on the ATM system itself, as well as its outcomes and outputs. In this context, the AGIE and EACAC experiments were limited to a judgement-based analysis of metrics measured during different scenarios. And, there were no explicit criteria to decide on whether or not the validation objectives were met.

7. COMPARISON BETWEEN AGIE, EACAC AND INTEGRA METRICS

This section performs a comparison of the metrics identified through the review of the AGIE experiment (CF. ANNEX A), the EACAC’2000 experiments (cf. ANNEX B), as well as the INTEGRA study applied to the EACAC experiments (cf. ANNEX C).

This comparison is made at different levels including:

- The characteristics of the metrics, including their different attributes, their means of calculation and decision criteria to conclude on whether or not the objective is met,
- The metrics scope and perspective with distinction between those related to ATM system outcomes (or outputs) and those related to specific ATM system components, and finally
- The different metrics related to common performance areas (either “internal” performance factors related to ATM system components or “external” ATM performance factors).

The purpose is to establish the relationship between the metrics used in both experiments to assess various aspects (e.g., human factors, efficiency or safety), to discuss the main differences and similarities between the various metrics, as well as to provide judgement about their relevance.

7.1. Comparison between metrics characteristics

7.1.1. Outline of AGIE, EACAC and INTEGRA metrics characteristics

The following table provides an overview of the main characteristics of the metrics used in both AGIE and EACAC experiments, as well as the INTEGRA metrics applied on EACAC.

	AGIE experiment	EACAC’2000 experiment	INTEGRA metrics on EACAC’2000
Metrics attributes	<p>More subjective and qualitative metrics; less objective and quantitative metrics</p> <p>Most subjective metrics with a scale of five level of ratings</p> <p>Very few intrusive metrics (using the WAK)</p>	<p>Both subjective and qualitative metrics, and objective and quantitative metrics</p> <p>Most of subjective metrics with four level of ratings</p> <p>Some intrusive metrics (ISA and physiological measurements)</p>	<p>Only objective and quantitative metrics, function of time</p> <p>No intrusive metrics</p>
Decision criteria	Expert judgement based on comparison between the four conditions	Judgement-based comparison of results with and without delegation	Judgement-based comparison of results with and without delegation

	AGIE experiment	EACAC'2000 experiment	INTEGRA metrics on EACAC'2000
Data collection (or measurements)	(Mainly) Post-run form questionnaires (subjective metrics) During-the-run forms and ratings (subjective metrics) Post experiment analysis of system recordings (objective metrics)	(Mainly) Post-experiment questionnaires (subjective metrics) During-the-run ratings (subjective metrics) Post experiment analysis of system recordings (objective metrics)	Post experiment analysis of system recordings
Data processing and analysis	To check validity of data level or significance: at least, means and standard deviation; Often ANOVA test Some descriptive statistics	Mean values of collected data per controller position, per exercise or per experiment	Weighted aggregation of collected data (i.e., the control actions and the traffic situation) on a given time span

Table 10: Outline of AGIE and EACAC metrics and measurements

AGIE metrics and measurements

Within AGIE, the method of analysis is generally based on expert judgement of the data collected whatever their origin (qualitative or quantitative). Both **subjective and objective metrics** were combined to assess some aspects of the ATM system (e.g. user's workload, or safety).

Due to the exploratory nature of the research and small sample sizes, inspection of the means and ± 1 standard error of measurement (SEM) was used as the primary method to analyse the data. The Analysis of Variance (ANOVA) statistical method was sometimes used to provide additional insight for future areas of research, particularly for the ground-side data.

This **statistical analysis of validity of the data collected** was useful to avoid to a certain extent misinterpretation of the results.

EACAC metrics and measurements

Within the EACAC'2000 experiments, the metrics almost always consisted in rough cumulative sums or averages of objective measurements or human ratings. There was **no statistical analysis** of the data collected through the experiments and the metrics derived from these data.

Both **subjective and objectives metrics** were used to assess some aspects of the ATM system (e.g. user's acceptance or safety). When related to the same aspect of the ATM system, the objective metrics did not always confirm the results obtained through the subjective metrics. This was particularly the case for the metrics related to controller's workload and safety assessment.

With regard to subjective metrics (collected through questionnaires), even number of levels of rating was almost always used, in order to force controllers to take position in favour or against evaluated items, although in some occasions no answer was given.

INTEGRA metrics

The INTEGRA metrics are designed to apply to any ATM system, regardless of which elementary control task is achieved by automated means or humans. As such they consist in a **weighted aggregation of objective measurements**, most of which were cumulative counts per exercise and per actor.

The measurements come from different sources, which must be measured accurately for the metrics to provide their full efficiency. This requires controller computer assistance tools whose outputs enable to discriminate between sources.

This was not fully the case for EACAC'2000 experiments. In addition, a number of weighting and scaling factors used to compute the INTEGRA metrics were set empirically within the study on EACAC'2000. Other simulations would have been required to fine tune them and make a sensitivity analysis.

Therefore, the absolute values of the INTEGRA metrics applied on EACAC'2000 were not significant and **only relative comparisons** (with and without delegation) could be fruitful.

7.1.2. Comparative analysis of experiments metrics and measurements

The characteristics of the metrics used in AGIE and EACAC'2000 experiments, as well as those used within the INTEGRA study on EACAC, are further compared hereafter.

7.1.2.1. Balance between subjective and objective metrics within AGIE and EACAC

The types of metrics are comparable in both experiments with a **majority of subjective data collected with non-intrusive data collection method**. This is coherent with the focus of the studies, which relied on human evaluation point of view.

The use of objective metrics is greater within EACAC than in AGIE. This might be explained by the slight difference in the development stage of the concepts under validation. Indeed, since the EACAC concept had already been investigated through previous experiment, the use of objective metrics might have been judged useful in the perspective of making an initial proof of the concept and its impact on the ATM system.

Similarly, with regard to subjective metrics collected through questionnaires, EACAC was much more "demanding" than AGIE with the use of even number of levels of rating (instead of odd number levels of rating within AGIE) in order to conclude positively or negatively about the evaluated items.

7.1.2.2. Data collection and analysis methods within AGIE and EACAC

Within both experiments, most of the data reported were collected by observation, questionnaires and debriefing. The main difference relies on the data processing method that involved **statistical calculation within AGIE, but not within EACAC**.

This is a useful process as it allows to potentially put aside some non-significant results or to emphasise significant ones. However, it does not allow to decide if it is relevant or not to generalise the results, i.e., to decide if the experimental conditions represent realistic ones. For that purpose, more sophisticated tests are required which are not necessarily relevant in early stages of the life cycle.

7.1.2.3. INTEGRA metrics on EACAC'2000

Unlike the EACAC'2000 metrics, which consisted in a combination of subjective and objective metrics, the INTEGRA metrics are by design only based on objective data.

In theory, the metrics also differ by their intended purpose of use. Whereas the EACAC metrics were defined and used for comparison on some aspects of the ATM system with and without delegation concept, the INTEGRA metrics are designed to apply for evaluation of the performance of any ATM system. As such, the INTEGRA metrics are much more ambitious and requires the establishment of absolute criteria to be used as threshold for evaluation of the metrics.

In the INTEGRA study on EACAC'2000, this was not possible due to the level of maturity of the metrics themselves (i.e., weights and scaling factors not yet validated), as well as the difficulty in collecting appropriate objective measurements from the experiments. This limited the use of the metrics as absolute indicators of the performance of the new ATM system experimented within EACAC.

7.2. Comparison between metrics scope and perspective

To support the identification of appropriate metrics regarding a given validation objective, the comparison of the metrics used within AGIE and EACAC is made according to their validation objectives (or performance areas). When required, reference is also made to the elements of the ATM system on which the metrics are applied.

The performance areas addressed within the AGIE and EACAC'2000 experiments, including the INTEGRA study on EACAC, include the followings:

- **Human performance (and acceptance),**
- **Efficiency** and
- **Safety.**

The various metrics related to these three performance areas are further compared and discussed respectively in sections 7.3, 7.4 and 7.5. Some of them may appear several times, being related to several validation objectives.

7.2.1. Outline of AGIE, EACAC and INTEGRA metrics

The following table provides an overview of the different levels at which metrics were defined (or interpreted) within the AGIE and EACAC'2000 experiments, including the INTEGRA study on EACAC.

		AGIE experiment	EACAC'2000 experiment	INTEGRA metrics on EACAC'2000
ATM system outcomes and outputs (“External” performance areas)	Expectations (or aims)	Safety Economics: flexibility and efficiency of the Airspace System	Safety Capacity Economics: flight efficiency	Safety Capacity
	Expected operational benefits (and constraints)	Safe a/c separations	Safe a/c separations Earlier sequencing of aircraft Flight efficiency	Safety significant events

		AGIE experiment	EACAC'2000 experiment	INTEGRA metrics on EACAC'2000
ATM system components ("Internal" performance areas)	Procedures (and rules)	Use of procedures	Use of procedures	
	Human actors	Perceived safety Strategies for conflict resolution Coordination (and communications) tasks Controllers/pilots workload and situational awareness	Perceived safety Users' acceptance Controller's availability Controller's activity: - Controller's workload - Controller's strategies - Controller's activity Controller-pilot communications	Information Processing Load (IPL)
	Systems	URET, CDTI	Controllers' Working Position interface	

Table 11: Outline of AGIE, EACAC and INTEGRA metrics

AGIE metrics

In order to identify operational issues linked to the three operational conditions under evaluation, AGIE defined a set of metrics related to **efficiency and safety** of the overall airspace system. Metrics were also defined to assess the impact of the new concept on **human workload and situational awareness**.

The AGIE metrics aimed at investigating both the controllers' and the pilots' point of view, sometimes with rather distinct scope and perspective. For example, the efficiency-related metrics were addressing both control efficiency and flight efficiency.

EACAC'2000 metrics

Most of the metrics defined in the EACAC'2000 experiments aimed at assessing the impact on **controller's workload, strategy and activity**, in order to investigate the potential impact on airspace capacity in relationship with controllers' availability.

Some metrics were also defined to assess the impact on **flight efficiency**, and on **safety** of air traffic operations. These metrics were not focused on technical aspects of safety, but rather on the human contribution to safety, as well as the achievement of safe separations of aircraft.

INTEGRA metrics on EACAC'2000

The INTEGRA metrics applied on EACAC'2000 were limited to the **capacity and safety** metrics. In particular, the INTEGRA efficiency metrics were not implemented at the time of the INTEGRA study on EACAC'2000.

Furthermore, different **factors limited the significance and interpretation** of the INTEGRA metrics applied on EACAC'2000.

- The metrics depend on a number of weighting and scaling factors which had been set empirically, thus limiting their level of significance as absolute metrics;
- The maximum “Information Processing Load” that the control can bear (to be used as a threshold for evaluation of the Capacity metric) was still to be determined;
- The abstractness of the Safety metrics.

Consequently, the INTEGRA study on EACAC’2000 could only conclude on the good responsiveness of the Capacity metrics, based on the expected assumption that delegation decreases the controller’s “Information Processing Load”.

Other applications of the metrics in more favourable conditions were considered necessary to further polish and assess them.

7.2.2. Comparative analysis of metrics scope and perspective

The metrics used within the AGIE and EACAC experiments have a lot of similarities, in particular with respect to:

- The balance between the metrics related to “external” and “internal” performance areas,
- The scope of the metrics spanning human-related metrics, efficiency and safety metrics. Although, different perspectives are sometimes taken into account to address the same performance area.

7.2.2.1. Balance between “external” and “internal” performance metrics

Almost no (direct) metrics was related to expectations and expected operational benefits (i.e., “external” performance areas) of the concept (for both AGIE and EACAC), except for safety through provision of safe aircraft separation and flight efficiency.

In particular within EACAC, the impact on airspace capacity, although mentioned as a “high-level” validation objective, was investigated through the impact on controller workload, strategy and activity.

Similarly within AGIE, although the targeted “high-level” validation objective was to assess efficiency and flexibility of the airspace system, the point of view taken was much more related to “internal” validation of efficiency from the controllers’ point of view.

Compared to the metrics used in both experiments, the INTEGRA **metric based on control activity** is closer to the “external” validation objective of assessing the system’s capacity. Nevertheless, it does not give a direct measure of, for example, how many aircraft can be dealt with by the control.

Actually, trying to perform an objective “external” validation (i.e., performance of the overall ATM system) would have not been very relevant since both experiments had low-level of realism and few collected data. Even in AGIE where simulators were of high-level of realism, there was no weather model, only “simple” conflicts and full equipage operations. Focusing on human activity, the “internal” validation deals with several objectives without explicitly stating:

- What are the assumptions behind them,
- Their mutual relationship, and
- The link with the “external” validation (i.e., overall ATM system performance).

7.2.2.2. Metrics scope and perspectives

Safety, efficiency and a number of human-related metrics are provided in both AGIE and EACAC experiments sometimes with different scope and perspective.

Human-related metrics

Although both AGIE and EACAC aimed at assessing human involvement and commitment of the new concepts, the scope of their **human-related metrics** is different:

- Within EACAC, focus was put on assessing the impact of delegation on controller’s activity. This impact was considered **at three levels**: individual workload, strategies in handling aircraft and resulting activity (e.g. instructions used in space and time). In addition, controller’s acceptability and availability were also assessed (more or less directly).
- Within AGIE, the effect of shifting separation authority was mainly investigated at the level of controllers/pilots workload, situational awareness and conflict detection and resolution strategies.

With respect to AGIE, some metrics have been classified in “**human activity**” metrics category when they are related to human strategy, even if they were not classified as such in the AGIE final report.

Actually, **all these human factors criteria are closely inter-related**. On one hand, strategy and impact on the human activity are closely influenced by the subject’s acceptability of the new operations, and also closely related to workload. On the other hand, human acceptability is influenced by the workload induced by the activities of each individual.

Capacity-related metrics

From both EACAC and INTEGRA perspectives, **capacity is approached by the evaluation of human workload**. The rationale behind this perspective is: the lower the controller’s workload is, the greater his ability to handle traffic (i.e., direct link with sector capacity).

In particular, the INTEGRA capacity metric made a direct link between the controller’s workload and the controller’s activity, through the notion of controller’s “**Information Processing Load**”.

Efficiency-related metrics

With respect to **efficiency-related metrics**, both AGIE and EACAC’2000 experiments investigated **flight efficiency** from a single aircraft perspective.

In addition, AGIE also aimed at investigating **control efficiency** from the overall airspace system perspective. More precisely, from the controllers' point of view, **efficiency** referred to the "usefulness and helpfulness of the shared separation concept and the information provided to perform their tasks".

Safety-related metrics

Although not easy to evaluate, safety is obviously a main validation objective as it is the purpose of ATM. Both AGIE and EACAC have used **safety-related metrics** focused on the **human contribution and evaluation of safety**, as well as the achievement of **safe separations** of aircraft.

Such "internal" perspective (i.e., human perceived safety) and "external" perspective (i.e., achievement of safety missions) are quite complementary, although closely inter-related and not easy to assess separately.

The INTEGRA safety metrics try to give an assessment of the ATM system performance by using a **probabilistic approach** of how a given air situation can be hazardous, and by determining if the control has a **significant activity margin** to deal with a hazardous situation.

More precisely, distinction is made between the notion of "**propensity**", which measures the likelihood of a safety significant event occurring during normal operations, and the notion of "**resilience**", which measures extent to which the ATM system responds to a safety significant event without causing more such events.

7.3. Comparison between AGIE, EACAC and INTEGRA metrics related to human performance

Both AGIE and EACAC used a majority of metrics related to **human performance and acceptance** of the concepts under validation, but with rather different related validation objectives. The purpose within AGIE was to evaluate the effect of shifting separation authority, whereas EACAC were much more focused on assessing the impact on controller’s activity.

The various human factor metrics defined within AGIE and EACAC experiments are discussed and compared hereafter. Distinction is made between the metrics related to:

- Human **workload**,
- Human **situational awareness** and
- Human **activity**.

The section also discusses the relationship between the INTEGRA **capacity metrics** applied on EACAC’2000 and the EACAC metrics related to human workload and activity.

Since the EACAC’2000 final report was focused on the controller’s point of view, the comparative analysis of AGIE and EACAC metrics performed hereafter is focused on the controller perspective.

7.3.1. Comparative analysis of AGIE and EACAC controller workload metrics

The following table provides an overview of the workload metrics defined within the AGIE and EACAC experiments.

Metrics Attributes	AGIE experiment	EACAC’2000 experiment
Objective Quantitative Intrusive Not binary		Controller’s physiological parameters: - Pupil diameter, - Dwell time and - Heart rate
Objective Quantitative Not intrusive Not binary	- Frequency of Controller Ground→Air and Land Line Push-to-Talk Transmissions - Duration of Controller Ground→Air and Land Line Push-to-Talk Transmissions	

Metrics Attributes	AGIE experiment	EACAC'2000 experiment
Subjective (expert observer) Qualitative Scale with 5 options or open questions	-Expert Observer Ratings of Controller Physical Task load	
Subjective Qualitative Non intrusive Scale with 5 options in AGIE and 4 in EACAC	-Controller Ratings for Physical, Mental, and Overall Workload	- Controller ratings of possible reduction in ATCO's workload - Controller ratings of perceived overall workload - Controller ratings of perceived workload (combining ratings of mental, physical and temporal demands, as well as performance, effort and frustration levels perceived) = NASA TLX
	- Controller Workload Ratings for Maintaining Aircraft Separation, - Controller Ratings for Land Line Coordination, - Controller Ratings for R-Side-to-D-Side Coordination, - Controller Ratings for Ground→Air Transmissions, - Controller Ratings for URET Coordination - Controller Ratings for Feeling Rushed and Bored	- Controller ratings of workload / mental effort required to monitor delegations - Controller ratings of workload and stress in monitoring of delegations - Controller ratings of factors contributing to workload (i.e., tasks associated with delegation)
Subjective Qualitative Intrusive Scale with 5 Choices	- Controller Interval Workload Ratings (WAK)	- Temporal distribution of perceived overall workload per controller position (ISA)

Table 12: Outline of AGIE and EACAC controller workload metrics

Both experiments have used different sources of data collection and analysis with both subjective and objective data. In general, the resulting workload metrics are quite homogeneous.

Comparison between AGIE and EACAC subjective metrics

Within both experiments, controller's subjective rating has been used during the runs (i.e., WAK or ISA techniques) to get continuous assessment of their "individual workload". Both techniques seem to be quite similar (except for the interval time). They rely on the notion of "**perceived workload**" (or strain).

In addition, both experiments made an attempt to confirm the impact on controller's workload through the use of other subjective metrics:

- Within EACAC, both ad-hoc questionnaires collected after experiment and NASA-Task Load Index (TLX) based on data collected after each exercise were used.
- Within AGIE, expert observer had a specific form to collect data related to physical workload of the experiment participants.

Within AGIE, there was an attempt to assess **workload in relation to the tasks** (e.g., inter-sectors co-ordination; controllers' team co-ordination, air-ground co-ordination, monitoring and planning). Similarly within EACAC, controllers were asked to rate the tasks associated with delegation that contribute much to their workload.

Comparison between AGIE and EACAC objective metrics

Within AGIE, workload has also been assessed through the number of actions performed (e.g. communication between air and ground). Within EACAC, although controller/pilot communications were measured, the corresponding metrics were used to assess controller's activity rather than workload.

Actually, this is an indicator of the tasks load that has an impact on the perceived workload. To get more relevant evaluation about task load, other tasks such as team communication or interaction with the tools should also be taken into account.

EACAC has used **physiological indicators**: pupil diameter, dwell time and heart rate. The limit of such indicator is the amount of runs and the number of participants it requires to obtain significant data. Another important issue is the meaning of this indicator: at least, it refers to the level of attention, but the link between attention and workload is not so straightforward.

In the November 2000 experiment, objective physiological measurements confirm feedback obtained through final questionnaires, whereas ISA and NASA-TLX techniques both invalidated the results obtained through the final questionnaires.

The difficulty with "workload" is that the human actors may understand it differently. For example, individual may reply to questionnaires (or instantaneous self assessment) while thinking about "task-load" or strain or feeling of stress. Then, the data collected during the experiments may have different meaning.

To avoid such situation, it is essential to clearly define this notion of workload during the experiment design and set-up (particularly, if subjective workload assessment is planned).

7.3.2. Comparative analysis of AGIE and EACAC controller situational awareness metrics

Whereas EACAC did not assess situational awareness as such⁴, AGIE explicitly defined one subjective and qualitative metric (based on questionnaire) related to **overall situation awareness** of both controllers/pilots.

Analysis of the AGIE metric

Situational awareness is a complex notion that is defined in AGIE as follows: “ The term **overall situation awareness** refers to what is commonly known as the controller’s “picture” and involves processing the relevant air traffic information to develop a thorough understanding of the current situation that facilitates appropriate air traffic actions in a timely manner.”

Since the three operational conditions investigated within AGIE imply a different sharing of tasks between controllers and pilots, each of them should require different situational awareness characteristics (e.g. more or less detailed information of traffic situation, leading to change in monitoring).

The **risk with subjective global evaluation** (as performed in AGIE) is to gather opinions reflecting more the change of situation awareness compared to today while the tasks have changed, requiring a different representation of the situation (and possibly different decision support tools).

7.3.3. Comparative analysis of AGIE and EACAC controller activity metrics

The following table provides an overview of the controller’s activity metrics defined within the AGIE and EACAC’2000 experiments.

Metrics Attributes	AGIE experiment	EACAC’2000 experiment
Objective Quantitative Not binary	- Descriptive Statistics for Altitude-Resolved Planned Conflicts - Descriptive Statistics for Vector-Resolved Planned Conflicts	- Number of ATC instructions per sector - Number of each type of ATC instructions per sector
	- Mean frequency of air↔ground transactions ⁵ - Mean duration of air↔ground transactions	- Number of controller/pilot calls - Mean duration of controller/pilot calls

⁴ Actually, the EACAC’2000 experiments partially and indirectly addressed the issue of situational awareness through the subjective assessment (based on questionnaire) of the usability of the new Controllers’ Working Position interface. Nevertheless, this assessment was clearly identified within EACAC as support to the design (rather validation) of the operational concept.

⁵ Transaction was defined as all communications initiated by a controller or pilot participant including acknowledgement.

Transaction duration was measured from the beginning of the first instruction, question, or comment made by any of the controller or pilot participants to the end of the final communication on the topic.

Metrics Attributes	AGIE experiment	EACAC'2000 experiment
	- Frequency and Type of Manoeuvres Issued by Controllers/pilots to Resolve Conflicts - Combination of Manoeuvres Issued by Controllers/pilots to Resolve Conflicts	- Number of (type of) ATC instructions according to distance to IAF - Number of (type of) ATC instructions over time
Subjective Qualitative Not binary	- Controller Role and Separation Responsibility Confusion	- Controller ratings of their confidence in determining the use of delegation instructions - Controller ratings of their hesitation when applying delegation instructions - Controller ratings of their ability to use delegation instructions - Controller ratings of workload and stress in monitoring of delegations
Subjective Quantitative Intrusive Not binary	- Controller Conflict Detection and Resolution Measures	

Table 13: Outline of AGIE and EACAC controller activity metrics

EACAC'2000 metrics

Within EACAC, these metrics were used to assess the impact of the delegation concept on their current activity (i.e. what is modified in the cognitive process and in the realisation of the tasks). As part of this activity, the notion of strategy referred to the way the controllers handled the traffic and what actions they implemented.

AGIE metrics

With regard to the AGIE experiment, metrics were classified in the previous table when they relate to controller/pilot strategies, although their original (and general) objective was “to identify operational issues that affect shared-separation operations”.

Some of these metrics, typically those related to the manoeuvres issued by controllers/pilots to resolve conflicts, were used in particular to assess the potential impact on “efficiency and flexibility” of the airspace system (cf. section 7.4.1).

Comparison between AGIE and EACAC metrics

Both experiments have used a combination of subjective and objective metrics **to assess the impact of the new concept on the activity of controllers** (and pilots):

- Within EACAC, these metrics either relate to their ability to effectively use new instructions, their new strategies in handling aircraft (i.e., the actions undertaken to provide ATC), and the controller/pilot communications.

- Within AGIE, these metrics include those aiming at investigating the relation between the initiator (flight crew versus controllers) and the strategy of conflict detection and resolution.

Although no discussion has been reported in the AGIE final report about the intrusiveness of such a technique, there was an attempt to assess the impact on controller’s activity (i.e. conflict detection and resolution measures), using controllers’ reply to questions during the runs. Due to problem when collecting the answers, the data were not accurate enough.

A less intrusive technique but more expensive in terms of resources would have been the use of video, together with an auto-confrontation method (i.e., replay of the target events with verbalisation of the participants). Actually, video was used during the air-side evaluation to assess the impact on pilot’s activity related to conflict detection and resolution.

As strategy and impact on the current activity are closely influenced by the level of skills & experience, trust and acceptability and closely related to workload, it should have been interesting to combine these factors.

Within AGIE, “Confusion in responsibility” was the only factor taken into account, whereas in EACAC there was an attempt (through the subjective metrics) to assess the link between specific activities (e.g. use of new instructions, monitoring activity) and workload, stress, or other factors such as confidence and skills (e.g. ability, hesitation).

7.3.4. Comparative analysis of EACAC and INTEGRA capacity metrics

The following table provides an overview of the relationship between the INTEGRA capacity metric components and the EACAC’2000 metrics.

The purpose is to highlight those metrics, which aimed at evaluating similar aspects of the ATM system.

INTEGRA Metrics on EACAC’2000		EACAC’2000 experiment	
Metrics	Attributes	Metrics	Attributes
- Information Processing Load (overall IPL combining elementary IPLs)	Objective Quantitative Scalar function (of time)	- Controller ratings of possible reduction in ATCO’s workload	Subjective Qualitative Non intrusive Scale of 4 levels of ratings
- IPL for Acquisition of a new flight, - IPL for Co-ordination with other control agencies.	Idem		
- IPL for Determining the forecast interactions,	Idem	- Number of (type of) ATC instructions over time	Objective Quantitative Not binary

INTEGRA Metrics on EACAC'2000		EACAC'2000 experiment	
Metrics	Attributes	Metrics	Attributes
- IPL for Planning resolutions, - IPL for Implementing the planned resolutions, - IPL for Other changes to trajectory		- Controller ratings of factors contributing to workload (i.e., tasks associated with delegation)	Subjective Qualitative Non intrusive Scale of 4 levels of rating
- IPL for Monitoring conformance to Plan	Idem	- Controller ratings of workload / mental effort required to monitor delegations - Controller ratings of workload and stress in monitoring of delegations	Subjective Qualitative Non intrusive Scale of 4 levels of rating

Table 14: Relationship between EACAC and INTEGRA capacity metrics

INTEGRA metrics

To determine the overall “**Information Processing Load**”, INTEGRA divides controller’s activity in elementary tasks and assumes that each elementary task contributes with a certain weight to the total processing load. The count of each elementary task must be assessed by taking into account the tactical instructions given by the controller in relation with the environment of each aircraft under control.

Comparison between EACAC and INTEGRA metrics

Unlike the workload metrics used in the EACAC’2000 experiments most of which are subjective, none of the INTEGRA capacity metric component is related to the controller assessment of his own workload/activity. Indeed, the INTEGRA capacity metrics are by design only based on objective data related to elementary control task.

The EACAC subjective workload metrics only allowed a global comparison between the experiment without and with delegation, whereas the INTEGRA IPL metric, which was evaluated as a function of time for both experimental conditions, gave greater granularity in the results. Of course, this also means that the INTEGRA results (graphics) required more interpretation efforts.

In addition, the exact weight of the different IPLs in the overall IPL (i.e. for example, does a resolution planning take as much effort as implementing a resolution) and the maximum IPL sustainable by the controllers were not yet determined.

This limited the interest of the metric as a “high-level” capacity performance indicator, from which the EACAC metrics were far more remote.

7.4. Comparison between AGIE and EACAC efficiency metrics

The various efficiency-related metrics defined within AGIE and EACAC experiments are discussed and compared hereafter. Distinction is made between the metrics related to:

- **Control efficiency** and
- **Flight efficiency.**

As already mentioned in section 7.2, the notion of “efficiency” from an overall airspace system perspective was only addressed within AGIE, whereas as both AGIE and EACAC investigated the notion of “efficiency” from an individual aircraft perspective.

7.4.1. Analysis of AGIE control efficiency metrics

The following table reports the metrics used in AGIE and related to control efficiency.

Metrics Attributes	AGIE experiment	
	Metrics	Objective/Hypothesis
Subjective Qualitative Not binary Not intrusive	- Controller mean ratings for URET conflict alert timeliness - Usefulness and frequency of air-air communication monitoring	Helpfulness of the shared separation concept To assess efficiency and flexibility of the airspace system
Objective Quantitative Not binary Not intrusive	- Mean frequency of URET conflict alerts - Mean duration of URET conflict alerts - Frequency and type of manoeuvres issued by controllers/pilots to resolve conflicts - Combination of manoeuvres issued by controllers/pilots to resolve conflicts	To assess efficiency and flexibility of the airspace system

Table 15: Outline of AGIE control efficiency metrics

Both objective and subjective metrics were used within AGIE to assess efficiency from the controllers’ point of view, depending on the four conditions. Nevertheless, **these metrics allowed assessing control efficiency only indirectly:**

- Subjective metrics aimed at assessing the helpfulness and usefulness of the concept and its associated tool (i.e., URET) assuming that efficiency increases if the concept is helpful and useful.
- Objective metrics are either related to safety (e.g., metrics related to URET conflict alerts) or to control activity (e.g., metrics related to manoeuvres issued by controllers/pilots to resolve conflicts). Actually, the objective metrics related to the

URET conflict alerts were also used to assess the potential impact of “shared – separation” on safety.

Actually, the objective metrics are more related to a **design and feasibility objective** (i.e., investigating the strategies employed by the actors depending on the four conditions) than to a concept validation objective.

Similarly, it seems that the subjective metrics did actually measure the adequacy of the decision support tool (i.e. URET) and information provided to the controllers to perform their tasks, rather than the helpfulness and usefulness of the shared-separation concept itself.

Once again, there is a lack of explicit relation between the validation objectives and then, a lack of combination of the metrics.

7.4.2. Comparative analysis of flight efficiency metrics

The following table provides an overview of the **flight efficiency** metrics defined within the AGIE and EACAC experiments.

Metrics Attributes	AGIE experiment	EACAC'2000 experiment
Subjective Qualitative Not binary Not intrusive	- Participant (pilots) ratings of flight efficiency	- Controller ratings of increased efficiency allowed by delegations (in terms of fuel consumption and time savings)
Objective Quantitative Not binary Not intrusive		- Total distance flown by aircraft (in the sectors of interest) - Total fuel consumption over the fleet (in the sectors of interest) - Total flight time over the fleet (in the sectors of interest)

Table 16: Outline of AGIE and EACAC flight efficiency metrics

Both experiments defined flight efficiency metrics with rather different perspectives and levels of details:

- Within AGIE, only subjective evaluation was performed from the pilots’ point of view. Collecting objective metrics was planned, but no analysis was done due to the too short segment of routes.
- Within EACAC, controller’s subjective evaluation of flight efficiency was used together with a set of complementary objective metrics.

It is worthwhile to note that within EACAC the controllers misunderstood the metric, and actually evaluated the “increased efficiency” **in terms of their reduced workload** to handle same amount of traffic.

7.5. Comparison between AGIE, EACAC and INTEGRA safety metrics

Both AGIE and EACAC defined metrics related to safety, but with rather different related validation objectives. The purpose within AGIE was to identify operational safety issues that may affect shared-separation operations, whereas EACAC were much more focused on the assessment of potential safety impact of the new procedures.

7.5.1. Comparative analysis of AGIE and EACAC safety metrics

The following table provides an overview of the safety-related metrics defined within AGIE and EACAC experiments.

Metrics Attributes	AGIE experiment	EACAC'2000 experiment
Objective Non-intrusive Quantitative Not binary	- Loss of Separation for Conflicts	- Number of "very serious"/ "serious"/ "minor" losses of separations - Mean duration of losses of separations
	- Mean frequency of URET conflict alerts - Mean duration of URET conflict alerts	- Maximum Aircraft Proximity Index (API)
	- Number of cancellation of Free Flight	- Number of "catching up" situations between two aircraft in sequence at transfer between sectors
Subjective Qualitative Not binary	- Controller/pilots Ratings for the Level of Safety for Procedures	- Controller ratings of potential ("could") increase in safety
	- Controller/pilot ratings for the time available to assure safe aircraft separation	
	- Controller/pilot ratings for the amount of time available for co- ordination and communication tasks	

Table 17: Outline of AGIE and EACAC safety metrics

Both experiments have used quantitative and qualitative data, as well as subjective and objective data, to assess the potential impact on safety. Nevertheless, the scope and focus of the metrics were rather different between both experiments:

- Within EACAC, the objective and quantitative metrics were more detailed looking at the **severity of the conflicts** and the potential of later conflict in the next sector (i.e., "catching up" situations). This last metric (although quite specific to the type of air traffic operations under assessment) is interesting since it considers the consequence of action (or non-action) in an extended area (potential conflict in the next sector).

- Within AGIE, both subjective and objective safety-related metrics were used, spanning **objective characteristics of conflicts** (i.e. frequency, duration and effective resolution) and controllers' rating of **perceived safety** and time available to assure aircraft separation. This last metric (although subjectively assessed) is interesting since it could be considered as an indicator of the ability of the controllers to perform their main safety mission, i.e. provide aircraft separation. In the same idea, the cancellation of Free Flight was considered as an objective indicator of the achievable level of safety (at least, as perceived by the participants)

The EACAC'2000 experiments also allowed identifying some factors that may affect safety (e.g., human overconfidence, failure of human expertise, human errors). However, these factors were not explicitly investigated through the use of specific metrics during the experiments.

Similarly, airspace violation, missed handoffs, human errors, potential mitigation means to recover errors have not been studied within AGIE (although, there was a checklist dedicated to the Expert Observers aiming at collecting such information).

7.5.2. Comparative analysis of EACAC and INTEGRA safety metrics

The following table provides an overview of the relationship between the INTEGRA safety metrics components and the EACAC'2000 metrics.

The purpose is to highlight those metrics, which aimed at evaluating similar aspects of the ATM system.

INTEGRA Metrics on EACAC'2000		EACAC'2000 experiment	
Metrics	Attributes	Metrics	Attributes
- Propensity (likelihood of a safety significant event occurring during normal operations) ⁶	Objective Non-intrusive Quantitative Scalar function (of time)	- Number of "very serious"/ "serious"/ "minor" losses of separations	Objective Non-intrusive Quantitative Not binary
- Resilience ⁷ (based on elementary IPLs)	Idem		

⁶ The "**propensity**" metric gives a probability of having a safety significant event at each time step, taking into account the quality of the airborne and ground equipment, the density and geometry of aircraft, and even the weather.

⁷ The "**resilience**" metric makes a link between the severity of the conflicts and the controller's workload. In a similar way to the capacity metric, the assumption is that there is an "Information Processing Load" associated with each hazard. This metric has to be associated with the INTEGRA capacity metric to check if the total workload does not exceed a maximum level (not determined in the INTEGRA study on EACAC'2000).

INTEGRA Metrics on EACAC'2000		EACAC'2000 experiment	
Metrics	Attributes	Metrics	Attributes
- IPL for Correcting a failure to make a state vector change for a potential hazard - IPL due to guidance errors	Idem	- Maximum Aircraft Proximity Index (API)	Objective Non-intrusive Quantitative Not binary
- IPL for hazard analysis	Idem		

Both sets of EACAC and INTEGRA safety metrics are complementary: INTEGRA metrics gave an assessment of the overall safety of airspace at a given time through a **probabilistic approach**, while EACAC metrics measure safety over the whole duration of the experiment with **plain figures**.

In addition, the EACAC metrics combined controller subjective assessment of safety and some objective assessment of aircraft separation, whereas the INTEGRA safety metrics provide by design objective assessment of the ATM system.

Although the purpose of the INTEGRA metrics is to provide a consolidated view of the ATM system performance, the need for fine-tuning some parameters did affect both safety metrics computed in the INTEGRA study on EACAC'2000.

This limited the use of the metrics as “high-level” safety performance indicators. In addition, it was recognised that the “propensity” metrics was difficult to translate in practical terms, and therefore were difficult to interpret.

8. CONSOLIDATION OF THE AGIE AND EACAC COMPARISON

This section consolidates the comparison between AGIE and EACAC experiments. It provides a synthetic description of the relationship between the various elements of comparison developed in the previous sections, and draws the main lessons learnt from this comparison.

This consolidation is performed at two distinct levels which respectively deals with:

- The validation approaches, objectives and technique used, and
- The metrics framework, characteristics and relevance. In particular, to support a common understanding of the metrics, the various performance factors to which metrics are related to are discussed.

8.1. Validation approaches, objectives and technique

This section provides a synthetic description of the similarities and differences between AGIE and EACAC'2000 experiments, and draws the main lessons learnt from this comparison, in terms of:

- Relationship between concept development and validation,
- Balance between “Internal” and “external” validation,
- Role of human-in-the loop experiments in validation, and
- Design of human-in-the-loop experiments.

8.1.1. Relationship between concept development and validation

The **design and the validation of new operational concepts are closely related processes**. In this respect, the comparison between AGIE and EACAC provides a good illustration of the relationship that exists between concept development strategy, stage in life cycle and validation approach and objectives.

8.1.1.1. Outline of the comparison

Although both experiments were dedicated to the design of the concept (i.e. Concept Phase), the underlying strategy for the concept development looks different within each of the experiments:

- The Air-Ground Integration Experiment consisted in a **confrontation between several operational conditions** in order to identify the most promising concept of operations – maybe composed of some of the three new conditions under investigation.

- The EACAC'2000 experiments were investigating not only the concept viability, but to a lesser extent its feasibility in terms of procedures and tools, through the **single comparison between the current situation and the new concept of operations**.

These different strategies could be related to the slight difference in concept maturity. Indeed, the “shared-separation” concept studied within AGIE was a bit less mature than the concept of “limited delegation” studied within EACAC.

The concept studied within AGIE was also more challenging than the one studied within EACAC because of the major changes it imply in terms of separation responsibility sharing between flight crews and controllers.

8.1.1.2. Consolidation and lessons learnt

Whatever the rationale for one of the other strategies for concept development, it is worthwhile to note that these strategies have an impact on the validation approach and objectives associated to the experiment.

In particular, **experimental design is closely related to the development history of the concept** either: confrontation of several operational conditions (developed separately) to select among concept options, or specific evaluation of operational changes to support incremental development of the concept.

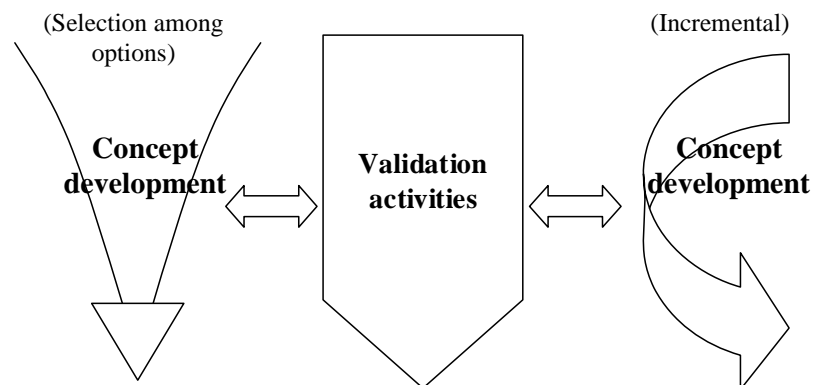


Figure 6: Relationship between concept development and validation

Both strategies could either be applied within one step of the concept development, as part of an overall iterative development process.

8.1.2. Balance between “Internal” and “external” validation

The comparison between AGIE and EACAC validation objectives also provides a good opportunity to discuss the dilemma that often exists in early stages of development between “external” validation of expected benefits of the concept and the “internal” validation of the new ATM system implied by the concept.

8.1.2.1. Outline of the comparison

Both AGIE and EACAC experiments were much more dedicated to the “internal” validation of the operational concept, rather than the actual evaluation of expected benefits and constraints of the operational concept.

In particular, both human-in-the-loop experiments were focused on the same perspective (mainly, human point of view) to support the design of the concept elements, such as appropriate information, procedures and decision support tools.

In this respect, AGIE was more complex than the EACAC experiments because of the “gap” between the ATM system implied by the concept(s) under assessment and the current ATM operational situation.

8.1.2.2. Consolidation and lessons learnt

The validation of new Operational Concept is to be performed using a stepwise approach allowing for progressive **design and validation** of a new ATM system (within a potentially new ATM environment) in line with the maturity of the Operational Concept and the development stage of the new ATM system.

Therefore, distinction has to be made between metrics related to “external” validation of expected benefits and constraints of the operational concept and “internal” validation of the design of the ATM system associated with the operational concept (e.g. appropriate decision support tools, appropriate procedures).

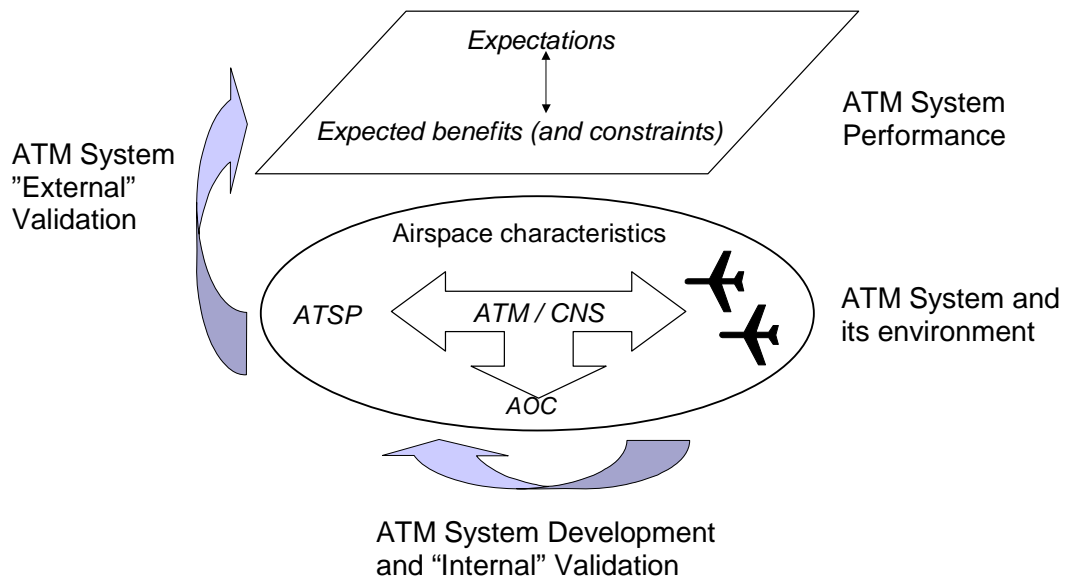


Figure 7: Development and validation of a new operational concept

Briefly said, “external” validation allows to support a “go/no go” decision about the next step of the concept development, whereas “internal” validation can be seen as a means to guide or evaluate design choices (i.e. without calling into question the relevance of the concept itself).

During early stages of concept development, **both “external” and “internal” validation objectives are often studied together.**

The challenge is that investigating “external” performance (i.e. expected benefits and compliance with constraints) requires a detailed specification of the concept, whereas deciding on to continue or not the development of a concept should ideally be done as soon as possible in the life cycle while specification are not mature.

Thus, within the concept design phase, investigation and evaluation is much more focused on the potential of the concept towards performance, rather than the ability of the concept to actually achieve a required level of ATM performances.

Taking these considerations into account, it appears essential to define the relationship that exists between the experiment objectives (generally focused on “internal” aspects of ATM performance) and more “external” validation objectives.

Then, and only then, experiment performed in early stages of development would help to:

- **Refine the concept** (e.g., feedback about possible improvements), and
- **Validate the concept** (e.g., feedback about potential external benefits).

8.1.3. Human-in-the loop experiments in the validation process

Based on the comparison between AGIE and EACAC experiment objectives, this section further discusses the adequacy of human-in-the loop experiments in early stages of concept development and validation:

8.1.3.1. Outline of the comparison

Despite the fact that both experiments have performed human-in-the-loop experiments, the experiment objectives were also rather different as already discussed in section 8.1.1.

Whereas AGIE was focused on the identification of operational issues related to the concept of “shared separation” (i.e. investigating the potential impact of the three operational conditions on pilot/controller activity), EACAC’2000 experiments were focused on the assessment of the (somehow anticipated) impact of the concept on the controllers’ activity.

8.1.3.2. Consolidation and lessons learnt

Defining human role, working method, procedures can be successfully supported by real-time simulation as a means to explore possible solutions. As such, real-time simulation is a means to support the concept design in the early stages of development (and so, of the validation process). Nevertheless, two main issues can be raised:

- The issue of the link between the validation technique and the life cycle: is the choice of technique more related to the stage of the concept in the life cycle or to the system components impacted (e.g. involving human actors or not)?
- The issue of the required level of realism: which level of realism (i.e. involving some level of sophistication and performance of the technical components; requiring some amount of resources) is necessary and sufficient to investigate the target objectives?

Taking into account these considerations, human-in-the-loop experiments could be used in early stages of development, as far as **right balance can be found between the level of realism achievable and the level of confidence and significance of the results.**

As illustrated in the figure below, it seems that the focus on “external” validation increases with the design life cycle, whereas the focus on the “internal” validation decreases (see trapezes in the figure).

For both “internal” and “external” validation, the reliability and accuracy of the validation results, as well as the level of fidelity of the experiment, should increase throughout the life cycle.

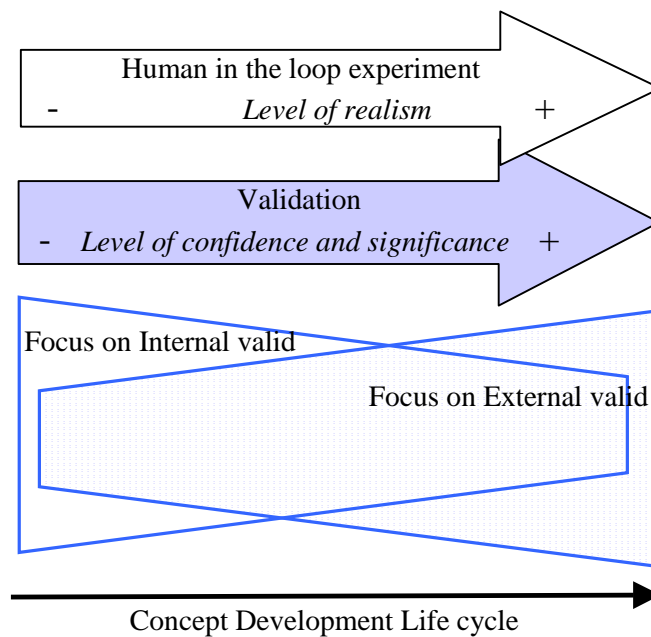


Figure 8: Role of human-in-the-loop experiment in validation

This is particularly true for concepts implying major changes in the ATM system. To get confidence in potential benefits achievable with such a concept, as well as to allow for refinement of the concept based on feedback from human actors, it should be beneficial to perform early small-scale human-in-the-loop experiments.

However, due to the limited maturity of the concept design, it would be pre-mature to invest in large-scale human-in-the-loop experiments. Indeed, this would require detailed analysis of the impact of the new concept on the ATM system, which is usually far from being possible in early stages of concept development.

8.1.4. Human-in-the loop experiment design

Based on the comparison between AGIE and EACAC experiments, some guidelines are developed, which support the definition and evaluation of metrics used during experiments in relationship with explicit validation objectives.

8.1.4.1. Outline of the comparison

Both experiments were focused on the assessment of human workload, human contribution to safety (and not technical aspects of safety) and controller/pilot strategies.

Within EACAC'2000 experiments, there was an attempt to define hypotheses about the relationship that may exist between some "internal" performance criteria, typically between controller's workload, strategies and activities. This was not the case within AGIE, which was looking for operational issues without the support of any explicit framework.

This discrepancy reflects a general feeling that within EACAC, there was a deeper analysis of the impact of delegation on of controllers' activity. This is also compliant with the difference that exists between both strategies of concept development.

Nevertheless, both experiments suffer from a lack of hypotheses about the impact of the concept on the ATM system itself, as well as its outcomes and outputs. Early and explicit formulation of such hypotheses would have lead to more efficient experiments (i.e. more positive balance between cost and benefice of the experiments).

8.1.4.2. Consolidation and lessons learnt

Starting from the "high-level" validation objectives, a top-down process dedicated to the identification of such hypotheses and then relevant metrics taking theses hypotheses into consideration should be performed. Without such a process, it is difficult to conclude on the experiment results, **i.e. to go back from the metrics measured during experiments towards "high-level" validation objectives.**

Indeed, experiment of a new operational concept requires the design of a new ATM system with potential changes in Human Commitment and Involvement (related to new roles and responsibilities), Organisations, Procedures and Rules, Infrastructure, Equipments/tools (related to new CNS/ATM functions). In addition, the experiment may also have to address potential change in the ATM system environment (e.g. traffic characteristics) at the timescale of planned/expected implementation of the operational concept.

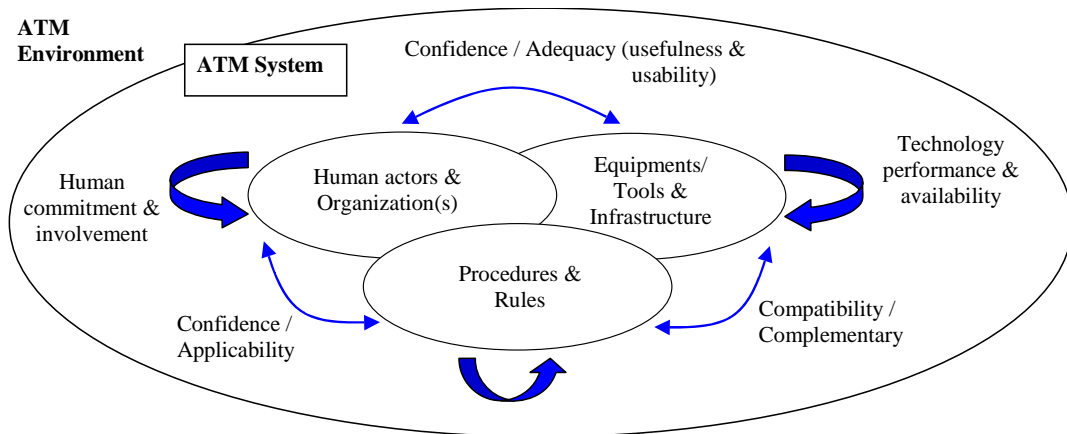


Figure 9: Model of the ATM system related to a new operational concept

To support the identification of relevant metrics and their analysis, there is **a need for a model relating human and technical components of the ATM system, their specific functions and expected performances, as well as their interactions**, as a basis for validation.

Several human performance factors influence the activity of human in positive or negative ways. The main ones are workload, confidence, motivation, skills, stress, teamwork, human error recovering.

As illustrated by the following figure, introducing a new concept possibly impacts all these factors with more or less emphasis. This impact will determine to which extent the tasks can be achieved or not by the human actors and at what cost. Finally, as part of the overall ATM system, the ability of the human to perform their missions at an acceptable cost will have an impact on the overall performance of the system.

Therefore, these human performance factors, and their inter-relationship, should be studied in light with the tasks to be achieved.

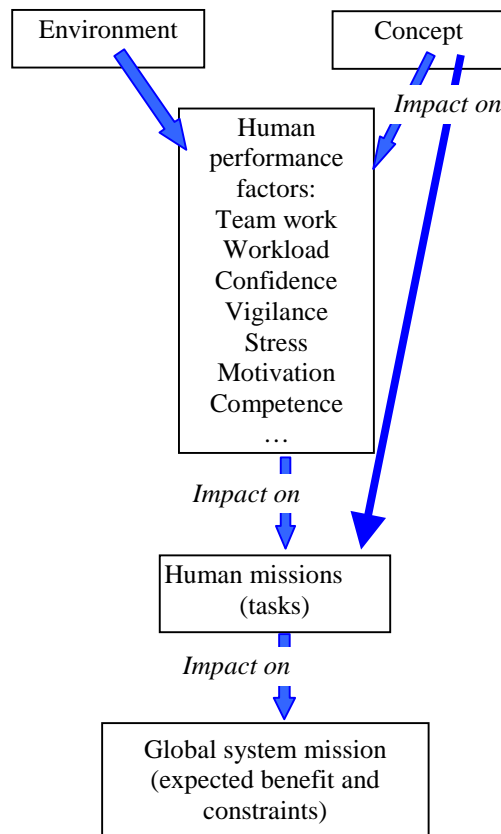


Figure 10: Model of human performance in relationship with a new concept

Therefore, to predict, analyse and validate the concept from an operational perspective, **there is a need to model human performance within the framework of the new ATM system model**, specifying the role and performance factors related to the humans (including team-work).

Such modelling should help anticipating the impact of a new concept on human tasks performance, and as such should support the formulation of relevant validation objectives (and metrics) related to human performance evaluation.

8.1.4.3. Guidelines for experiment design

Based on the experience gained from the comparison between the AGIE and EACAC experiments, methodological elements that support optimum and correct interpretation of experiment results include the followings:

- **Clear identification of the validation objectives of the experiment, and the way to conclude on these objectives, to be able to:**
 - Derive metrics from the validation objectives (top-down approach),
 - Go back from data collected to the validation objectives (bottom-up approach),
 - Possibly, relate some metrics to different validation objectives,
 - Combine some metrics to get information on complex validation objectives.

- **Clear scope of the validation activities to be able to conclude on:**
 - Assumptions and limitations (e.g. simplified ATM environment, focus on some ATM sub-system) of the experimented ATM system,
 - The robustness and quality of the results with respect to the numbers of runs, of expertise/training of the participants and of the number of conditions compared,
 - Level of significance (i.e. possible generalisation) of the experimentation results with respect to the Operational Concept under validation.

- **Well controlled dependant variables of the experiment to be able to distinguish between:**
 - Inadequate experimentation set-up (e.g. inappropriate training of participants, unrealistic ATM environment),
 - Inadequate design of the experimented ATM system (e.g. inappropriate Human/Machine interface, inappropriate Decision Support Tools), and
 - Inadequate Operational Concept (i.e. no possible evolution of the existing ATM system that would achieve the stakeholders' objectives expected from the Operational Concept under validation).

8.2. Metrics framework, characteristics and relevance

This section provides a synthetic description of the similarities and differences between the metrics used within AGIE and EACAC'2000 experiments, and draws the main lessons learnt from this comparison, in terms of:

- Scope and focus of metrics used in human-in-the-loop experiments;
- Specific metrics (and validation objectives) related to human performance;
- Applicable metrics (and validation objectives) related to efficiency;
- Applicable metrics (and validation objectives) related to safety.

Due to the nature of the AGIE and EACAC experiments, the discussion of metrics developed hereafter is focused on metrics applicable to human-in-the-loop experiments.

Capacity-related metrics are not discussed as such, since the impact on airspace capacity have only been investigated indirectly during the experiments through the metrics related to human performance, and more precisely those related to controller's workload and activity.

8.2.1. Balance between “internal” and “external” performance metrics

Based on the comparative analysis of the metrics defined within the AGIE and EACAC'2000 experiments, some general considerations about the scope and focus of the metrics applicable within human-in-the-loop experiments are provided hereafter. The need for models to allow for combination of these metrics into more consolidated indicators of either “internal” or “external” performance areas is also discussed.

8.2.1.1. Outline of the comparison

To support the identification of appropriate metrics regarding a given validation objective, the metrics used within AGIE and EACAC, including the INTEGRA metrics applied on EACAC, have been compared on the basis of their respective validation objectives (or performance areas).

AGIE and EACAC'2000 metrics

Both experiments have used a **majority of subjective data collected with non-intrusive data collection method**, including observation, questionnaires and debriefing. Such subjective metrics were not always related to validation objectives, but also aimed to support design objectives.

With regard to validation objectives, the use of objective metrics was greater within EACAC than in AGIE, typically for metrics related to workload, possibly due to the slight difference in the development stage of the concepts under validation. The main difference relies on the data processing method that involved statistical calculation within AGIE, but not within EACAC.

Regardless of the relevance of the metrics, the following figure provides an overview of the focus (i.e., number of metrics) of the validation metrics used in AGIE and EACAC experiments: a number of safety, efficiency and human-related metrics are provided in both AGIE and EACAC experiments sometimes with different scope and perspective.

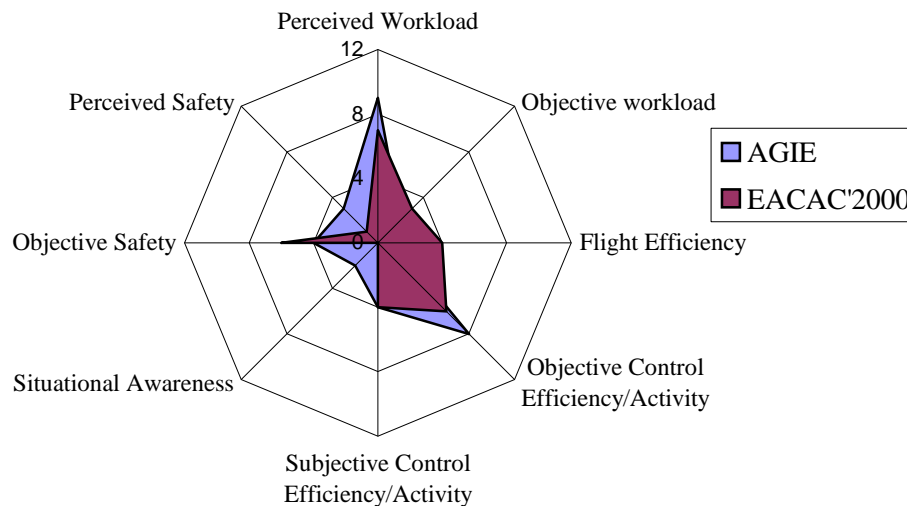


Figure 11: Number of AGIE and EACAC metrics per validation objectives

With respect to metrics related to human performance (and acceptance), both experiments defined a set of metrics to assess the impact on controller's **workload**, but only AGIE defined specific metrics to assess impact on **situational awareness**.

With respect to efficiency-related metrics, both AGIE and EACAC'2000 experiments investigated **flight efficiency** from a single aircraft perspective, whereas only AGIE investigated the impact on **control efficiency**. Actually, this was done through the use of metrics related to the controllers' activity (or strategies in handling aircraft), also used in EACAC but to further assess the impact on controller's workload (and ultimately controller's acceptability and availability).

Finally, with respect to safety-related metrics, both AGIE and EACAC have used a combination of subjective metrics to assess the participants **perceived safety**, and **objective safety** metrics mainly related to safe aircraft separations.

Both experiments suffer from a lack of hypotheses about the impact of the concept on the ATM system itself, as well as its outcomes and outputs. Consequently, all these performance factors were investigated alone, and there was almost **no attempt to correlate the metrics** related to factors that may interact with each other.

Furthermore, only judgement-based analysis of metrics measured during different operational conditions was performed. And, there were no explicit criteria to decide on whether or not the validation objectives were met.

INTEGRA metrics on EACAC'2000

The INTEGRA metrics, which aim at evaluating the performance of any ATM system, are by design only based on objective data. As such, the INTEGRA metrics are much more ambitious and requires the establishment of absolute criteria to be used as threshold for evaluation of the metrics.

The INTEGRA metrics applied on EACAC'2000 were limited to the **capacity** and **safety** metrics. In particular, the INTEGRA efficiency metrics were not implemented at the time of the INTEGRA study on EACAC'2000.

Furthermore, different factors limited the significance and interpretation of the INTEGRA metrics applied on EACAC'2000, including the difficulty in collecting appropriate objective measurements from the EACAC'2000 experiments, as well as the level of maturity of the Capacity metrics and the abstractness of the Safety metrics.

Consequently, the absolute values of the INTEGRA metrics applied on EACAC'2000 were not significant and **only relative comparisons** (with and without delegation) could be fruitful.

8.2.1.2. Consolidation and lessons learnt

As illustrated in the following figure, the scope of human-in-the-loop experiment experiments is usually focused on the assessment of human factors criteria rather than objective assessment of the outputs of human activity in the ATM system.

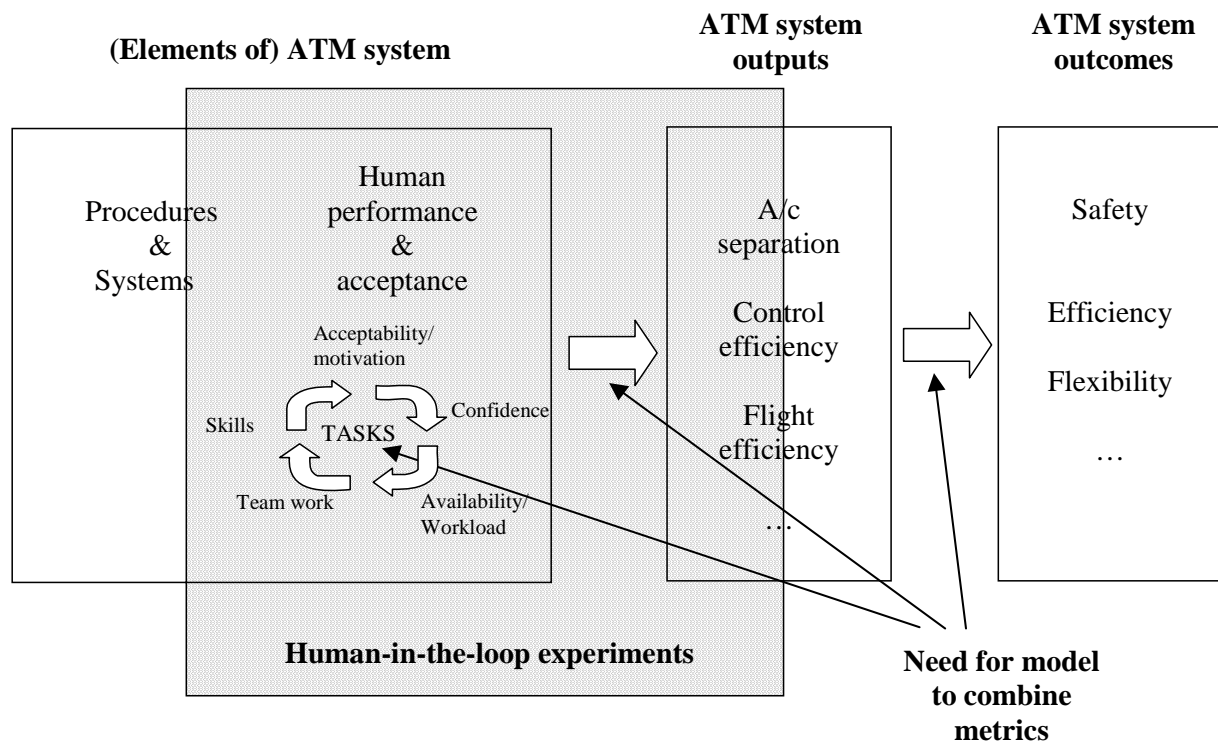


Figure 12: Scope of metrics measurable within human-in-the-loop experiments

Actually, experiments usually suffer from the lack of functional and performance modelling of the experimented ATM system. Although required to support consolidation of the metrics measurable within human in-the-loop experiments, such modelling is not easy to develop in early stages of (new) concept development. In addition, the definition and use of such models require significant amount of effort to collect and analyse data, which is not always possible during experiments.

Taking these considerations into account, it might be useful to investigate various hypotheses about the potential relationship that would exist between the level of performance from individual elements of the ATM system (typically, the human participants) and the level of performance expected from the ATM system outputs and outcomes.

More precisely, there is a need for a model describing the relationship between “internal” measurable performance metrics (related to elements of the ATM system), and “external” performance areas (e.g. safety, efficiency, capacity). Such explicit relationship would serve as a basis for validation, as an adequate framework for the identification of relevant metrics and their analysis.

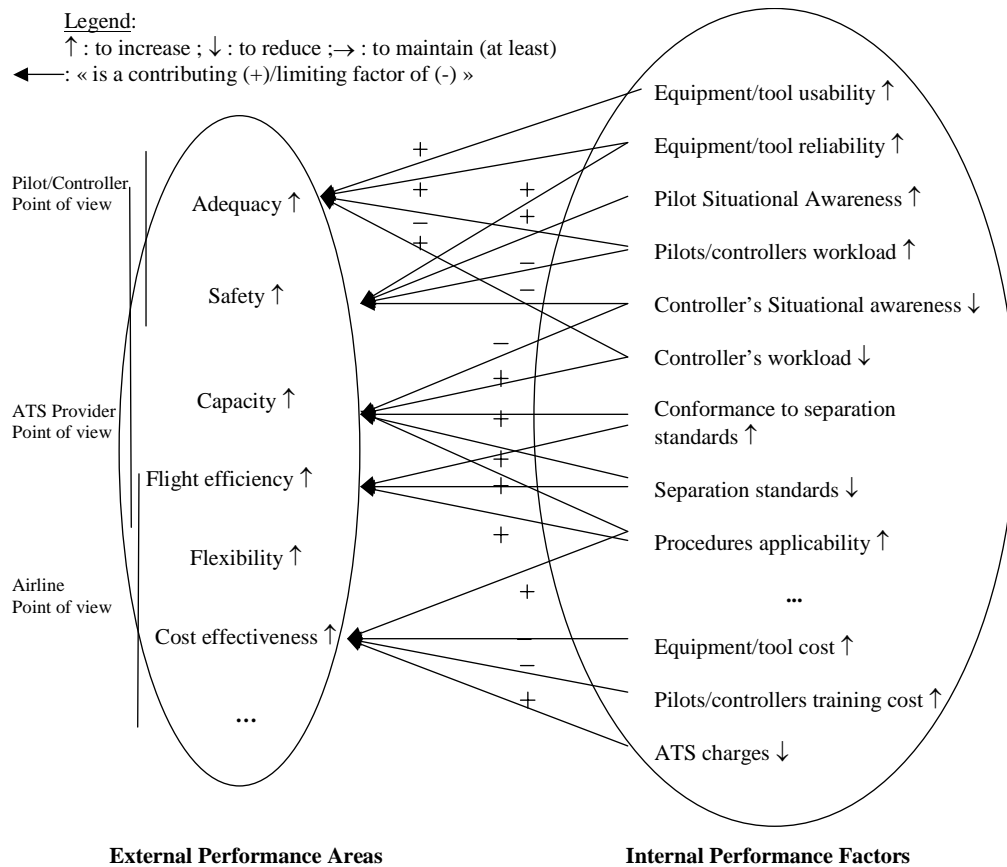


Figure 13: Illustration of the relationship between “external” and “internal” performance criteria

Another major issue is the statistical significance of the experiment results. Without sufficient amount of data, or insufficient robustness and quality of the data, it would not be possible to get confidence in the consolidated metrics derived from those experimental measurements.

In those circumstances, one should prefer the use of relative comparison of metrics rather than the establishment of absolute performance metrics, which would be unreliable and possibly misunderstood.

8.2.2. Metrics (and validation objectives) related to human performance

Based on the comparative analysis of the human-related metrics defined within the AGIE and EACAC'2000 experiments, some general conclusions about validation objectives and metrics dealing with human performance and acceptance are provided hereafter.

The INTEGRA study on EACAC'2000 also provide an opportunity to discuss the relationship between these metrics focused on “internal” performance of the ATM system, and more “external” performance metrics such as airspace capacity-related metrics.

8.2.2.1. Outline of the comparison

Both AGIE and EACAC used a majority of metrics related to **human performance and acceptance** of the concepts under validation. The various human factor criteria evaluated within both experiments included human **workload**, human **situational awareness** and human **activity**.

AGIE and EACAC metrics related to controller's performance

In general, the workload metrics used in both experiments were quite homogeneous, including both objective and subjective metrics. Subjective metrics included **participant ratings of perceived workload**. There was also an **attempt to assess workload in relation to the tasks** allocated to controllers and pilots depending on the concept under assessment. The EACAC'2000 experiments also used **physiological indicators to assess workload**, although the relationship was difficult to establish due to the limited amount of runs.

Unlike the EACAC'2000 experiments, AGIE also intended to assess impact on **situational awareness, through subjective global evaluation** by the experiment participants. Considering the major changes in controllers' and pilots' roles and responsibilities implied by the “shared-separation” concept”, the relevance of such global evaluation of situational awareness is questionable.

Finally, both AGIE and EACAC'2000 experiments defined both subjective and objective **metrics to assess the impact of the new concept on human activity**, typically controllers and pilots. Within EACAC, the purpose was to investigate the impact of the delegation concept on the current **controllers' activity** (i.e., potential changes in the cognitive process and in the realisation of their tasks), as well as the impact on controllers' strategy in handling aircraft. Within AGIE, similar (although dedicated) metrics related to pilots' and controllers' strategies for conflict management were used to identify operational issues related to “shared-separation” depending on the four conditions.

INTEGRA capacity metrics applied on EACAC

Within the EACAC'2000 experiments, as well as in the INTEGRA study within EACAC, the rationale for investigating the impact on controller workload, strategy and activity is to assess the potential impact on airspace capacity.

In particular, the INTEGRA capacity metric made a direct link between the controller's workload and the controller's activity, through the notion of control “**Information Processing Load**”. Although the INTEGRA metrics are designed to apply as absolute metrics for the evaluation of any ATM system capacity, different factors (including the need for refinement of the INTEGRA metrics themselves) limited their effective use as “high-level” performance metrics within the EACAC study.

Consequently, the INTEGRA study on EACAC'2000 could only conclude on the good responsiveness of the Capacity metrics, based on the expected assumption that delegation decreases the controller's "Information Processing Load".

8.2.2.2. Consolidation and lessons learnt

In following, consolidation focuses on controller's workload and situational awareness since they were the two main human performance factors studied within AGIE and EACAC. Other human factors (such as confidence, teamwork, motivation/acceptability, human error and error recovering) being slightly studied in one or the other experiment, it was not possible to draw lessons on all these aspects from the comparison.

Metrics related to workload (and activity)

The notion of human workload is complex covering the "task load" and the "perceived workload" (strain). As such, it is composed at least of tasks-load, time pressure, predictability of events and influenced by the experience, the teamwork quality, and individual factors.

The workload is an important factor **because it affects performance** of the subjects. Nevertheless, the relationship between workload and human performance is not obvious. For instance, operational errors have been reported under conditions of over-workload, as well as under-workload (lack of vigilance).

It means that assessing workload requires combining subjective evaluation (as feeling and reaction to a situation is specific to each individual) and objective evaluation related to tasks load.

The challenge with subjective evaluation of workload is that the human actors may understand it differently. To allow for correct interpretation of the data collected during experiments, the training (or briefing) before simulation should clearly define the notion of workload to the participants.

Metrics related to situational awareness

Situational awareness is also a complex notion, which plays a major role in the human activity. Managing situational awareness is a **continuous task** that aimed at collecting, gathering and interpreting information whose nature and detail are **related to the tasks** to be performed:

- Sub tasks of situational awareness are detect conflict, detect deviation, detect aircraft problem, be aware of the traffic load and its progress, integrate an entrance aircraft.
- Means to "build and maintain situational awareness" are for example monitoring the radar image, communicating with other controllers or pilots (to get information) and monitoring decision support tools.

According to the change the new concept provides on the human actors activity, an important issue is that the required situational awareness characteristics may change accordingly. Therefore, relevant metrics should allow checking whether or not people are able to build and maintain situation awareness **appropriate** for the tasks they have to perform.

As for all the tasks allocated to human, expectation about characteristics of the required situation representation should be defined during the design of the concept, in order to formulate validation hypotheses before simulation and to guide the identification of pertinent metrics.

Actually, rather different metrics may bring relevant information to assess situational awareness. For example, metrics related to safety could be used as indicator of situational awareness when they are related to detection of problem. Objective metrics such as eye tracking are also an efficient means to understand how situational awareness is built and updated.

Subjective metrics may also be used. Nevertheless, their interpretation should be done with caution, bearing in mind that the experiment participants will have natural difficulty to adapt their research of information required to perform their new tasks.

8.2.3. Metrics (and validation objectives) related to efficiency

Based on the comparison between AGIE and EACAC metrics, this section discusses the notion of “efficiency”.

8.2.3.1. Outline of the comparison

Both AGIE and EACAC experiments defined efficiency-related metrics, but with different scope and focus. Whereas the EACAC metrics was only addressing flight efficiency, the AGIE metrics aimed at investigating efficiency from both the ground-side and the air-side.

From the ground-side, although the AGIE targeted “high-level” validation objective was to assess **efficiency and flexibility of the airspace system**, the point of view taken was much more related to “internal” validation of efficiency from the controllers’ point of view.

On the other hand within EACAC, the metrics related to (flight) efficiency were actually misunderstood by the controllers, who evaluated the “increased efficiency” **in terms of their reduced workload**.

8.2.3.2. Consolidation and lessons learnt

Considering the various scope of the efficiency-related metrics defined within both experiments, it clearly appears that the notion of “**efficiency**” fully depends on the point of view taken.

From human evaluation point of view, efficiency would rather be related to workload (e.g. time available to detect, anticipate and plan, act and check action), motivation (e.g. to optimise trajectories), error and recovering, teamwork (e.g. compatible goals).

From airspace users point of view, efficiency would either be related to flight efficiency from an individual aircraft perspective, or control efficiency from an overall airspace perspective.

Therefore, when defining efficiency-related metrics and depending on the validation objectives, distinction should be made between efficiency of the overall ATM system, i.e., ability to provide the expected outcomes in a cost-effective manner and efficiency from the perspective of a single element of the ATM system, typically the human involved.

8.2.4. Metrics (and validation objectives) related to safety

Finally, the comparative analysis of the AGIE and EACAC metrics related to safety, including the INTEGRA safety metrics applied on EACAC'2000, provide an opportunity to discuss the kind of safety assessment that could be performed within human-in-the-loop experiments.

8.2.4.1. Outline of the comparison

Both AGIE and EACAC'2000 experiments defined safety-related metrics, but with rather different scope and focus. The EACAC safety metrics consisted in a combination of subjective and objective metrics, whereas AGIE metrics were focused on subjective assessment of **perceived safety**. In both experiments, the objective assessment of safety was done through metrics related to **losses of aircraft separation**.

Although sometimes addressed by the experiment participants in their (textual) answers to the questionnaires, the relationship between these "internal" and "external" perspectives of safety were not investigated directly through the use of dedicated metrics.

The INTEGRA safety metrics applied in EACAC'2000, which provide by design objective assessment of the ATM system, were an attempt to do so. In particular, the "**resilience**" metrics aimed at assessing the extent to which the ATM system responds to a safety significant event, taking into account the amount of control "information processing load" required to deal with these events.

However, the level of maturity of the INTEGRA metrics at the time of the EACAC'2000 experiments did not allow their effective use as "high-level" safety performance indicators.

8.2.4.2. Consolidation and lessons learnt

In the ATM system, the **human remains responsible for the safety** of air traffic operations. Evaluating safety from an operational (or human) perspective implies to explicit the safety missions of the human (e.g. avoid mid-air collision, comply with aircraft separation minima, avoid violation of restricted area), so as define appropriate metrics to assess the human contribution to safety.

To which extent the human are able to achieve the safety missions is under the influence of their workload, situation awareness, confidence, skill and experience in both nominal and non-nominal situations. Once again, these relationships should be more explicit to support evaluation of safety through human-in-the-loop experiments.

In addition, in view of the limited amount of data collected through experiments, the safety-related metrics should focus on the identification of potential issues related to inadequate procedures, human training or decision support tools.

In any case, safety related results obtained during human-in-the-loop experiments would not permit to conclude on the achieved level of safety. However, **some trends about potential safety issues or factors supporting safety** might be identified.

9. CONCLUSIONS

The comparison of EACAC and AGIE metrics has been performed in the context of harmonisation of validation work, and to support exchange of information, between US and Europe within the framework of FAA/Eurocontrol Action Plan 5.

During the CEAMCA study, independent team performed analyses of both AGIE and EACAC experiments and metrics. The study identified the main characteristics of the metrics and their context of use in both experiments, as well as in the INTEGRA study applied on EACAC.

The comparison of AGIE and EACAC metrics highlighted the difficulty to compare metrics related to “internal” validation (i.e., related to performance factors of ATM system elements). Indeed, such metrics are quite specific to the concept under assessment, and the functional and performance characteristics of the elements the ATM system is composed of, typically the human actors.

The issues raised throughout the comparison are not specific of the studied experiments, stress general issues of human performance validation and the dilemma between:

- The expectations and requirements of ATM stakeholders regarding the quality of human performance validation results, and the ability to conclude from these results about the potential improvements achievable by the operational concept under assessment, and
- The requirements towards the ATM R&D regarding the development of models, and the amount of resources to be allocated to experiments (i.e., to collect and analyse data and to control dependant variables)

Meanwhile addressing these issues, the establishment of a common framework of metrics may have to focus on “external” metrics (as envisaged by the INTEGRA metrics), at least based on the current knowledge and modelling of the ATM system.

In the perspective of comparing metrics, it was considered necessary to also compare the concept under assessment, as well as the validation approach and objectives sustaining the experiments. Indeed, the comparison of metrics was made, as far as possible, on the basis of their validation objectives, rather than the element of the ATM system on which they fall on.

Although the study highlighted the difficulty to perform a one-to-one comparison of metrics, it should be possible to develop common framework supporting the identification of relevant metrics and their interpretation during experiments.

In this respect, the comparison between AGIE and EACAC experiments allowed highlighting some general guidelines for experiment design in line with MAEVA and VALSUP, together with elements for consideration when defining the validation objectives and metrics for human-in-the-loop experiments.

These considerations constitute a first step toward the definition of a common framework for metrics definition. Nevertheless, due to the limited scope of the study, further work would be required to refine the set of relevant performance areas and associated metrics measurable through experiments.

10. FUTURE WORK

The added value brought by the FAA support along the performance of the CEAMCA project has underlined the interest of an extended partnership between US and Europe in the performance of possible follow-up actions.

These actions may come within the scope of future activities of the FAA/Eurocontrol Action Plan 5, in charge of developing a validation and verification strategy spanning R&D activities and implementation. They would aim at exploiting and consolidating the CEAMCA outputs by:

- I. Further consolidation and validation of the CEAMCA framework for metrics comparison, for instance through its comparison with the framework supporting the Validation Data Repository or the Validation Master Plan;
- II. Further development of the guidelines for experiment design identified through the CEAMCA study to support the identification of relevant metrics and their interpretation;
- III. In depth investigation of the role of experiments in early stages of development and validation, especially with regards to the following questions:
 - Is the choice of the technique more related to the stage of the concept life cycle or to the elements of the ATM system impacted by the concept?
 - Which level of realism is necessary and sufficient to investigate depending on the validation objectives?
- IV. Further investigation of the need for models (at various levels) to support the consolidation of experiments design and results, in particular with regards to human performance, but not only:
 - A model defining the role of the human performance factors in the performance of the human missions (and tasks), as well as the inter-relationship between these human performance factors;
 - A model defining the relationship between “internal” measurable performance criteria (related to human and technical elements of the ATM system) and “external” performances areas.

ANNEX A ANALYSIS OF THE AIR-GROUND INTEGRATION EXPERIMENT

This annex contains the detailed analysis of AGIE [1], performed using the framework described in 3 which addresses:

- The operational concept under validation,
- The validation approach and objectives,
- The experiments scope, objectives, and finally
- The indicators, metrics and measurements.

A.1 Outline of the concept under validation

A.1.1 Concept studied and its rational

A.1.1.1 Concept of operations

The concept studied is in the framework of the “free flight” concept: it suggests shifting aircraft separation responsibility from air traffic controllers to flight crews. This creates a ‘shared-separation’ authority environment.

Three operational concepts have been defined on the basis of “free-flight”. Their characteristics are summarised in the following table. (An operational concept is named condition in the following of the document).

The first condition presented in the next table represents the current operational situation. It has been used as the reference (baseline) during the experiment.

Characteristics	CO (Current Operation)	CO: CDTI	SS: L1 (Shared Separation Level 1)	SS: L2 (Shared Separation Level 2)
Separation standards of 5 nm horizontal or 1000/2000 ft vertical	X	X	X	X
URET was available to controllers	X	X	X	X
Controllers coordinated URET red alerts with other sectors	X	X	X	X
Controllers had full separation responsibility	X	X		
Pilots required to <i>request</i> clearance from controllers prior to manoeuvring	X	X		
CDTI-AL was available to pilots		X	X	X
Pilots and controllers shared-separation responsibility			X	X
Air↔air frequency was available			X	X
Pilots used right-of-way rules while resolving potential conflicts			X	X
Pilots could cancel free flight			X	X

Characteristics	CO (Current Operation)	CO: CDTI	SS: L1 (Shared Separation Level 1)	SS: L2 (Shared Separation Level 2)
Controllers could cancel free flight			X	
Pilots could initiate any manoeuvre but were required to first <i>inform</i> controllers prior to manoeuvring			X	
Controllers were required to issue traffic alerts to aircraft concerning URET red alerts			X	
Pilots <i>did not have to inform</i> controllers prior to manoeuvring				X

Table 18: AGIE Experimental Condition Characteristics Summary

A.1.1.2 Operating environment characteristics

The concept studied within AGIE is focused on High altitude En route airspace.

Main CNS/ATM assumptions related to the targeted operational environment include:

- Change on controller working positions: Display System Replacement System (DSR) consoles, no paper strips, and the integration of the User Request Evaluation Tool (URET) for the controllers,
- Change on pilot flight deck: Cockpit Display of Traffic Information (CDTI), airborne alerting logic added to cockpit automation, ADS-B technology included for data-link communication of aircraft parameters,
- No controller-pilot data-link communications,

A.1.1.3 Expected benefits and constraints (“high-level” objectives of validation)

“Free flight” is intended to provide:

- Increased flexibility and efficiency throughout the global airspace system (i.e., more flexibility to manage flight operations)
- Improved safety through enhanced conflict detection and resolution capabilities and redundant traffic monitoring procedures.
- Improved decision making thanks to better decision-making tools for air-traffic controllers and flight crews

A.1.2 Stage in the life cycle

According to FAA/EUROCONTROL MoC on R&D: Action Plan 5: Operational Concept Validation Strategy Document, AGIE experiment takes place in the first stage of design: “Development of the operational concept specification”.

According to EMERALD RTD (Research and Technical Development) Plan, AGIE takes place in the “User Requirement or Concept Phase”.

A.2 Outline of the validation approach

A.2.1 Stage in validation (possibly per validation objectives)

Prior to AGIE, the concept of “shared-separation” had started to be evaluated separately at NASA and the FAA. A variety of studies have examined the effect of free flight operations on controllers and pilots from 1997 to 2000. These studies have been done on air issues, ground issues and supported tools individually. Then, there was a need to investigate how all the elements might work together in a shared separation environment.

The Air-Ground Integration Experiment provided an initial examination of the effect of shared-separation authority on flight operations when both air and ground have enhanced traffic and conflict alerting systems. This study is considered as a preliminary investigation, and the results should not be generalised or accepted as conclusive.

As such, AGIE takes place in validation stage 1 (V1): agreement about “Basic principles of a new concept”.

A.2.2 Validation objective(s)

According to the ATM 2000+ validation objectives, the validation objectives of AGIE were mostly to assess:

- Human involvement and commitment (mainly, procedures recommendations⁸)

And, to a lesser extent, to assess impact on:

- Safety
- Economics: flexibility and efficiency of the Airspace System

A.2.3 Validation point of view

The point of views taken in the validation exercises are mostly that of:

- Air Traffic Controllers, and
- Flight crews.

A.2.4 Object(s) under validation

The validation was focused on the assessment of:

- The use of shared separation operations under nominal conditions.

A.2.5 Validation activity (or technique as named in MAEVA)

The validation activity/technique was:

- Real-time simulation concerning both air and ground side.

⁸ At this early stage of validation, there was no attempt to evaluate or validate the procedures created for the experiment. The objective was to provide recommendations to the FAA for procedures should such a concept be implemented.

In summary,

Validation objectives	Validation phase	Validation point of view	Object under validation	Validation activity (or technique)
Human involvement and commitment (mainly, procedures recommendations) And, to a lesser extent, to assess impact on: Safety, Economics flexibility and efficiency of the Airspace System	V1: Basic principles of a new concept are agreed.	Air Traffic Controllers, and Flight crews	The use of shared separation operations under nominal conditions (focusing on human factors issues and looking for recommendations for procedures, and decision support tools)	Real-time simulation involving controllers and pilots

A.3 Experimentation of the concept

A.3.1 Experiment characteristics

A.3.1.1 Scope of the experiment

The main assumptions and limitations of the experimented ATM system and environment simulated within AGIE include the followings:

Airspace characteristics

- Two adjacent en-route sectors (high altitudes).

Traffic characteristics

- Traffic samples derived from two traffic recordings,
- Resulting traffic close to moderate to high-density traffic,
- Traffic flows preserved, but Planned conflicts were created (eight 2 a/c conflicts of similar complexity were planned for each sector),
- Full (CDTI) equipped traffic (thus allowing for the maximum opportunities for shared separation operations).

Controller working environment

- No paper strips,
- Instead, use of the User Request Evaluation Tool (URET), which assists the controller in predicting and evaluating potential conflicts. At the time, it was installed as a prototype system in the real operational environment. It provides the controller 5 levels of automated problem detection alerts with a look-ahead time of 20 minutes.
- High fidelity Display System Replacement (DSR) controller positions.

Cockpit display of traffic information

- High fidelity cockpit simulator,
- Use of a prototype CDTI (including new display components and input control devices) with a prototype airborne alerting logic designed to help flight crews manage the more strategic shared-separation responsibilities. The alerting logic automation is based on the expected ADS-B capabilities that include extended traffic depiction and improved update rates for the navigation display. It overlays the existing TCAS logic and provides an additional alerting zone beyond that of TCAS. (A CDTI-AL alert was triggered for the flight crews when the logic predicated a pending violation of the protected zones (or minimum separation requirement)).

A.3.1.2 Experimental conditions

The main characteristics of AGIE set-up and performance include the followings:

- AGIE was conducted concurrently at the WJHTC on the east coast and NASA ARC on the west coast. Therefore, simulation procedures were conducted simultaneously at both locations.
- Each group of controllers and the Expert Observers (EO) participated for a 3-day simulation period. Each flight crew participated in the study for two 8-hour days. Four groups (of controllers, EOs, flight crews) participated in total.
- Participant training to procedures and airspace (during the first day and a half of simulation).
- The three operational conditions are compared to the current one: All participant groups (controller and pilot) participated in all four runs. The order of condition presentation was counter-balanced across the four data collection groups.

Experiment participants

- Number of participants: 12 experienced controllers and 6 experienced pilots, (all controllers were familiar with the simulated airspace and all the pilots were rated for the aircraft).
- Two controllers (planning and tactical) per measured sector,
- Two members of the simulation experiment team staffed the “ghost sector” controller position for all adjacent, non simulated sectors; one member staffed the automatic data-link operator position during specific runs (update the host computer so that CDTI AL and URET (respectively air and ground decision support tools) remained current and consistent with flight plan updates),
- Ten pseudo pilots; one intruder simulation pilot staffed the intruder aircraft that was scripted to be involved in planned conflicts.

Data collection during experiments

Subjective and objective data were collected throughout the study from the ground-side (participant controllers, EOs, and the ATC environment), and the air-side (pilot participants and the flight deck).

1) Ground side Data:

- Subjective data were collected from controllers and Experiment Observers: before the training phase form, during-the-Run forms, Interval Workload Ratings, Post-Run Form, Exit Form, De-briefing.
- Objective data related to URET alerts and trial plans, voice communication data, Host data, and audio and video data were collected:

2) Air side data

- Subjective data were collected from the individual pilot participants primarily through post-run and exit forms. Additional data were collected during a debriefing session and workload data were collected following each flight segment.
- Air-Side Objective Data: from the NASA ARC simulator computer systems, including data on use of various aircraft flight systems, use of flight deck display controls, and communication equipment. Video and audio recordings of flight crew interactions on the flight deck were also collected.

Experiment limitations

The following describes recognised limitations and constraints of AGIE:

- This study had a limited number of participants and therefore limited power for the use of inferential statistics.
- Traffic scenarios: Due to time constraints and the unexpected complexity of the design process, only two data collection traffic scenarios were created. (Aircraft call signs and the destination airports of conflict aircraft were changed to create four unique data collection runs). In the exit forms, participants reported that the runs were familiar.
- The aircraft simulator used for this study was a Boeing 747-400. This aircraft is typically used for long-haul oceanic flying. The flight crews used were those qualified on this aircraft type to insure minimal training. The use of commercial pilots who typically flew long oceanic routes may have affected the flight crew results.
- Due to relatively short flight segments (20 minutes), most air-side efficiency measures were not possible to analyse.
- Some of the aircraft entered the simulation too close to the sector boundary. This was a technical limitation in the laboratory.
- The controllers were able to distinguish the pilot participants from most of the simulation pilots due to phraseology, style of communications, and clearer frequencies from the simulation pilot laboratories.

A.3.1.3 Experiment objectives (“low level” objectives of validation)

This experiment provided an initial examination of the effect of shared-separation authority on flight operations when both air and ground operators have enhanced traffic and conflict alerting systems. There was a strong emphasis on identifying and evaluating human factors issues. The specific objectives were the followings:

- To identify operational issues (e.g., communications and procedures) that affect shared-separation operations,
- To provide recommendations for the information requirements and procedures necessary to facilitate shared-separation operations.
- To evaluate the effect of shifting separation authority on controller and pilot workload and situation awareness.

Note: The second objective is of “internal objectives” type, i.e., objectives related to the design in its global sense.

In summary,

Scope of experiment (and limits)	Experimental conditions	Experiment objectives
<p>Limited number of participants, four unique data collection runs based on two traffic scenarios.</p> <p>The aircraft simulator used for this study was a Boeing 747-400.</p> <p>Due to relatively short flight segments (20 minutes), most air-side efficiency measures were not possible to analyse.</p> <p>Some of the aircraft entered the simulation too close to the sector boundary.</p> <p>The controllers were able to distinguish the pilot participants for most of the simulation.</p>	<p>AGIE was conducted concurrently at the WJHTC on the east coast and NASA ARC on the west coast.</p> <p>Four sessions (of one week each) between November 1999-February 2000.</p> <p>Each group of controllers and the Expert Observers participated for a 3-day simulation period.</p> <p>Each flight crew participated in the study for two 8-hour days.</p> <p>Participant training to procedures and airspace.</p> <p>Three operational conditions to be compared to the current one: Current Operation (CO); CO + CDTI; Shared Separation Level 1; Shared Separation Level 2.</p>	<p>To identify operational issues (e.g., communications and procedures) that affect shared-separation operations.</p> <p>To provide recommendations for the information requirements and procedures necessary to facilitate shared-separation operations.</p> <p>To evaluate the effect of shifting separation authority on controller and pilot workload and situation awareness.</p>

Table 19: Outline of the Air-Ground Integration Experiment

Description of each experiment objective with:

- Hypotheses to be validated (or invalidated) through the experiment (either implicit or explicit),
- Related expected benefits (or constraints), or
- Related ATM system element under validation.

The following tables concern the ground-side analysis and the integrated ground and air sides data analysis. As mentioned previously, information that pertained only to the air side analyses was not included in this comparison report.

Note: The hypotheses have been inferred from the chapter 3 – Results of the AGIE final report. The document does not mention the hypotheses explicitly. For example, there is no assumption on the result (an assumption could be: we expect more increase of workload in condition A than in condition B). Similarly, the relationship between validation objectives and hypotheses has been inferred during the review (no explicit link in the report).

To identify operational issues that affect shared-separation operations		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation
The three operational conditions tested have an impact on shared separation operations regarding: <ul style="list-style-type: none"> - The perceived amount of time available to assure safe a/c separation and required co-ordination - The perceived level of safety - The frequency and duration URET alerts - The minimum separation distance and cancellation of free flight - Pilot and controller manoeuvre strategies for conflict resolution. 	Safety Safety Efficiency and Flexibility of the Global Airspace System Safety Human Involvement and Commitment & Efficiency and Flexibility of the Global Airspace System	Procedural recommendations and decision support tools Procedural recommendations and decision support tools Procedural recommendations Procedural recommendations an decision support tools

To provide recommendations for the information requirements and procedures necessary to facilitate shared-separation operations		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation
The three operational conditions tested have an impact on the information required and procedures (respectively for PC and TC) regarding: <ul style="list-style-type: none"> - The amount of information available to resolve a conflict - The adequacy of the timing of conflict probes - The usefulness of air-air communication monitoring - The helpfulness of the shared separation concept - The cancellation of free flight 	Safety Safety & Efficiency and Flexibility of the Global Airspace System Safety	Procedures, Data information & decision support tools

To evaluate the effect of shifting separation authority on controller and pilot workload and situation awareness.		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation
The control conditions (i.e., the three operational conditions tested) have an impact on the workload and situational awareness (respectively for PC and TC, and captain and first officer)	Human Involvement and commitment	Procedures, Data information & decision support tools

Table 20: Description of each experiment objective of AGIE

A.3.2 Indicators, Metrics and measurements

Within the AGIE final report, the metrics are classified per ATM system on which the metric is performed (i.e., ground or air side), and then by related experiment objectives.

The synthesis of airside related metrics is not reported in this document, as the EACAC’2000 experiments have not performed airside metrics analysis. As such, metrics comparison related to the flight deck’ perspective cannot be done.

Note: In the following review, metrics having same characteristics (i.e., attributes, related objectives and means of calculation) were gathered into the same tables.

A.3.2.1 Pre-simulation questionnaire

After an initial briefing, a questionnaire was proposed to get information about the background of the controllers involved in AGIE.

A.3.2.2 Metrics related to identification of operational issues

Both subjective and objectives metrics were used to identify operational issues that affect shared separation operations depending on the three operational conditions under assessment.

Metrics names:			
- Controller/pilot ratings for the time available to assure safe aircraft separation			
- Controller/pilot ratings for the amount of time available for co-ordination and communication tasks			
Attributes	Related experiment objective(s)/hypothesis	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To identify operational issues that affect shared-separation operations: through the time available	Judgement-based analysis based on comparison between the four conditions, and the actors	Post run form question Analysis of data: Variance test (ANOVA)

Metric name: Controller/pilots Ratings for the Level of Safety for Procedures			
Attributes	Related experiment objective(s)/ hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To identify operational issues that affect shared-separation operations: through the perceived safety	Judgement-based analysis based on comparison between the four conditions, and between controllers and pilots responses	Post run form question Analysis of data: Variance test (ANOVA)

Metrics names: -Mean frequency of URET conflict alerts (per run) -Mean duration of URET conflict alerts (per run)			
Attributes	Related experiment objective(s)/hypothesis	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Non intrusive Not binary	To identify operational issues that affect shared-separation operations through the number and duration of URET alerts: safety and efficiency point of view	Comparison between the three operational conditions taking account of the type of alert and of the controller position	Collect of data and computation: number of alerts/run Analysis of data: SEM, ANOVA. For duration, a Tukey HSD post-hoc comparison was used

Metric name: Loss of Separation for Conflicts			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary Not intrusive	To identify operational issues that affect shared-separation operations in terms of separation minima	Criteria = minimum separation defined as 5nm horizontally or 1000/2000 ft vertically Comparison between three operational conditions	Collect of data and computation: number of separation violation (simulator output) and observation + video recording Analysis of data: replay of video

Metrics names: - Descriptive Statistics for Altitude-Resolved Planned Conflicts - Descriptive Statistics for Vector-Resolved Planned Conflicts			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Non intrusive Not binary	To identify operational issues that affect shared-separation operations through: - The minimum separation distance, and - Cancellation of free flight	Judgement based analysis based on statistical comparison between the four conditions Comparison between pilots and controllers strategies	Collect of data and computation: means and SD (standard deviation) through ANOVA for: 1) Pilots (only) resolve conflict 2) Free flight cancelled 3) Controller (only) resolve conflict

Metrics names: - Mean frequency of air↔ground transactions (*) - Mean duration of air↔ground transactions (*)			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary Non intrusive	To identify operational issues that affect shared-separation operations	Judgement-based analysis based on Statistical Comparison between the relevant conditions	Collect of data and computation: means frequency and SD of the transactions in each condition. Frequency: summation of all the transactions during the three flight segments for each condition. Mean frequency: for each condition = average of the three flight crew frequencies.

(*) Transaction was defined as all communications initiated by a controller or pilot participant including acknowledgement.

Transaction duration was measured from the beginning of the first instruction, question, or comment made by any of the controller or pilot participants to the end of the final communication on the topic.

Metrics names: - Frequency and Type of Manoeuvres Issued by Controllers/pilots to Resolve Conflicts - Combination of Manoeuvres Issued by Controllers/pilots to Resolve Conflicts			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary Non intrusive	To identify operational issues that affect shared-separation operations through the relation between the initiator and the manoeuvre strategies	Judgement-based analysis based on Comparison of strategy according to the initiator	Collect of data and computation: computation of types of manoeuvres initialised by the controllers: heading, altitude, speed (Idem for the pilots)

Metric name: Controller Conflict Detection and Resolution Measures			
Attributes	Related experiment objective(s)/hypothesis	Decision criteria	Means of calculation (or measurements)
Subjective Quantitative Intrusive Not binary	To identify operational issues that affect shared-separation operations through the relation of the initiator and conflict detection and resolution strategies	Judgement based analysis based on the comparison between the four conditions	During the run observer questionnaire; collected controller conflict detection times and times when controllers would have started conflicts resolution if a/c were not self-separating (**)

(**) Results and analyses of the metric were not reported within the AGIE final report due to data precision issues.

Metric name: Controller Role and Separation Responsibility Confusion			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Binary Not intrusive	To identify operational issues that affect shared separation operations through responsibility awareness	Judgement based analysis	Post simulation form (scale with 2 options); percentage

Metric name: Participant (pilots) rating of flight efficiency			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To identify operational issues that affect shared separation operations through flight efficiency	Judgement based analysis based on: Pair comparison of the conditions in terms of flight efficiency	Post-run questionnaire Using AHP method (Analytical Hierarchy Process)

Metric name: Fuel burn			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
N/A	N/A	N/A	(**)

(**) Due to short flight segments and different routes through the sectors, fuel comparison could not be made. Therefore, the metric description was not available in the AGIE report.

A.3.2.3 Metrics related to information requirements and procedures

Subjective feedback from the AGIE participants was collected, in order to get recommendations for the information requirements and procedures.

Metrics names: - Controller mean ratings of information to resolve conflicts - Controller mean ratings for URET conflict alert timeliness - Usefulness and frequency of air-air communication monitoring - Helpfulness of the shared separation concept			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To provide recommendations for the information requirements and procedures	Judgement based analysis based on: - Statistical Comparison between the four conditions - Comparison between information required for PC and TC	Post-run questionnaire Analysis of data: ANOVA.

A.3.2.4 Metrics related to Workload and Situational Awareness

Mainly subjective metrics, together with some objective ones, were used to evaluate the effect of shifting separation authority on participants' workload and Situational Awareness.

Metrics names: - Controller Ratings for Physical, Mental, and Overall Workload - Controller Workload Ratings for Maintaining Aircraft Separation, - Controller Ratings for Land Line Coordination, - Controller Ratings for R-Side-to-D-Side Coordination, - Controller Ratings for Ground→Air Transmissions, - Controller Ratings for URET Coordination - Controller Ratings for Feeling Rushed and Bored			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To evaluate the effect of shifting separation authority on controller's workload	Judgement based analysis based on Statistical Comparison between the four conditions And comparison between pilots and controllers rating	Post-run questionnaire Analysis of data: ANOVA.

Metric name: Controller Interval Workload Ratings (WAK)			
Attributes	Related experiment objective(s)/hypothesis	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Intrusive Scale with 5 Choices	To evaluate the effect of shifting separation authority on instantaneous workload.	Judgement based analysis based on Statistical Comparison between the four conditions	During the run electronically recorded data (WAK); computation of the mean. Analysis of data: ANOVA.

Metric name: Expert Observer Ratings of Controller Physical Task load			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Subjective (expert observer) Qualitative Non intrusive Scale with 5 options or open questions	To evaluate the effect of shifting separation authority on workload.	Judgement based analysis based on Statistical Comparison between the four conditions	Post-run form for expert observer; computation of the mean - 6 open questions for comments Analysis of data: ANOVA.

Metrics names:			
- Frequency of controller Ground→Air and Land Line Push-to-Talk Transmissions - Duration of controller Ground→Air and Land Line Push-to-Talk Transmissions			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not intrusive Not binary	To evaluate the effect of shifting separation authority on workload	Judgement based analysis based on Statistical Comparison between the four conditions	Collect of data and computation: computation of frequency and duration of communications + SEM through ANOVA.

Metric name: Controller Ratings of Overall Situation Awareness			
Attributes	Related experiment objective(s)/hypothesis	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Non intrusive Question with a scale of 5 choices	To evaluate the effect of shifting separation authority on controller situation awareness	Judgement based analysis based on Statistical Comparison between the four conditions And comparison between pilots and controllers rating	Post-run questionnaire Analysis of data: ANOVA.

A.3.2.5 Post- simulation questionnaire

General feeling of the participants about the adequacy and realism of the simulation was collected (through post-simulation forms) at the end of the experiment.

Metrics names: - Controller rating of realism of the simulated flight crew responses - Participant rating of overall realism of the simulation - Participant rating of adequacy of simulation training			
Attributes	Related experiment objectives/hypotheses	Decision criteria	Means of calculation (or measurements)
Qualitative Subjective Not Binary Not intrusive	<i>No explicit relation to one of the 3 high-level objectives</i>	Judgement based analysis	Post simulation form (scale with 5 options); percentage

ANNEX B ANALYSIS OF THE EACAC'2000 EXPERIMENTS

This annex contains the detailed analysis of the EACAC'2000 experiments [2], performed using the framework described in 3 which addresses:

- The operational concept under validation,
- The validation approach and objectives,
- The experiments scope, objectives, and finally
- The indicators, metrics and measurements.

B.1 Outline of the concept under validation

B.1.1 Concept studied and its rationale

B.1.1.1 Concept of operations

The concept studied with EACAC consists in limited delegation of separation tasks (i.e., implementation and monitoring tasks) from controllers to flight crew through:

- Sequencing applications in extended terminal areas (from cruise to initial approach fix), and
- Crossing and passing applications in en-route airspace.

The concept is based on two key principles related to human and technology aspects:

- Minimum change in current roles and working methods of controllers and flight crews,
- Keep technology as simple as possible.

The concept implies the use of new ATC instructions upon controllers' initiative.

Level of delegation	Sequencing applications	
	In-trail	Merging
Report separation	<i>Report in-trail</i>	<i>Report merging distance</i>
Maintain separation	<i>Remain behind</i>	<i>Merge behind</i>
Resume then maintain separation	<i>Heading then remain behind</i>	<i>Heading then merge behind</i>

Level of delegation	Crossing / passing applications	
	Lateral	Vertical
Report separation	<i>Report clear of target</i>	<i>Report clear of target</i>
Maintain separation	<i>Resume navigation</i>	<i>Resume climb</i>
Provide separation	<i>Pass behind</i>	<i>Pass below / above</i>

Table 21: EACAC Operational Procedures Summary

B.1.1.2 Operating environment characteristics

The concept is intended to apply within both extended terminal areas and en-route airspace in Europe core area under radar control services. No explicit timescale (for traffic forecast) was anticipated within the EACAC'2000 experiments.

The main CNS/ATM assumptions related to the targeted operational environment include:

- No (major) change on controller working positions,
- No controller-pilot data-link communications,
- No data-link communication of aircraft parameters (e.g. no “intent” information from aircraft),
- No automation on board (e.g. no coupling to the autopilot nor to the FMS),
- Cockpit display of traffic information with display cues (to achieve selected separation from selected aircraft).

B.1.1.3 Expected benefits and constraints (“high-level” objectives of validation)

Expected benefits from the operational concept under assessment include the followings:

- Increase in airspace capacity through increase of controller’s availability (mainly in terms of mental workload) to handle more traffic, depending on airspace conditions, airspace constraints and practice level.
- Increase in safety (or at least, the same as today) through better organisation of tasks, redundant separation monitoring ensured by both controllers and flight crews, and possible increase in flight crew’s situational awareness.
- Increase in flight efficiency through anticipation of trajectory changes (less time-critical instructions to follow) and optimisation of trajectories.

B.1.2 Stage in the life cycle

According to FAA/EUROCONTROL MoC on R&D: Action Plan 5: Operational Concept Validation Strategy Document: Stage 1: “Development of the operational concept specification”.

According to EMERALD RTD (Research and Technical Development) Plan, EACAC takes place in between:

- “User Requirement or Concept Phase”, and
- “User Requirement Analysis or Feasibility Phase”.

B.2 Outline of the validation approach

B.2.1 Stage in validation (possibly per validation objectives)

Initial feedback (i.e., qualitative indications through questionnaires) about the operational concept was already obtained from both controllers and flight crews during previous experiment (June 1999) conducted in a simple environment. This initial validation (phase V1) was focused on operational feasibility and potential interest, and resulted in refinement of some applications.

The purpose of the year 2000 experiments was rather to perform an “initial proof of the concept (V2)” through real-time experiments conducted in more realistic and comprehensive operational environment.

B.2.2 Validation objective(s)

The validation objectives of the EACAC’2000 experiments were mostly to assess:

- Human involvement and commitment

And, to a lesser extent, to assess impact on:

- Safety
- Capacity
- Economics: flight efficiency.

B.2.3 Validation point of view

The point of views taken in the validation exercises were mostly that of:

- Air Traffic Controllers, and
- Flight crews (to a lesser extent).

Note: The EACAC'2000 final report focused on the controller's point of view. Feedback from pilots was covered in a separate report not reviewed in this study.

B.2.4 Object(s) under validation

The validation was focused on the assessment of:

- The use of the delegation procedures under nominal conditions.

And to a lesser extent,

- The (new) controller working position.

B.2.5 Validation activity (or technique as named in MAEVA)

The validation activity/technique was:

- Small-scale real-time simulation focused on the controllers' side and with some insight into the airborne side (in November only).

In summary,

Validation objectives	Validation phase	Validation point of view	Object under validation	Validation activity (or technique)
Human involvement and commitment And to a lesser extent, Safety, Capacity, Economics: flight efficiency	Initial proof of the concept (V2)	Air Traffic Controllers, and Flight crews (to a lesser extent)	Procedures, CWP.	Small-scale real-time simulation

B.3 Experimentation of the concept

B.3.1 Experiment characteristics

B.3.1.1 Scope of the experiment

The main assumptions and limitations of the experimented ATM system and environment simulated within the EACAC'2000 experiments include the followings:

Airspace characteristics

- Two distinct airspace organisation (one for extended TMA and another for en-route),
- For each airspace organisation, four existing sectors from the Paris area were combined into two measured sectors (allowing for reduced need for ATC co-ordination),
- For each airspace organisation, two feed sectors were simulated,
- Two controllers (planning and tactical) per measured sector, and one controller per feed sector.

Traffic characteristics

- Traffic samples derived from two traffic recordings,
- Traffic slightly increased and adjusted to create clusters of aircraft in the extended TMA organisation (resulting traffic close to high-density traffic),
- Traffic flows preserved, but considerably augmented in en-route organisation to create numerous and varied conflict situations (resulting traffic more complex than today),
- Full (CDTI) equipped traffic (thus allowing for the maximum opportunities for the use of delegation).

Controller working environment

- In June experiment, paper strips and no advanced tool (to support identification of controllers' need for specific information),
- In November experiment, simplified version of EATCHIP validation platform (strip-less environment) with no decision support tool, and with some marking for aircraft involved in delegation procedures (new interface not fully validated).

Cockpit display of traffic information

- No traffic information related to surrounding traffic,
- Display cues about selected traffic to support pilots in their specific new tasks.

B.3.1.2 Experimental conditions

The main characteristics of the EACAC'2000 experiments set-up and performance include the followings:

- Two sessions (of two weeks each) in June and November 2000 (one week to experiment sequencing applications in ETMA, and another week for crossing/passing applications in en-route),
- Each week of experiment consisting in briefing session, set of exercises (training, qualitative, measured) and final debriefing session
- Participant training to procedures, airspace (through presentations only) before simulation
- Each exercise was simulated twice: once without delegation (i.e., conventional control) and once with (in a second step).

Experiment participants

- Experienced air traffic controllers (2x6) from different European countries (most of them were not familiar with the simulated airspace),
- Pseudo-pilots (2x4), and (5) test and airline pilots in November experiment only.

Data collection during experiments

- (Individual) questionnaires (before, during and after experiments)
- In June, questionnaires (60 items) only at the end of each week of experiment
- In November, questionnaires at the end of briefing/training session (7 items), after each exercise with delegation (7 items) and at the end of each week of experiment (70 items)
- In November only, (subjective) workload assessment continuously during the experiment (ISA technique) and at the end of each exercise (NASA-TLX technique)
- Collective debriefings (during and at the end of each session)
- System recordings (during experiment)

Experiment limitations

The following describes recognised limitations and constraints of the EACAC'2000 experiment:

- Due to the short training, the exercises without delegation were performed before the exercises with, which could have introduced some bias favourable to the concept under assessment.
- All the traffic was equipped to receive delegations, thus offering maximum opportunities to use it.
- In the November experiment, due to technical problems (i.e., interface trouble shooting), and lack of controllers' training to the new simulated environment, the experiment participants did not manage to get familiar enough with the delegation procedures.

B.3.1.3 Experiment objectives (“low level” objectives of validation)

The validation objectives for the EACAC' 2000 experiments (either related to “external” validation or “internal” validation of the new ATM system) include the followings:

- To assess users (i.e., controllers and pilots) acceptance.
- To assess the impact on controller activity.
- To assess the impact on flight efficiency.
- To assess the impact on safety.

Subsidiary validation objectives (related to “internal” validation of the new ATM system) were:

- To assess the usability of the new controllers' working position interface.
- To improve the procedures and the phraseology.

In summary,

Scope of experiment (and limits)	Experimental conditions	Experiment objectives
Two distinct airspace organisation Due to the short training, the exercises without delegation were performed before the exercises with. All the traffic was equipped to receive delegations.	Two sessions (of two weeks each) in June and November 2000. Each exercise was simulated twice: once without delegation. Experienced air traffic controllers (2x6) from	To assess users' (i.e., controllers and pilots) acceptance. To assess the impact on controller's activity. To assess the impact on flight efficiency.

Scope of experiment (and limits)	Experimental conditions	Experiment objectives
In the November experiment, due to technical problems and lack of controllers' training to the new simulated environment, the participants did not manage to get familiar enough with the delegation procedures.	different European countries. Participant training to procedures, airspace (through presentations only) before simulation.	To assess the impact on safety.

Table 22: Outline of the EACAC experiment

Description of each experiment objective with:

- Hypotheses to be validated (or invalidated) through the experiment (either implicit or explicit),
- Related expected benefits (or constraints), or
- Related ATM system element under validation.

To assess users' (i.e., controllers and pilots) acceptance		
Hypotheses (if any)	Related "high-level" validation objective(s)	Related ATM system element(s) under validation
Effective use of delegation by controllers	Human involvement and commitment	Delegation procedures
Delegation procedures accepted by pilots (Implicit)	Human involvement and commitment	Delegation procedures and cockpit interface

To assess the impact on controller's activity		
Hypotheses (if any)	Related "high-level" validation objective(s)	Related ATM system element(s) under validation
Increase in controller's availability, mainly in terms of cognitive resources	Potential increase in airspace capacity (not directly measurable in this early stage of validation)	Controller's activity
Reduction of the manoeuvring instructions	<i>Not explicit.</i>	Controller's activity, and controller-pilot communications
Earlier sequencing of aircraft	<i>Not explicit.</i>	Controller's activity
Less time critical instructions (considered as time consuming since they require active monitoring)	<i>Not explicit.</i>	Controller's activity
Reduced workload monitoring	<i>Not explicit.</i>	Controller's activity
Smoothing over time of controllers' instructions	<i>Not explicit.</i>	Controller's activity

Note: The relationship between Increase in controller’s availability and potential increase in airspace capacity is evoked in the EACAC’2000 final report, but was not further addressed, nor taken into account in the exploitation of the experiments results.

To assess the impact on flight efficiency		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation
No adverse effect on (individual) flight efficiency	Economical viability of the operational concept	Aircraft operations

To assess the impact on safety		
Hypotheses (if any)	Related “high-level” validation objective(s)	Related ATM system element(s) under validation
Increased, or at least same, level of controllers’/pilots’ perceived safety	Increase in safety	Controllers’/pilots’ confidence in new air traffic operations
Increased, or at least preserved, safety margins in air traffic operations	Increase in safety	Aircraft separation, proximity and transfer conditions

Table 23: Description of each objective of the EACAC’2000 experiments

Note: The EACAC’2000 experiments were not focused on technical aspects of safety, but rather on the human contribution to safety, as well as the achievement of safe separations of aircraft.

B.3.2 Indicators, Metrics and measurements

The different metrics from the EACAC’2000 experiment are described hereafter, and are classified mainly according to their related experiment objective(s):

- Pre-simulation questionnaire
- Metrics related to user (ATCO) acceptance
- Metrics related to controller activity
- Metrics related to flight efficiency
- Metrics related to safety
- Metrics related to “internal” design (i.e., CWP interface)

All metrics derived from the EACAC’2000 experiment are **relative metrics** used to conclude on trends. As such, there were no precise decision criteria to decide on whether the validation objectives/hypotheses are met. Instead, judgement based comparison of results (e.g. with and without delegation) was performed to assess the impact of the operational concept on various items/aspects of the ATM system.

Note: In the following review, metrics having same characteristics (i.e., attributes, related objectives and means of calculation) were gathered into the same tables.

B.3.2.1 Pre-simulation questionnaire

Controllers' initial expectations and feeling about the operational concept were collected (through questionnaires) prior to the simulation (but, after the initial briefing session), and were later confronted to their answers (to the same questionnaires) at the end of each exercise.

In the following tables, the experiment objectives related to each questionnaire item were not explicitly stated, but was quite easy to determine during this review of the EACAC'2000 experiment final report.

Metric name: Controller ratings of their background knowledge about EACAC			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Binary	None	Used in judgement-based analysis of others metrics	Questionnaire pre-simulation

Metrics names: - Controller ratings of systematic risk ("always dangerous") associated with delegation - Controller ratings of decrease in safety (associated with delegation)			
Attributes	Related experiment objective(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating (strongly agree, agree, disagree, strongly disagree)	To assess the impact on safety	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire pre-simulation (items 1, 6) Cumulative sum per level of rating, per briefing session

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Metrics names: - Controller ratings of possible ("some") benefits of delegation - Controller ratings of benefit in capacity and efficiency (associated with delegation)			
Attributes	Related experiment objective(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating	To assess the controllers' acceptance	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire pre-simulation (items 2, 7) Cumulative sum per level of rating, per briefing session

Metrics names:			
<ul style="list-style-type: none"> - Controller ratings of requirement for permanent (“always”) air traffic control by ATCO - Controller ratings of possible reduction in ATCO’s workload - Controller ratings of possible increase in ATCO’s monitoring effort 			
Attributes	Related experiment objective(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating	To assess the impact on controller activity	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire pre-simulation (items 3, 4, 5) Cumulative sum per level of rating, per briefing session

B.3.2.2 Metrics related to user (ATCO) acceptance

Both subjective metrics to get feedback from ATCO about their acceptability of the operational concept of delegation, and objective metrics about the effective use of delegation procedures, were used during the EACAC’2000 experiments.

Subjective metrics were collected using final questionnaires at the end of each week of experiment. Not all the items (70) of these questionnaires are listed in this review, but only the major ones used in the discussion developed in the EACAC’2000 experiment final report.

Metrics names:			
<ul style="list-style-type: none"> - Controller ratings of usefulness of the “concept” (i.e., delegation procedures) - Controller ratings of understanding of the “concept” (i.e., delegation procedures) - Controller ratings of usability of the “concept” (i.e., delegation procedures) - Controller ratings of compatibility of delegation procedures with current working methods 			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating (totally/completely, generally/mostly, partially, not at all)	To assess the user acceptance	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per level of rating, per each week of experiment, and overall sum

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Metric name: Rate of use of delegation procedures per aircraft			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative (%)	To assess the user acceptance through effective use of delegation	Judgement based comparison of results between sectors and experiments	Post-experiment counting per sector and per exercise Ratio of number of “delegated” aircraft “/ total number of “concerned” aircraft

Note: “Delegated” aircraft refer to those aircraft that received at least one delegation instruction, whereas the “concerned” aircraft correspond to potentially delegable aircraft. In extended TMA, a concerned aircraft is an arrival flight.

Metrics names: - Time ratio of delegation procedures per sector (i.e., “concerned aircraft”) - Time ratio of delegation procedures per aircraft (i.e., “delegated aircraft”)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative (%)	To assess the user acceptance through effective use of delegation	Judgement based comparison of results between sectors and experiments	Post-experiment counting per sector and per exercise Ratio of total delegation time / total flight time of aircraft (either “concerned” or “delegated” ones)

Metric name: Number of use of each delegation procedures per sector			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative	To assess the user acceptance through effective use of delegation	Judgement based comparison of results between sectors and experiments	Post-experiment counting per sector and per experiment Cumulative sum of each instructions

B.3.2.3 Metrics related to controller’s activity

Both subjective metrics to get feedback from ATCO and objective metrics were used during the EACAC’2000 experiments to assess the impact of delegation on controller’s activity.

The impact on controller’s activity was considered at three levels: individual workload, strategies in handling aircraft and resulting activity (in terms of instructions used in space and time).

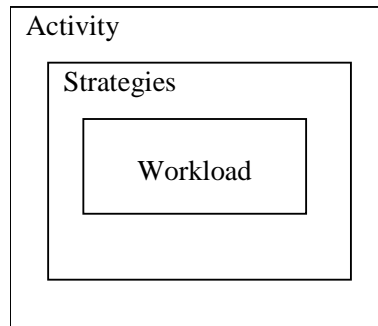


Figure 14: EACAC complementary levels of controller’s activity analysis

Subjective workload metrics were based on different sources of data collection and analysis (ad-hoc questionnaires collected after each week of experiment, continuous Instantaneous Self-Assessment (ISA) during the exercises, and NASA-Task Load Index (TLX) based on data collected after each exercise).

In the EACAC’2000 final report, it s mentioned that in November experiment different metrics related to controller workload gave different results: Objective physiological measurements confirm feedback obtained through final questionnaires, whereas ISA and NASA-TLX techniques both invalidated the results obtained through the final questionnaires.

Metric name: Controller ratings of workload/mental effort required to monitor delegations			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating (much lower/lower/higher/much higher)	To assess impact on controller activity through their workload	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per level of rating, per experiment

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Metric name: Controller ratings of factors contributing to workload (associated with delegation)			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Binary (for each factor)	To assess impact on controller activity through their workload	Relative metric to conclude on trends + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per factor, per experiment

Metric name: Controller ratings of perceived overall workload			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 5 levels of rating (very high /high/fair/low/very low)	To assess impact on controller activity through their workload	Judgement-based comparison of distribution with and without delegation <u>Note:</u> Including judgement-based correlation between controller's perceived workload and controller's familiarity of the simulated environment (CWP, sector, type of traffic)	Continuous (each 2 mn) ISA measurement during experiment Percentage of each level of rating per sector, and per controller position (executive and planning)

Metric name: Temporal distribution of perceived overall workload per controller position			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 5 levels of rating (very high /high/fair/low/very low)	To assess impact on controller activity through their workload	Judgement-based comparison of distribution with and without, and of both controller position	Continuous (each 2 mn) ISA measurement during experiment Cumulative sum per level of rating per controller position (executive and planning), and per exercise

Metric name: Controller ratings of perceived workload			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary	To assess impact on controller activity through their workload	Judgement-based comparison of controllers' scores (with and without delegation)	Questionnaire post-exercise NASA TLX score combining ratings of mental, physical and temporal demands, as well as performance, effort and frustration levels perceived

Metrics names: Controller's physiological parameters: pupil diameter, dwell time and heart rate			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller activity (or workload) through physiological parameters	Judgement-based analysis of measurements	Measurements during experiment

Metrics names: - Number of ATC instructions per sector - Number of each type of ATC instructions per sector			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller activity through their strategies in handling aircraft (i.e., air traffic control)	Judgement-based comparison of results with and without delegation	Post-experiment counting of HMI selections by pseudo pilots Cumulative sum of ATC instructions, per sector and per experiment

Metrics names: - Number of controller calls - Mean duration of controller calls - Number pilot calls - Mean duration of pilot calls			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller / pilot activity through their communications	Judgement-based comparison of results with and without delegation, and between controller's and pilot's sides	Post-experiment analysis of radio occupancy

Metric name: Sequences of aircraft at IAF			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller activity through their strategies in handling aircraft (i.e., air traffic control)	Judgement-based comparison of results with and without delegation	Post-experiment analysis of system recordings

Metrics names:			
<ul style="list-style-type: none"> - Controller ratings of their confidence in determining the (appropriate?) use of delegation instructions - Controller ratings of their hesitation when applying delegation instructions - Controller ratings of their ability to (appropriately?) use delegation instructions - Controller ratings of workload and stress in monitoring of delegations 			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of rating (much lower/lower/higher/much higher)	To assess impact on controller activity through their use of delegation instructions	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per level of rating, per experiment

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Metric name: Number of (type of) ATC instructions according to distance to IAF per sector			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller activity through their use of delegation instructions <u>Hypothesis:</u> earlier sequencing of aircraft (i.e., better anticipation in implementing controller's strategies)	Judgement-based comparison of results with and without delegation (*)	Post-experiment analysis of system recordings Cumulative number of (type of) instructions (according to distance to IAF) per sector and per exercise

(*) Different representations of the metric were used to make this comparison. These include cumulative curves and density lines of total number of instructions, and histograms distinguishing the type of instructions.

Metric name: Number of (type of) ATC instructions over time per sector

Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on controller activity through their use of delegation instructions <u>Hypothesis</u> : smoothing of controllers' tasks over time and less time-critical instructions (considered as time consuming since they require active monitoring) and thus, reduced monitoring workload	Judgement-based comparison of results with and without delegation (*)	Post-experiment analysis of system recordings Cumulative number of (type of) instructions per period of time, per sector and per exercise

(*) Different representations of the metric were used to make this comparison. These include cumulative curves and density lines of total number of instructions, and histograms distinguishing the type of instructions.

B.3.2.4 Metrics related to flight efficiency

Both subjective metrics to get feedback from ATCO, and objective metrics were used during the EACAC'2000 experiments, to assess impact on flight efficiency.

Metric name: Controller ratings of increased efficiency allowed by delegations (in terms of fuel consumption and time savings)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary	To assess impact on flight efficiency	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per level of rating, per experiment

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Note: Actually, it is reported in the EACAC'2000 final report that the controllers misunderstood the metric, and evaluated the increased efficiency in terms of their reduced workload to handle same amount of traffic.

Metrics names:			
- Total distance flown by aircraft (in the sectors of interest) - Total fuel consumption over the fleet (in the sectors of interest) - Total flight time over the fleet (in the sectors of interest)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on flight efficiency	Judgement-based comparison of results with and without delegation	Post-experiment analysis of system recordings Cumulative sum of each parameter per experiment

B.3.2.5 Metrics related to safety

Both subjective and objective metrics were used to first assess if safety is impacted by the use of the new ATC instructions, then to identify causes that lead to unsafe situations.

Metric name: Controller ratings of potential (“could”) increase in safety			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 levels of evaluation (totally, generally, partially, not at all)	To assess impact on safety	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-experiment Cumulative sum per level of rating, per experiment

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

Note: According to ATCO, factors that supports the increase in safety include:

- More accurate maintenance of separations,
- Increase in pilot situational awareness,
- More closely verification of the traffic situation enabled by the reduction of ATCO workload to handle the traffic.

According to ATCO, factors that may affect the safety of the delegation procedures include:

- Pilots’ ability to comply with the new ATC instructions,
- Pilot/controller misunderstanding due to phraseology,
- Reduction of the safety margins currently applied to cope with approximation of aircraft separation distance by ATCO.

Metrics names: - Number of “very serious” losses of separations - Number of “serious” losses of separations - Number of “minor” losses of separations - Mean duration of losses of separations			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on safety	Judgement-based comparison of results with and without delegation	Post-experiment analysis of system recordings Computation of each parameter per exercise

Metric name: Maximum Aircraft Proximity Index (API)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on safety	Judgement-based comparison of results with and without delegation	Post-experiment analysis of system recordings Weighted measure [8] of conflict intensity (where 100 is a mid-air collision and 1 is a minor violation of the separation standards)

Note: According to the judgement-based analysis of the results, factors that may lead to unsafe situations included:

- Controller’s overconfidence and excessive expectations from aircraft/pilot capabilities,
- Failure or lack of controller’s expertise in using delegation procedures,
- Pilots/controllers’ errors either in acting or making decisions during delegation procedures.

It is not clear whether these safety-influencing factors (“some observed, others suspected”) were hypotheses to be confirmed through the experiments, or rather conclusions from the experience gained through the experiments.

Metric name: Number of “catching up” situations between two aircraft in sequence at transfer between sectors			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Quantitative Not binary	To assess impact on safety	Judgement-based comparison of results with and without delegation	Post-experiment analysis of system recordings

Note: According to the EACAC’2000 final report, relevance of such a metric was not obvious, since the next sectors were not controlled in the simulations.

B.3.2.6 Metrics related to “internal” validation (i.e., CWP interface)

Subjective feedback from ATCO was collected along the EACAC’2000 (at the end of each exercise performed in November), in order to validate the CWP interface developed to support the operational concept.

Metrics names:			
<ul style="list-style-type: none"> - Controller ratings of encountered difficulty in the CWP interface - Controller ratings of easy understanding of marking information (added to CWP) - Controller ratings of easy finding of required information (through CWP) - Controller ratings of additional work complexity due to marking facilities - Controller ratings of more easy work due to delegation (**) - Controller ratings of essential support got through marking facilities - Controller ratings of additional interfering tasks (with respect to “main activity”) due to delegation (**) 			
Attributes	Related experiment objective(s)	Decision criteria	Means of calculation (or measurements)
Subjective Qualitative Not binary Scale of 4 values of agreement with assertion (strongly agree, agree, disagree, strongly disagree)	To assess usability of new CWP interface	Relative metric to conclude on trends (*) + Judgement-based analysis of (textual) comments	Questionnaire post-exercise Cumulative sum per level of rating, per briefing session

(*) Even number of rating positions, in order to force controllers to take position in favour or against evaluated items.

In this review of the EACAC’2000 experiment, such metrics are classified as metrics related to “internal” validation, although this is not obvious some of the questionnaire items (**).

ANNEX C ANALYSIS OF INTEGRA METRICS ON EACAC'2000 EXPERIMENTS

This annex contains the detailed analysis of the INTEGRA metrics applied to the EACAC'2000 experiments [3] based on the framework described in 3. The annex addresses:

- The study scope and objectives, and
- The INTEGRA metrics based on the EACAC metrics and measurements.

C.1 Scope of the study

When the INTEGRA metrics were applied on the EACAC'2000 experiments, not all the INTEGRA metrics were specified. Therefore, the scope of the study was focused on the assessment of the impact on:

- Capacity and.
- Safety.

Actually, the study was also an opportunity to assess the performance of the INTEGRA metrics and methodologies (as implemented at that time). It also allowed some fine-tuning of the metrics before their final delivery.

Data analysis after EACAC'2000 experiments

Dedicated software was used to read the data recorded during the EACAC'2000 experiments in order to compute the factors that are part of the INTEGRA metrics calculation.

C.2 INTEGRA metrics applied in EACAC

The different metrics from INTEGRA in EACAC'2000 experiments are described hereafter, and are classified mainly according to their related experiment objective(s):

- Metrics related to controller's activity (i.e., the capacity metrics),
- Metrics related to safety.

The Capacity metrics and the Resilience metrics are both supposed to be compared with a threshold representing the maximum controller processing capacity. However such a capacity was not determined in INTEGRA. So, all INTEGRA metrics applied in EACAC'2000 experiment remain **relative metrics** used to conclude on trends. So, judgement based comparison of results was performed to assess the impact of the operational concept on the ATM system.

C.2.1 Metrics related to capacity

Metric name: Instantaneous IPL (Information Processing Load)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Post-experiment analysis of system recordings Weighted (*) sum of processing contributions from 7 causes in a given time step: - Acquisition of a new flight, - Determining the forecast interactions, - Planning resolutions, - Implementing the planned resolutions, - Monitoring conformance to Plan, - Other changes to trajectory, - Co-ordination with other control agencies.

(*) Tuning of the weights was not performed in INTEGRA. All weights were set to 1.

The following tables detail the metrics that contribute to the aggregated metric above:

Metric name: IPL for Acquisition of a new flight			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum of aircraft arriving under control

Metric name: IPL for Determining the forecast interactions			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum of all tactical changes of trajectory

Metric name: IPL for Planning resolutions			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Weighted (estimated resolution difficulty) sum of tactical changes when another aircraft is within a defined interaction volume

Metric name: IPL for Implementing the planned resolutions			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum of tactical changes when another aircraft is within a defined conflict range.

Metric name: IPL for Monitoring conformance to Plan			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Average number of aircraft under control

Metric name: IPL for Other changes to trajectory			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum of events where an aircraft turns and climbs simultaneously

Metric name: IPL for Co-ordination with other control agencies			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum of tactical changes of trajectory

C.2.2 Metrics related to safety

Metric name: Propensity (likelihood of a safety significant event occurring during normal operations)			
Attributes	Related experiment objective(s) / hypothes(e)s	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on safety	Judgment based comparison of metric with and without delegation	Post-experiment analysis of system recordings Calculation involving the variances of the probability distribution of aircraft on the three axes, which are estimated by multiplying a value based on the FMS control capability by a set of multipliers taking into account: - The presence of advanced tools, - The aircraft density, - Higher than normal information processing load, and - Weather.

The following tables detail the metrics that contribute to the aggregated metric above:

Metrics name: Resilience (extent to which the ATM system responds to a safety significant event without causing more such events)			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on safety	Judgment based comparison of metric with and without delegation	Post-experiment analysis of system recordings Calculation involving the maximum controller IPL, the instantaneous information processing load and a contingency information processing load computed as the sum of: - The information processing load needed to correct a failure to make a state vector change for a potential hazard, - The information processing load due to guidance errors, and - A hazard analysis information processing load.

Metric name: IPL for Correcting a failure to make a state vector change for a potential hazard			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum for each reference aircraft of the sum, of their total proximity factor (= Sum for all aircraft in a defined interaction area, of a function of the distance between the reference aircraft and the proximate aircraft, with values in [0,1])

Metric name: IPL due to guidance errors			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Cumulative sum, for each reference aircraft in interference with another aircraft, of the probability density mapping multiplied by the total proximity factor around both aircraft in the pair D

Metric name: IPL for hazard analysis			
Attributes	Related experiment objective(s) / hypothese(s)	Decision criteria	Means of calculation (or measurements)
Objective Non-intrusive Quantitative Not binary Scalar function (of time)	To assess impact on controller activity through their workload	Judgment based comparison of metric with and without delegation	Arbitrary value

***** END OF DOCUMENT *****