EUROPEAN ORGANISATION FOR THE SAFETY OF AIR NAVIGATION

EUROCONTROL

Technical Annex: An Experimental Methodology for Selecting Fonts for Next Generation Air Traffic Management Systems

HRS/HSP-006-REP-02

Edition	**:**	**1.0**
Edition Date	**:**	**31.08.2000**
Status	**:**	**Released Issue**
Class	**:**	**EATMP**

EUROPEAN AIR TRAFFIC MANAGEMENT PROGRAMME

DOCUMENT IDENTIFICATION SHEET

DOCUMENT DESCRIPTION

Document Title

Technical Annex: An Experimental Methodology for Selecting Fonts for Next Generation Air Traffic Management Systems

WORK PACKAGE REFERENCE NUMBER: HRS/HSP-006

DELIVERABLE REFERENCE NUMBER:	**EDITION:**	1.0
HRS/HSP-006-REP-02	**EDITION DATE:**	31.08.2000

Abstract

This technical document reports work carried out as part of the CORE Requirements for Air Traffic Management (ATM) Working Positions Project conducted by the EUROCONTROL Experimental Centre (EEC) on behalf of the EUROCONTROL Human Factors and Manpower Unit* (DIS/HUM) within the EATMP** Human Resources Domain (HUM).

It provides an empirical methodology for selecting and evaluating screen fonts suitable for use with the technologies to be employed in the emerging generation of ATM working positions.

The requirements and rationale on which this methodology is based are described in a companion EATMP Report 'Font Requirements for Next Generation Air Traffic Management Systems' (EATMP, 2000).

* Formerly known as the 'ATM Human Resources Unit'
** European Air Traffic Management Programme

Keywords

Human Factors	Air Traffic Control (ATC)	Display Technology	Font Selection
Typography	Text Processing	Reading	Visual Search
Eye movement	Design	Part-task Experiments	Experimental Design
Human-Machine Interface	Graphical Interaction	Controller Working Position (CWP)	Evaluation Methodology

CONTACT PERSON: A. JACKSON	**TEL:** +33-1-6988.7544	**UNIT:** EEC / ATM R&D CoE

AUTHOR: Stefana BROADBENT for CARA, BROADBENT & JEGHER

DOCUMENT STATUS AND TYPE

STATUS		**CLASSIFICATION**	
Working Draft	☐	General Public	☐
Draft	☐	EATCHIP	☑
Proposed Issue	☐	Restricted	☐
Released Issue	☑		

ELECTRONIC BACKUP

INTERNAL REFERENCE NAME: G\Own_use\Delvrabl\Released\HFs\Fonts_tech.doc

HOST SYSTEM	**MEDIA**	**SOFTWARE**
Microsoft Windows	Type:	**MicroSoft Office 97 (MS97)**
	Media Identification:	

DOCUMENT APPROVAL

The following table identifies all management authorities who have successively approved the present issue of this document.

AUTHORITY	NAME AND SIGNATURE	DATE
Manager CORE Project ATM Research and Development Centre of Expertise (ATM R&D CoE)	A. JACKSON	28.08.2000
Manager Human Factors Sub-Programme (HSP) Human Factors and Manpower Unit (DIS/HUM)	V.S.M. WOLDRING	28.08.2000
Manager Human Resources Programme (HRS) Human Factors and Manpower Unit (DIS/HUM)	M. BARBARINO	30.08.2000
Chairman Human Resources Team (HRT)	A. SKONIEZKI	31.08.2000
Senior Director Principal EATMP Directorate (SDE)	W. PHILIPP	31.08.2000

DOCUMENT CHANGE RECORD

The following table records the complete history of the successive editions of the present document.

EDITION	DATE	REASON FOR CHANGE	SECTIONS PAGES AFFECTED
0.1	09.11.1999	Draft	All
0.2	10.03.2000	Proposed Issue	All
1.0	31.08.2000	Released Issue	All (document configuration)

TABLE OF CONTENTS

Page intentionally left blank

EXECUTIVE SUMMARY

This technical document reports work carried out under contract by CARA, BROADBENT AND JEGHER as part of the CORE Requirements for Air Traffic Management (ATM) Working Positions Project. This Human Resources Domain (HUM) project is conducted within Work Package 6 of the Human Factors Sub-Programme (HSP) – part of the larger EATMP Human Resources Programme (HRS) - by the EUROCONTROL Experimental Centre (EEC) on behalf of the EUROCONTROL Human Factors and Manpower Unit[1] (DIS/HUM). It provides an empirical methodology for selecting and evaluating screen fonts suitable for use with the technologies to be employed in the emerging generation of ATM working positions.

The requirements and rationale on which this methodology is based are described in a companion report 'Font Requirements for Next Generation Air Traffic Management Systems' (EATMP, 2000), to which this technical report is a supplement. EATMP (2000) was also prepared by CARA, BROADBENT AND JEGHER as part of the same project on CORE requirements for ATM Working Positions.

Chapter 1 establishes the context of the document, states the objectives and describes the structure.

The remainder of the document is divided into five main sections (A to E), each with subsidiary chapters.

SECTION A provides the framework of the methodology.

Chapter 2 describes the process of selecting appropriate candidate fonts for the context of evaluation.

Chapter 3 discusses the characteristics of text processing in Air Traffic Control (ATC) and their consequences for evaluation.

Chapter 4 describes the heart of the methodology, the 'controlled part-task experiment', and identifies the main components which must be managed.

The following four sections (B to E) apply this framework to the three principal types of text processing previously identified as dominant in ATC applications.

SECTION B analyses and proposes a detailed methodology for evaluation of fonts for use in radar labels.

SECTION C pursues a similar process for fonts to be used in lists and tabular formats.

SECTION D develops the methodology for menu applications.

[1] Formerly known as the 'ATM Human Resources Unit'

SECTION E describes the integrated evaluation of these different elements in a common experiment to consider trade-offs and interactions.

Some Recommendations on visual display ergonomics, References, a list of the Acronyms and Abbreviations used in this report and their full designation, as well as a list of the Contributors to this document, are annexed.

1. INTRODUCTION

The CORE Requirements for ATM Working Positions Project is managed at the EUROCONTROL Experimental Centre (EEC) as part of the Human Factors Sub-Programme (HSP) of the EATMP Human Resources Programme (HRS). The main objectives of this project are concerned with supporting the development of ATM in ECAC[2] area through providing methods and material to assist, the requirements capture, design, specification, development and evaluation of Controller Working Positions (CWPs) for European ATM.

In the domain of Human-Machine Interaction (HMI) the homogeneity of solutions and methods can greatly reduce the time of development and integration of new tools and can ensure that common human factors concepts are developed for all of Europe. In order to provide guidance on recurring problems, CORE wants to establish a reliable, stable and up-to-date set of reference studies on various aspects of the CWP.

In 1997, following an initial synthesis of HMI material for en-route working positions (Jackson & Pichancourt, 1995), a workshop was held at the EEC to identify issues and recurrent problems in the area.

One of the issues identified was that of **font selection** for the new generation of raster scan displays. In particular, the substitution of the current generation of stroke written displays which provide a high-definition cursive character set, opens up a number of questions regarding the best font candidates for the new generation of ATC equipment.

As a result a study contract was launched to:

- study the requirements for fonts for the next generation of ATM systems;
- establish a methodology for the identification and evaluation of suitable fonts against such requirements.

The contract was awarded to CARA, BROADBENT AND JEGHER Associates at the end of 1998 and resulted in the following two major deliverables:

- an EATMP Report, ‘Font Requirements for Next Generation Air Traffic Management Systems’ (EATMP, 2000), which reviews and explains the issues and requirements for fonts in ATM working positions;
- the current document, ‘Technical Annex: An Experimental Methodology for Selecting Fonts for Next Generation Air Traffic Management Systems’, which describes a detailed procedure for the selection and evaluation of fonts for ATM applications.

[2] European Civil Aviation Conference

1.1 Context of the Document

One of the commonest problems of interface design is **font selection and the evaluation of fonts** for the new generation of raster scan displays. Future ATM systems will display considerably more textual information than in the past; text will be used to designate at least the following:

- the function of buttons and controls;
- the functions or values in menus;
- the call sign, basic and extended label information on the radar screens;
- aircraft data on tabular lists (electronic strips, sector inbound list, communication lists, etc.).

This significant increase of the amount of text in ATC windows-based interfaces will mean that the screen fonts will have to meet high standards of legibility and readability. There is currently no standard set of fonts for ATC; local authorities are selecting their own set of fonts or are accepting those proposed for the screen by the manufacturers of the video displays.

1.2 Objectives of the Document

The objective of this report is to provide an experimental methodology for the evaluation and selection of font sets for ATM systems. The document will attempt to give a set of structured guidelines on how to test fonts in the context of ATC activities.

1.3 Structure of the Document

To address the various factors that affect the selection of screen fonts, we have divided the document in two main elements:

⇨ A framework for the methodology (Section A):

- a definition of the overall experimental paradigm;
- a rationale for selecting the features of fonts to measure;
- a rationale for selecting the relevant behavioural measures to compare and evaluate fonts.

⇨ A detailed presentation of the experimental procedure to select the appropriate set of fonts for each type of text (Sections B to E):

- choosing the fonts to test;
- selecting the criteria on which to compare fonts;
- deciding what measures to take;
- implementing these choices in an experimental design.

SECTION A: A FRAMEWORK FOR AN EXPERIMENTAL METHODOLOGY

When approaching the evaluation of any user interface feature through experimentation (be it a font, a new screen, a new coordination tool), there are four main questions that have to be addressed from the outset:

1. What mental processes is the interface feature affecting?

2. What are the criteria on which to test and evaluate the interface features?

3. What form of experimentation will yield the most effective and reliable results?

4. What measures can be taken to evaluate the effect of the interface features on users' behaviour?

The requirements analysis presented in the companion document (EATMP, 2000) provides a series of conclusions which lead towards an answer to these questions. We will briefly recall the main points and expand on the implications they carry for an experimental paradigm for font selection.

Page intentionally left blank

2. PRE-SELECTING THE FONT SET TO EVALUATE

The issue of selecting a set of acceptable fonts for ATC displays raises a fundamental methodological problem: that of assessing the role of each of the multiple typographical features and feature combinations that compose what we generally call a **font**. In fact, as we discussed in EATMP (2000), the generic term of font has come to include the typeface, the style and the layout of characters. Text displayed on the screen generally has a certain typeface (Times, Verdana, Helvetica), a size, a colour, an intercharacter and interline space, a background, etc.

A **typeface** usually defines:

- the stroke width of characters,
- the intercharacter spacing,
- the x height of the characters[3].

The **layout** of the text includes features such as:

- the size of the characters,
- the interline spacing,
- the use of colour and highlighting,
- the number of characters per line.

A font, therefore, is made of a cluster of features and in turn each of these classes of features contain attributes that can be further decomposed into parameters that can take one of multiple values. Hence, it is impossible to think of addressing the issue of font selection in terms of a compositional approach, where one could design the perfect font composed of all the best features. Even though the constraints imposed by the technical characteristics of ATC screens reduce somewhat the range of possibilities, typography offers such a large number of features that the exploration of all the of them and of their combination is unachievable.

The solution to this methodological problem is, first, to reduce the combinatory space to a limited number of relevant features, second, to group them into significant clusters and, finally, to proceed by comparing clusters' legibility as a function of the context of use and typology of the text items to be displayed. This means first of all accepting that a huge effort of clustering has already been carried out by typographers over decades, to create typefaces that group

[3] The x height of a typeface is the size of the body of the characters represented by the letter x (x is the only letter that reaches out to all four corners of space). The x height is the height of a lowercase x in any font. Variations of this height can make a font appear large or small in contrast to other fonts, independent of size. This is a complementary measure to the type size. When a typeface is defined as being a certain size, for instance 24 point, this is only the approximation of the actual number of points from the top of the characters' ascenders to the bottom of the descenders. Within that size the ascenders or descenders can be any size. The x height therefore provides a supplementary measure and constraint.

features such as shape and x heights and intercharacter spacing, in effective and harmonic ways.

The task for the HMI expert seeking to provide a satisfactory solution for fonts on radar screens, must be seen as one of selecting among existing typefaces rather than designing a new type. As it stands, given the enormous variety of typefaces available and the numerous and stringent conditions that ATC activity imposes on all the elements of the visual display, experimenting on existing fonts will give rise to a sufficiently challenging research plan. However, **fine-tuning of individual fonts** can be an objective after a first evaluation of the most relevant candidates has been carried out.

The process of selecting the fonts to evaluate will therefore consist of various steps:

1. Pre-selecting a set of font-types as good candidates for comparison.
2. Deciding on a set of relevant layout values (size, colour and spacing).
3. Combining the layout features with the font-types.

By combining a ‘pre-selection’ strategy which eliminates non-relevant features, with a ‘contextual’ strategy that analyses the relevant properties as a function of the kind of textual objects displayed (text for menu, for labels, etc.), the number of clusters of features can be reduced significantly. The first step of the selection process for fonts in ATM systems will therefore consist of finding the relevant clusters for each text type that will be experimented upon.

3. THE CHARACTERISTICS OF TEXT PROCESSING IN AIR TRAFFIC CONTROL AND THE IMPLICATIONS FOR THE EVALUATION CRITERIA OF FONTS

It is important to realise that text processing in ATC can only be partially described as 'reading', because the nature of the text being processed is different from the continuous text that is the object of what is typically described as reading. On ATM screens there are no sentences or paragraphs, no syntactical constructions. Most of the text in fact consists of single words, numerals or alphanumeric codes. When there are more units the combination of units is by juxtaposition of alphanumeric elements.

On the new generation of CWPs there are mainly five categories of text that will be displayed:

- the call sign, basic and extended label information on the radar screens,
- aircraft data in tabular lists or electronic strips,
- messages in message windows,
- the function of buttons and controls,
- the functions or values in menus.

Each one of these text items has a set of functional characteristics which means that they are processed by the controllers in a very specific way (the type of text processing specific to each text category has been discussed in detail in EATMP, 2000.):

- labels need to be visually searched among the other labels, found and recognised;

- lists are processed sequentially in a way closest to what is generally defined as reading (a series of words and codes which enrich the description of an aircraft, for instance, are interpreted as soon as they are encountered and the new information is progressively integrated to what is already known about the aircraft);

- menus are processed as a part of a perceptual-motor routine.

The fact that different text items on the ATC screen involve different forms of text processing means that the criteria with which fonts must be evaluated should be specific to each type of text. It is therefore essential to test the fonts for each of these contexts independently, with a specific experimental technique, in order to assess the adequacy of the font for the type of activity that the controllers carry out with that text item. The criteria which will be used to decide if font A is more legible than font B will vary. Different text types will require fonts that support different combinations of the following:

- effective visual search,
- effective recognition,
- analytic and synthetic processes of reading,
- effective comprehension.

Page intentionally left blank

4. THE EXPERIMENTAL PARADIGM: CONTROLLED EXPERIMENTS

On the basis of the conclusions reached in the requirements phase of the study, the experimental paradigm for font evaluation should consist of controlled testing. This paradigm carries a series of implications in terms of the type of data that will be collected and the nature of the tasks required of the controllers. In fact, in order to maintain control over the specific effects of font design on the overall interaction between controllers and their tools, it is necessary to reduce the number of variables involved in an experiment and limit the number of tasks that the subjects carry out during the evaluation.

4.1 Part-task Experiments

An essential characteristic of the controlled laboratory approach therefore is that it implies 'part-task' studies (on this topic see 'Harmonisation of Man-machine Interface experiments in the context of PHARE advanced tools' [Broadbent, 1993]). By part tasks we mean that the experimental subjects will be engaged in a set of tasks that constitute only a fraction of their usual work activity: the experiment will isolate sub-tasks or elements of the controllers' activities which subjects will be asked to perform. Part-task experiments are used for subsystem or feature tests and evaluation[4].

Part-task experiments reproduce a particular activity segment and particular subsystem functions in comprehensive detail. Observed decrements or improvements of human performance can then be related to particular features of the specific subsystem or feature being tested.

The difficulty of evaluating interface options in real-time simulations is tightly linked to the measurement problem. It is not generally sufficient to measure the outcome of a scenario; tracking the process is also necessary because interface systems are multidimensional and outcome measures cannot alone determine which of the multiple potential factors active in the evaluation of an interface contributed to a specific outcome result.

[4] An example of such an approach was a study by Makins (Makins & Broadbent, 1993) to assess an operational scenario for a Multi-Sector Planning (MSP) environment developed with the rapid prototyping facilities of the EEC. Among the tools available in the MSP prototype was a filtering function allowing the suppression from the radar display of aircraft beyond a certain range from the aircraft involved in a specific conflict situation. The filtering tool was the object of an experimental evaluation to determine whether filters assist controllers in their performance of a conflict resolution task. Two different filters were compared to the traditional unfiltered situation. The test consisted of a problem-solving task on a semi-static display. Subjects were presented with a frozen display on the screen showing a traffic sample in which a conflict situation had developed. At the same time a written solution to the conflict was displayed; the solution was either 'correct' in the sense that it did not engender another conflict in the following five minutes, or 'incorrect' in the sense that it generated a conflict in the following five minutes. Subjects were asked to observe the traffic display and judge whether the proposed solution was correct or incorrect. The most important result of the study was that the use of a filter reduced the decision time by 25% without affecting the accuracy of the response. There were no statistically significant differences between the two filters.

Attempting to evaluate the role of a font type on the overall performance of an operator during a simulation is in fact practically impossible. Fonts are embedded in a complex interactive interface making it impossible to determine just what behaviours they are affecting. The effect of an interface option such as font type must therefore be evaluated with respect to specific text processing activities.

This means that tests must be carried out on partial simulators covering for instance only one control position. Most of the experimentation on fonts could, for instance, be carried out simply with one position on the radar screen only. It is obviously essential that the screen used to display the text be the same type and quality than those of the raster screen which will be used by the controllers eventually. The quality of the display plays in fact a major role on the perceptual legibility of the displayed font.

However, the part-task approach is not a substitute for real-time simulations; on the contrary, it implies that a second stage of experimentation be carried out to examine the interaction of the different features of the interface that have been chosen. An adequate experimental strategy for the evaluation of all interface items should in fact combine rigorously controlled experiments to examine the effect of specific interface features with more complex experimental contexts to establish the effect of the interrelation of factors.

4.2 Collecting the Data[5]

Running controlled experiments also has implications on the type of data that will be collected. The focus will be on **objective** rather than subjective **data** although this will not exclude the possibility of collecting subjective assessments as well. Objective measurements of performance such as reaction time and accuracy should constitute the main data to compare and evaluate fonts.

Measuring the duration of operations or events in general is a classical way of measuring controller performance. The time taken per task or per sub-task is some reflection of the amount of effort the controller is making. It is generally expected that given a task a subject that takes longer is experiencing some difficulty, and that it is thus a poorer interface. In general, the behavioural data most frequently collected in HMI evaluation studies are related to interaction with the system. Measures on these variables are usually taken either as event frequencies, time duration or event sequences.

[5] For convenience please note that in this section and those following the masculine gender (he, him, his) has systematically been used to refer to professional categories (as opposed to individuals) such as controllers, pilots, students, learners and instructors, which also include females and should be considered as such.

Data collection facilities can be included in the internal architecture of the system to support the process of data collection. Thus frequency/time data of the use of specific elements of the display (clicks on buttons or text items) can be recorded[6].

The computer system can be logged to record and **time-stamp** every click on an item of the display:

- click on waypoints in a list,
- click on flight levels on a label,
- click on aircraft data on a menu,
- etc.

At the end of a session a file is generated, with a session identifier, a position identifier and a list of all the interaction events that happened during the test, and it is easy to analyse the data on frequency and duration of each event.

Another type of data that can be usefully recorded in the context of controlled experiments is eye movement. Tracking eye movement provides interesting data on fixation points, fixation times and paths. It is a very rich data, often difficult to analyse, but when used as a companion to other data (such as logs or latencies) it provides very useful insights on controllers' strategies.

In the context of font selection we will see in further chapters, that eye tracking can solve some specific measurement issues that appear when we attempt to find a starting point for some text processing behaviours. In fact, when measuring latency of events that are very short (a few milliseconds) it can be difficult to measure the on-start of a process such as listening to an auditory signal and then starting to look on the screen for an item. An elegant solution can be to measure the moment in which a subject moves his eyes away from a stimulus or an element on the screen to look for the target item.

4.2.1 Debriefing Interview and Questionnaire

Subjective data can complement the objective data. A recording can be taken of comments made during the experimentation session. This data can complement the system recordings and can also provide an estimate of controllers' subjective impressions of the fonts for each context being tested.

[6] When counting events it is quite problematic to distinguish noise events from intentional events. Although the counts of events can be easily interpretable, one should bear in mind that some events may be noise and that events may receive a different interpretation for different subjects. An extension of event counting is event density, i.e. the number of given events per unit time (e.g. how many times a window has been opened in ten minutes). Event density is a more uniform metric, which controls for the fact that some users will take longer in a task than others.

It is also useful to include a debriefing session at the end of the experimental session. The debriefing session should be carried out as a structured interview. The experimenter follows a sequence of fixed points or questions that are put to the controller, verbally. A rather informal approach can be taken to allow the conversation to flow. The questions are subdivided in a set of topics covering various aspects of the task and interface:

- task complexity,
- comments on specific examples,
- preference of a certain font and format,
- specific difficulties with some items.

4.3 The Traffic

As a general remark we wish to point out that, in order to ensure the success of part-task experiments, it is advisable that **synthetic sectors** be created with eventually standard traffic samples. In current simulation studies the traffic samples on which controllers operate are modified from study to study and from centre to centre. In general, the experimental traffic samples chosen for a study are compatible with the traffic with which each centre is used to dealing.

However, using existing sectors and realistic traffic can introduce a very strong bias on the results of this type of study. Different levels of familiarity with the sector can enormously affect the speed and accuracy with which items are found and understood. The basis for any comparison of data across experiments and across establishments is that the traffic samples be equivalent. Standardising the traffic samples for experiments guarantees that the experimental tasks are common. It is in fact very difficult to compare the performance of two controllers if the events they have dealt with are radically different.

An effective solution is to create an arbitrary sector with a set of well structured rules governing the routes and traffic and let the controllers (subjects of the experiment) discover and carry out all the experimental tasks within the new airspace. Although the sequence of events that each controller will face will in fact be different given that the effects of each control decision will modify the situation and as the simulation progresses each controller will de facto be controlling a different traffic. Also, none of the controllers would be familiar with the sector. Therefore, all subjects would start with an equivalent level of knowledge. Subjects could be trained on this arbitrary airspace which would provide a common basis for comparisons of performance. The traffic samples may then vary but the standard sectors would introduce a common environment that may lead to a better control of experimental results.

4.4 Conclusions

In conclusion, the experimental approach that we are proposing for the evaluation and selection of fonts is based on the following principles:

1. Fonts will be evaluated within part task controlled experiments.

2. The experiments will be run with controllers performing a subset of their usual control activities on simulators including realistic display equipment (in order to ensure the same quality display) but synthetic sectors.

3. A set of different experiments will be run for each of the main types of text objects present on the screen (labels, lists, menus).

4. The experiments can be seen as incremental in terms of task complexity and ecological realism.

5. A pre-selection of fonts will be carried out, for each text type, on the basis of existing experimental results and contextual relevance principles.

Page intentionally left blank

5. THE EXPERIMENTAL DESIGN

There are different ways of designing experiments for the purpose of comparing or evaluating interface solutions:

- randomised subjects designs,
- correlated groups designs,
- mixed designs.

In the next few paragraphs we will provide a brief introduction of these different types of design protocols. A detailed presentation of experimental design procedures is found in Broadbent (1993).

5.1 Randomised Subjects Designs

The simplest way of conducting a comparative experiment is to use two **independent groups** of subjects, one group being tested under one condition and another being tested under the other condition (if relevant a third group, the control group, can be introduced which does not receive the treatment conditions). For instance, Group 1 would perform a set of tasks with interface A and Group 2 will perform the same set of tasks with interface B. The basic principle to observe in such a design is that there should be no bias in the groups. For the results to have any value the two groups of subjects must not differ in any significant way. For example, results could be biased if subjects of Group 1 had more experience than Group 2 subjects in a task that required a certain type of expertise.

The principle behind the randomised subjects designs is that independent groups of subjects are administered varying amounts or degrees of the variable that is being studied (e.g. the font type). For the experiment not to be biased it is necessary that the groups are formed by allocating subjects in random way. This type of design ensures that the carry over effects from one experimental condition to another are minimised. This type of design is employed when there are numerous conditions and it would not be appropriate to have subjects repeat the conditions, e.g. for fear that there would be a learning effect. Let us imagine for instance that a study was to be carried out comparing a series of input devices (keyboard, touch-screen, mouse and writing pad) for modifying entries in an aircraft sequence proposed by an approach aid. A series of operations involving the displacement and insertion of an aircraft in the proposed sequence is designed to test the efficacy of each input device. In order for the comparison to be effective it is essential that the task and the sequence of flights be identical for all four input devices. It is however considered important that subjects be unfamiliar with the approach sequence. Having the same subject repeat the task for each input device could lead to a learning effect that could seriously bias the results. Randomising the order of presentation of the devices would control for the familiarity effect.

Most studies on fonts will include more than one independent variable and often the **interaction** between the independent variables is an important research question. This is why in HMI experiments it is not unusual to encounter a **factorial design.** The factorial design is an experimental design in which two or more independent variables are simultaneously studied to determine their independent and interactive effects on the dependent variable.

Assuming that we have two independent variables A and B (type of font and level of foreground/background contrast) and that each variable has two levels of variation (one typeface Verdana (A1) and a second typeface Georgia (A2)) and there is a heavy contrast (B1) and a low contrast (B2)), there are the following four possible treatment combinations:

- Verdana and low contrast (A1B1),
- Georgia and low contrast (A2B1),
- Verdana and heavy contrast (A1B2),
- Georgia heavy contrast (A2B2).

Each of these treatment combinations is referred to as a cell. there are therefore four cells in this design. Subjects will be randomly assigned to the four cells.

	A1 Verdana	A2 Georgia
B1 high contrast		
B2 low contrast		

Subjects who were randomly assigned to the A1B1 would receive the A1 level of the first independent variable and the B1 level of the second independent variable. Similarly, the subjects assigned to the other cells would receive the designated combination of the two independent variables.

In a design such as this one with four cells, there are two effects to be analysed. The influence of the two independent variables has to be analysed statistically to determine if the different levels of each produced significantly different results. Whether in other words performance is better with Verdana than with Georgia. Equally important is to detect the interactive effects that may result from the simultaneous presentation of the two independent variables. In other words, detecting whether the performance achieved with a certain font depends on the intensity of the contrast. Analysing the independent effects of each independent variable requires that the effects of the different levels of variation of A and B be analysed separately. This means comparing the two A mean scores and determine whether they differ significantly and then compare the two B scores and determine whether they differ significantly.

Analysing the interactive effects of the two independent variables requires one to determine if the difference between the means of the levels of variation of one independent variable vary as a function of the levels of variation of the other independent variable. This means establishing whether the difference between A1 and A2 depends whether you are at level B1 or B2. In other words, by comparing the means for the Verdana and Georgia (i.e. comparing A1 to A2 separately from B) it may appear that Verdana allows users to perform more efficiently. Analysing the results in terms of the interaction between the two independent variables may show that Verdana is better with light contrast but worst than Georgia with strong contrast. This means that the effect identified comparing just the two fonts is in a sense rectified by analysing it in interaction with the other independent variable.

A factorial design can be constructed with three or more independent variables; we have used the example of two variables for simplicity. Statistically, there is no limit to the number of independent variables that can be included in an experiment. Practically, increasing the number of independent variables means increasing the number of cells especially if each variable has different levels. This means that the number of subjects must increase as well and this can be a problem in HMI experiments. In fact, if a 2x2 design in which ten subjects were allocated to each cell requires only forty subjects, a 2x2x2 arrangement yields eight cells and eighty subjects is necessary if the ten subjects per cell allocation is maintained. Another problem with factorial design is that with multiple variables there can be significant interactions, which are difficult to interpret. A three variable interaction means that the effect on variable A is a joint function of variables B and C. The experimenter must understand exactly what combination of B and C produced that effect.

Regardless of the aforementioned drawbacks, the advantage of manipulating more than one independent variable in an experiment and testing more than one hypotheses makes factorial design very useful in HMI experiments. The importance of studying interactive effects of independent variables upon the dependent variable was made clear by the previous example. Testing main effects does not need factorial design but testing interactions does and in HMI where many factors seem to interact in determining human performance this type of testing is essential. Moreover, factorial design allows the researcher to control for extraneous variables by building in the design possible confounding variables. In order to increase the reliability of the individual measures, and avoid noise effects, the number of measures or observations for each condition should be multiplied when possible.

A solution to the problem of reducing the very large sample of subjects needed for factorial design as soon as there are more than a few independent variables and conditions, is the method of **Latin Square design**. The principle behind a Latin Square design is that all treatment combinations are not represented. However, each level of each variable appears equally often. Let us take an example in which we have three independent variables: input device (I), colour of text (C), font (D). Each factor has three levels (1, 2, 3); for instance, (I) can be a mouse I1, touch-screen I2 or keyboard I3. If we were to

construct a factorial design there would be 27 combinations or cells as each level of I would be crossed with each level of C and D. In the Latin Square design for I1 will be investigated all dialogue types and all colours of display but not in combination. A choice is made of which combination of dialogue type and colour is most important to investigate with input device 1 (e.g. the mouse). All other 3x3-1 combinations are excluded. The following table shows one way in which the possibilities may be permuted:

	C1	**C2**	**C3**
D1	I1	I2	I3
D2	I2	I3	I1
D3	I3	I1	I2

Therefore, for input device 1 we only investigate the combination of colour 1 and font 1, for input device 2 we investigate only the combination of colour 2 and font 2. Only one third or nine of the possible 27 combinations are investigated. This type of design allows cutting down the number of subjects but at some expense with regard to the generalisability of the experiment. It is important to consider that if taking a Latin Square design from a book, conditions should be randomised and not just repeated in the order presented in the text.

5.2 Correlated Groups Designs

When subjects are admitted to all the conditions in the study, we have what is called a **repeated measures or within subjects** type of design. In repeated measures design each subject is measured in each of the conditions. For instance, if there are three interface solutions being compared, a subject will perform the same task with all three interfaces. In this type of design a correlation is introduced into the dependent variable measure; there is in fact a correlation between the various sets of subjects' scores.

This design allows reduction of the number of subjects and also allows us to control for the level of performance and 'performance style' of each subject. A subject that is not as proficient as the others (for instance, because the traffic sample he is working on does not correspond at all to his habitual sectors) will perform poorly in all three conditions. There are statistical procedures that will allow us to discount the individual attainment level of each subject when calculating the overall effect of the experimental conditions (for more details see Kirk (1982). The difficulty with repeated measure designs is that order effects may influence the results. There may for instance be a learning curve that may condition performance levels. One way of controlling this, is to give each subject a random assignment of experimental condition orders. Another solution is counterbalancing for order so that at the end of the

study the same number of subjects have done condition 1 as their first, second and third condition.

The within subjects type of design is very powerful when we can control the sequencing effect (learning or familiarity) or when the matching of subjects is done effectively, because it is very sensitive and more capable of detecting the existence of a treatment effect. In fact, the experimenter can eliminate the systematic effect produced by the correlated variable and be left with a less contaminated dependent variable measure, which increases the probability of detecting any effect of the independent variable. If we think of our example with four input devices for which four different groups of subjects were constituted, there exists a risk that, although we had randomly allocated subjects to groups, some extraneous variable linked to the groups may come to contaminate the results. If we could have compared the four devices by letting each subject carry out the tasks with each of the devices, we could have augmented our confidence in the results. As it is, eliminating the familiarity effect due to the repetition of the task was considered more important than the risks associated with any uncontrolled subject variable. The choice of design is often a matter of tradeoffs between advantages and disadvantages of each design within the constraints of a particular research question.

5.3 Combining Randomised and Correlated Groups Designs

One of the major constraints in human factors research in ATC is the limited availability of controllers for experiments. This means that often fewer than ten subjects are available for a given study. We have talked in a previous section of the consequence this has on the sampling of subjects and on the possibility of controlling for subjects' individual differences. With respect to experimental design the constraint on the number of available subjects has as a main consequence that of reducing the complexity of the design. When subjects are few it is more convenient to reduce the number of independent variables and avoid multiple interactions. Furthermore, it is more cost effective to augment the number of measurements or scores for each subject. This means employing repeated measures designs in which each subject carries out each experimental condition. This is feasible if the order or the repetition of the tasks has no real consequence on the results (no learning or habituation effect). Counterbalancing and randomisation of the order of presentation are techniques that can be very effective to reduce the order effect; however, there are cases in which repeating the same task or condition several times leads inevitably to a confounding of the results. In this case increasing the number of subjects and designing an independent subjects experiment is the only solution.

A design that is widely used in psychology and in HMI experiments is a **mixed design**: a factorial correlated groups design. It is a design in which some experimental factors are measured repeatedly, and form a part experiment, which is of, repeated measures design, within a larger experiment in which other factors are given to different groups of subjects. This is an ideal design when there are multiple independent variables, some of which would fit into a

within group design and others into an independent groups design. In such a case one factor can be represented by repeated measures (e.g. two different font types are presented to all subjects) and the other factor by independent subjects (three different levels of foreground background contrast presented to three independent groups of subjects). Measures are therefore repeated over tasks, independent over displays. See Siegel (1988) for more details.

In each chapter regarding the methodology for experimenting with a certain text type we will suggest the most appropriate type of experimental design, in order to compensate for the problem of experimenting with a very limited number of controllers.

SECTION B: AN EXPERIMENTAL METHODOLOGY FOR LABELS

Flight labels are dynamic text items present on the radar screen that include a call sign and, according to their state, a certain number of other items. Call signs are displayed as alphanumeric codes on the radar screen. They are composed of up to nine alphanumeric and have a well-defined morphology as they are composed systematically of the company code and flight number. The flight label can be in various states, e.g. standard, selected or extended. For each state the amount of information regarding the flight plan displayed is different (ranging from two lines to five if all the details are provided such as the requested flight level, exit flight level, etc.) However, this can vary in different ATM systems. These items can be numerals, alphanumeric codes, or character symbols (such as arrows).

Page intentionally left blank

6. SELECTING THE INDEPENDENT VARIABLES: THE FONTS

6.1 A Pre-selection for Typeface

As we will see in detail in the next section labels on the radar screen are the object of a visual search and have to be quickly identified and discriminated against a background of similar items. Legible fonts in this situation are those that support rapid global discrimination and identification. Furthermore, the information on the label and extended label after having been found, must be read and understood. This implies that the font also supports the process of text comprehension.

Firstly, candidate fonts for this type of text should have what we have called 'iconic legibility'. Iconic legibility is the global property of text that make it stand out against a dense background. Iconic legibility ensures immediate recognition of the text item as a whole enabling discrimination between whole items.

Secondly, candidate fonts for this context should have 'internal legibility'. Internal legibility is a function of character brightness, contrast between letter and background, font size, interword spacing, line spacing, line length. Internal legibility should enable the discrimination between characters (in particular, the problematic X and K, T and Y, I and L, N, M, W, I and l, O and 0 and Q, S and 5, U and V) and the identification of single characters.

Thirdly, if they are to be selectable objects they must have a minimum size, which is greater than the minimum size needed for read-only items.

Given what we know from existing literature on fonts and given the context of the ATC displays, there are whole families of font types that can be excluded as candidates because they are not legible on a computer screen. These are all Decorative, Oldstyle, and Serif font types.

- Font types should be chosen preferably from the Sans Serif style family. Experimental literature suggests that Sans Serif is more legible than Serif especially on the screen. However, in the case of numerals and upper case characters, Serif seems to provide a better legibility and intercharacter discrimination.

- Modern style families are to be preferred over other styles (which are more Decorative and have been created for paper supports). Modern styles tend to include the font types created for the screen.

For the same reasons fonts in condensed, or oblique must be excluded. Fonts in bold must be evaluated carefully as there are mixed results on the positive or negative effects of bold on legibility of text on the screen[7].

[7] Colour contrast is affected by bold. Bold text is heavier than normal intensity text.

As typefaces are very much determined by x height there is a clear rule that has emerged relative to the most legible x height: exclude fonts that have extreme x height (as a high x height or a low one, reduce distinctiveness).

Intercharacter spacing is another important component by which there are some types that can be immediately excluded from any experimentation because they reduce the legibility of words:

- Exclude fonts that have reduced width between characters.
- Monospaced fonts can be employed for input fields, but proportional fonts must be used for all the other text types. Although labels can be modified and thus are technically input fields as well it is preferable for recognition to select proportional fonts.

Here again, the results reported in the literature on proportional fonts apply mainly to continuous text. Even if, in the case of one word item such as labels, the problem of creating alignments that reduce the possibility of quickly isolating items does not exist, we still recommend to consider the issue when selecting candidates.

In conclusion, a reasonable set of typefaces to test first should be those designed for the screen such as Verdana and Georgia. Both should have a very good level of internal legibility. They have not however been designed to support global iconic legibility. Therefore, it is essential to examine experimentally whether they can support this function as well, providing controllers with text items that can be easily recognised.

It may also be appropriate to compare results with classic fonts such as Helvetica to create a benchmark reference value from which to compare the results of all other options.

6.2 Selecting the Relevant Layout Parameters

There is a number of layout features that can interact with the typeface and that must be experimented upon:

- the size of type[8] (we have to establish the best size for a type, a typeface may be very readable at point 14 but not so at point 11);
- the contrast level given by the colour of the font relative to the colour of the background.

[8] There is a considerable amount of evidence showing that a font size below point 12 is perceptually difficult to discriminate on a screen. However, in ATC the issue of screen real estate is very important and controllers request the smallest possible font size in order to avoid overlapping labels. It is therefore useless to experiment on sizes inferior to point 8 for moving items. As screen real estate is precious a size larger than point 16 would probably be unacceptable. It is necessary, therefore, experimenting within a range of point 12 to point 16.

6.2.1 Size

To be precise font size should be treated in terms of visual arc. By visual arc we mean the angle subtended by the area being viewed at the apparent focal point of the eye. The measurement of displayed information cannot in fact take the physical size of the character only into account, but also the distance of the character from the eye. Character size is measured by the height of the character in terms of its visual angle. The visual angle is used as a unit of measurement and is specified in terms of minutes of arc or degrees (one degree is equal to sixty minutes of arc). Character legibility approaches 100% as the visual angle exceeds ten minutes of arc.

Typically, controllers sit at a viewing distance of sixty to eighty centimetres. An EEC Report by Jackson & Pichancourt (1995) suggests that such a viewing distance, corresponding to a visual angle of 21'', requires characters of four to six millimetres. This corresponds to character thirteen pixels high. However, the size must be bigger for elements that must be selected. Jackson and Pichancourt suggest the following:

> *The ability to point at an object is governed by Fitt's Law which relates the Movement Time (MT) taken to acquire a target to the Width of the target along the direction of approach (W) and the Distance (D) which the hand must be moved.*
>
> $MT = a + b \log_2 (2D/W)$ *where a and b are empirically derived constants.*
>
> *For the mouse this can be roughly expressed as:*
>
> $MT = 100 \log_2 (D/W + 0.5)$
>
> *A worst case value for D on the Intergraph or Sony displays is 68.58 cm (27" diagonal) although a more typical value would be around 30 cm for the RPVD and tools. Under these conditions a minimum target width of 4mm along the direction of the movement would result in reasonable MT values of 620 and 750 milliseconds. Note that this size also corresponds to the minimum character height recommended.*
>
> Jackson & Pichancourt (1995), section 3.3.2.2

However, manipulation of font size is typically done in terms of points and it will be easier for the experimenter to manipulate the variable in terms of point sizes. It would be a good practice especially in order to compare fonts sizes (point 12 in Times does not correspond to point 12 in Verdana) to always calculate the font size also in degrees.

6.2.2 Contrast

Contrast conditions are very important to control because in operational situations labels have different states that are signalled essentially by a change of configuration of the background/foreground colour combination. The legibility of alphanumeric characters is affected by the colour and the luminance of the characters and the colour and luminance of the background. Coloured text should differ from coloured backgrounds by a minimum of hundred colour distance units as measured by an international lighting and colour standard, CIE method.

The contrast of foreground text and symbols against the background luminance is referred to as contrast ratio, which is measured by dividing the luminance of the foreground by the luminance of the background.

We suggest choosing the two most extreme cases of contrast on the HMI environment being used and testing the fonts systematically in both conditions.

Extreme cases can be the **highest and lowest** contrasted foreground/ background combinations. As an example, a high contrast may be offered by a label in unselected mode, assuming a general background (e.g. white characters constituted of 100% red, 100% green, 100% blue on grey background constituted of 29% red, 32% green, 29% blue). A low contrast is offered by a Transferred label on a general background (grey characters constituted of 48% red, 45% green, 45% blue on a grey background made up of 29% red, 32% green, 29% blue).

6.3 Conclusions

In conclusion the independent variables that must be included in the experiments are the following:

- **typeface** (two or three different typefaces),
- **character size** (two or three sizes),
- **contrast levels between font and background**.

7. SELECTING THE DEPENDENT VARIABLES: WHAT TO MEASURE

Deciding what is the best candidate means measuring the effect of the font on controller' activities. This in turn means deciding what will be the parameters measured to assess such an effect (the **dependent variables***).* An analysis of the processing involved in controllers' treatment of flight labels will allow us to identify the behaviour that is most relevant to measure.

7.1 Controllers' Text Processing of Labels

Call signs and labels are the object of constant monitoring. They are read and reread many times in the course of this monitoring activity.

7.1.1 Searching for a label never seen before

In the process of assuming a flight, there are instances in which the controller is searching for a call sign he has not seen before following a call from a pilot entering his sector. In this case the controller has an auditory trace of the call sign and some knowledge of the position which the aircraft is coming from and thus can restrict his search on the screen.

7.1.2 Searching for a label already seen

After a controller has assumed a flight he may not need to consider it for some time. At a certain point he may either want to verify the position of the aircraft, or carry out some operation on the menus. This implies scanning the display to identify the new position of the label.

In this instance the controller knows what he is looking for, has a mental representation of the label and is looking at the display to identify a configuration that matches his mental representation. In this type of search the controller is probably looking for a global configuration resembling his mental representation and is not looking at the single letters.

7.1.3 Recognizing a label already seen

In many cases the controller will be monitoring the screen and regularly check or encounter a label he has already seen and to which he has already attended. While not actively searching for this item, the controller recognizes the label when he sees it which means that recognizing the configuration triggers a set of associated information about that code (its route, destination, aircraft type, etc.).

7.2 Operational Measures

As we have seen above, a significant aspect of the processing of labels is searching and recognising text items; therefore, the main criteria which must be satisfied by font in this context is to effectively support the identification and discrimination of the label. This means that the measures taken to compare two font types should be relating to the accuracy and time, in order to be able to identify and recognise a label among other similar items.

Fonts will be compared on accuracy and the time it takes controllers to spot the label corresponding to the aircraft to which he wants to attend. As mentioned in a previous section eye tracking can solve the problem of finding a starting point for the visual search process. In fact, when measuring latency of events that are very short (a few milliseconds) it can be difficult to measure the on-start of a process such as listening to an auditory signal and then starting to look on the screen for an item. An elegant solution can be to measure the moment in which a subject moves his eyes away from a stimulus or an element on the screen to look for the target item.

The dependent variables or measures taken will be the following:

- **Time to identify a target** (latency measured between the moment the call sign is heard or visualised, and the moment the controller clicks on the label or from the first eye movement away from the last fixation point).
- **Accuracy** (measured as the number of correct identifications in a series of repeated searches)[9].
- **Eye movement** (fixation points, fixation times and paths).

[9] As mentioned in a previous section eye tracking can solve the problem of finding a starting point for the visual search process. In fact, when measuring latency of events that are very short (a few milliseconds) it can be difficult to measure the on-start of a process such as listening to an auditory signal and then starting to look on the screen for an item. An elegant solution can be to measure the moment in which a subject moves his eyes away from a stimulus or an element on the screen to look for the target item.

8. CONTROLLING VARIABLES

There is a number of variables to be controlled during this experiment:

- the traffic,
- the type of distractors (similar or different),
- the targets,
- the presentation input (auditory or visual),
- the other text items.

8.1 The Traffic

In order to avoid a bias in the results produced by different levels of familiarity with the target labels, distractors and the airspace as a whole, it is necessary to define a common set of arbitrary beacons and keep the airspace structure constant across experiments. In other words, in order to control for controllers' previous experience it is advisable to create a synthetic airspace. Subjects could be trained and validated on this arbitrary airspace which would provide a common basis for comparisons of performance. The traffic samples may then vary but the standard sectors would introduce a common environment that may lead to a better comparison of experimental results.

8.2 The Distractors: The Other Labels

The target label will appear in the midst of other labels. The role of the other labels in the search and recognition process is very important. In fact, we can imagine that the greater the number of labels and the more they are similar, the more difficult is the recognition task. As subjects will be repeating the search task a large number of times, the target and distractors will change at each trial.

To control for the effect of the distractors we suggest the following strategy:

- keep a constant number of labels (twenty distractor labels simulate a heavy traffic load and offer an extreme situation which should satisfy the most stringent conditions);

- there should be an equal proportion of very similar and dissimilar distractors;

- similar distractors are for instance call signs that have at least five out of nine characters identical to the target;

- dissimilar distractors have less than three characters in common with the target.

8.3 The Target

To avoid habituation and learning effects, the target label should be different and should appear in different positions at each trial. Although in real-life situation aircraft that have not yet been seen appear at the boundaries of the sector, we must, in this experimental situation, also control for those cases in which a label has to be found within a sector, as happens when an aircraft is attended to after having been assumed.

- the target position will change at every trial,
- the positions will include both sector boundaries and within sector,
- each trial will present a different target,
- the target will be dynamic as are the distractors.

8.4 The Presentation Conditions

In the real world controllers can be spurred to search for a label on the screen either by an auditory input, a visual input or by their own representation of the aircraft they want to attend to. Typically, they can be notified of the arrival of new aircraft in their sector both by auditory or by visual means. It is therefore important to design the visual search experiment with two presentation conditions:

- the target is presented to the subject as an auditory input (for instance, in headphones);

- the target is presented to the subject as a visual input (for instance, displayed on a window on the screen).

The effect of the means of presentation must be controlled to ensure that the recognition process is as effective when the subject starts with a visual or an auditory trace of what he is searching.

8.5 Other Text Items

All the elements on the label will be the same in terms of font type and layout conditions (they will however follow all specific principles present on the generic HMI specification document [Jackson & Pichancourt,1995]).

All the text items that could be present on the screen should have the same font as the font being tested for the labels. This in a sense is the most conservative approach or the worst case scenario, in which also other unrelated text items are behaving as distractors. The result with these test conditions will allow us to be confident of the fact that if other fonts are used for other text items they should interfere less than when the same fonts are used for all the screen items.

9. DESIGNING THE EXPERIMENT

9.1 The Tasks

The activities on labels that are normally carried out by controllers on a radar screen are very close to the type of tasks that are generally performed during visual search experiments. We can therefore transpose rather easily this experimental paradigm to font selection.

The experimental paradigm most often employed in the domain of visual search and recognition is that of asking subjects to identify as quickly as possible an item (with a certain set of features) displayed within a set of distractors. Measures are taken of the reaction time and accuracy of response. However, if eye movement recording is available, search strategies might be analysed. Eye movement data can in fact enrich the other data with information about search patterns or fixations and thus help explain some of the accuracy or latency results.

The experimental situation can be designed on the basis of the following case:

- A target call sign is presented to the experimental subject either as an auditory input (read or heard in headphones) or as a visual one (presented on a tabular list on the screen such as a communication list or as a small pop-up window in the middle of the screen). The visual presentation must always be in the same position and for the same duration. Ideally, the stimulus should be in the centre of the screen so that the visual search starts from the same position.

- After the brief auditory or visual presentation the screen is animated with traffic and the search can begin.

- The target label is displayed on the radar screen in the midst of a set of background distractors. Contrary to traditional visual search experiments where the experiments must carefully decide what the characteristics of the distractors[10] should be, in ATC the background distractors are given (other labels which have the same structure but vary in terms of the alphanumeric that are displayed).

- Time will be measured from presentation to the moment the controller clicks on the label[11]. In fact, as very often a controller looks for a label to perhaps

[10] Most often in visual search experiments the object of study is both the feature of the target object and the features of the distractors. In fact, it is widely agreed that the difficulty of a search task can be explained by the similarity relationships between targets and distractors and between different types of distractors.

[11] Data collection facilities can be included in the internal architecture of the system. Thus frequency/time data of the use of specific elements of the display (clicks on buttons or text items) can be recorded. The computer system can be logged to record and time-stamp every click on an item of

act on it, we will ask the experimental subjects to go and click on the label with his mouse, rather than just pointing at it.

More specifically, we can envisage the following set of tasks:

9.1.1 Finding a call sign never seen before

Subjects are either presented with an auditory input or a visual display which presents a call sign and are asked to find and click on the corresponding call sign on the radar screen as quickly as possible:

- between each trial the screen goes blank;
- a different set of aircraft is present on the screen at each trial;
- there is always the same number of aircraft on the screen (e.g. twenty aircraft);
- the target label is presented in a random position within the airspace.

The other distractor labels on the radar screen are for 50% of them similar in the sense of sharing at least five (out of nine) characters of the call sign. The remaining 50% of the labels are dissimilar in the sense that they share fewer than three characters.

9.1.2 Finding an information on a label previously seen

Subjects are asked to listen to an auditory input or look at a visual input which asks a question relative to an information displayed on one of the labels.

Questions could be:

- What is the exit level of flight xxxxxx?
- What is the exit waypoint of flight xxxxx?

In order to be able to answer the question the controller must find and consult the label by extending it. We will therefore log the moment in which the mouse enters the label field.

In this case the radar screen does not go blank between trials but the controller can monitor the evolution of the traffic for a few minutes so as to familiarise with the traffic. Only at that point are the questions asked and the controller must find and answer the query. This situation attempts to simulate the case in which controllers must attend to a label they have already seen and have analysed previously. In order to avoid a simple recall (where controllers answer on the basis of what they remember and, therefore, do not

the list, click on waypoints, flight levels, aircraft data, etc. At the end of a session a file is generated, with the controller's name, the date, the time, the group and the set of call signs associated to the condition.

need to consult the label), we will define a set of questions, which actually demand that the controller look at the extended label.

If eye movement is recorded measures could be taken from the moment the first eye movement is made after the audio presentation or when the eye moves away from the visual presentation.

All other conditions can be the same as the previous experiment.

A variation of this experiment could be designed to integrate with the experiments on menus. In this case the subject will be asked to modify an element on the label (for instance, by changing a level by using a menu) rather than answering a question.

9.2 The Experimental Design

The combination of the two font types with two size conditions, two presentation conditions (auditory and visual) and two background/foreground conditions (high and low contrast) make for a relatively complex design. In experimental terms we have four independent variables which each have at least two levels.

Font	Size	Stimulus	Contrast
Verdana	12 pt	auditory	highest contrast
Georgia	14 pt	visual	lowest contrast

Each font type and font size must be combined with the contrast conditions. Each condition should be tested with at least ten trials in order to provide reliable data. Considering that it is difficult to envisage an experiment with more than twelve controllers, a correlated group design should be opted for, where each subject is admitted to all the conditions and measures are repeated for each condition. This design allows us to have a small number of subjects and to control the 'performance style' of each controller.

In a repeated measures design the conditions should all be presented to all the subjects. This means that each subject will carry out at least 160 trials, which may be too much. Should the experimenters consider that they want to reduce the number of trials for each subject, it would be advisable, therefore, to opt for a mixed design in which a group of subjects will be presented for instance with one type of background and another group with another.

It is important to choose accurately what factors are being repeated and what factor is attributed to different groups. As the groups will be small comparisons between groups will be less reliable than comparisons within groups, so it is better to attribute to different groups the less critical variables (in this case, for instance, the auditory/visual mode of presentation).

9.3 Selecting the Subjects

Subjects should be experienced air traffic controllers. Given the nature of the task, competence in planning and control in general, is required. Furthermore, the overall ecological validity of results will be increased by having end-users of the system test the fonts that are being evaluated.

9.4 Controlling for Order

Some of the most important factors that must be controlled, when running an experiment that includes more than one task, are learning, practice and habituation. In fact, in any experiment with a repeated measures design, where all subjects serve in each of several experimental conditions, there is the risk of an order effect or of a carry over effect. In the former case if a sequence of tasks is presented in a fixed order there may be a practice effect whereby subjects perform better in the last few tasks than in the first few because they are learning or habituating to the tasks. In the case of the comparison of two fonts that do not differ very significantly it may be the case that subjects always perform better with the second one, because they have used the first one as a training phase or because they are more experienced by the time they start interacting with the second one. The second type of sequencing effect that can occur is the carry-over effect whereby performance in one treatment condition is partially dependent upon the condition that precedes it.

The solution we suggest in this study is to counterbalance for order so that at the end of the study the same number of subjects have done condition 1 as their first, second and third condition. However, to counterbalance for order we need to have a certain number of subjects; therefore, we suggest to use incomplete counterbalancing.

Incomplete counterbalancing is an intra-group counterbalancing technique in which some of the possible sequences of treatment are enumerated. In this case only some sequences are enumerated and under the condition that each treatment condition must appear an equal number of times in each ordinal position. In the case of four interface options, for instance, we could enumerate the following four sequences which respect the rule that each condition is presented the same number of times (in this case once) in each ordinal position: 1234, 2341, 3412, 4123. Clearly, the number of subjects required diminishes radically with this method with respect to a full counterbalancing in which there would have been 24 groups.

10. EXPERIMENTAL PLAN

The experimental session should be composed of various phases, starting with a training phase, followed by the experiment proper and closed by a debriefing interview.

10.1 Experimental Procedure

10.1.1 Training

The training schedule includes various stages. The first step consists of a dynamic demonstration of the experimental display by the experimenter with the controller in a passive role. This includes a specific presentation of the different tools, of the airspace of the functions available and an explanation of the task controllers are to perform. The experimenter demonstrates the system on a sequence of six conditions, showing how or where the visual or auditory input will appear, and how to select the label. The controllers then get a 'hands-on' experience of the equipment and the experimental task and are given the same sequence of six conditions.

10.1.2 Experimental Phase

After the training phase is completed, the controller is given a set of sixty tasks in one of the conditions depending on what group he belongs to. At the end of this first sequence a short pause (for the resetting of the system) allows him to rest, make some comments and prepare for the following set.

10.1.3 The Debriefing

A short debriefing session can be envisaged at the end of the experiment.

The actual time taken for each part of the session depends on the precise tasks that are designed and on individual controllers.

Page intentionally left blank

SECTION C: AN EXPERIMENTAL METHODOLOGY FOR LISTS

Tabular lists can contain multiple items and are organised according to a fixed order. In the case of electronic strips, for instance, a line on the list contains a set of codes, abbreviations, numerals, words, that all normally pertain to the same aircraft. It provides an organised set of information on the aircraft. The messages list is a similar list that contains syntactically more varied text.

There are other tools in which a list may contain relational information regarding two or more aircraft (Short-Term Conflict Alert [STCA]). There are lists in which the aircraft are simply present within the list when they are in a certain state (e.g. incoming aircraft list).

Page intentionally left blank

11. SELECTING THE INDEPENDENT VARIABLE: THE FONTS

11.1 A Pre-selection for Typeface

Candidate fonts for this context should have internal legibility. Internal legibility is a function of character brightness, contrast between letter and background, font size, interword spacing, line spacing, line length. Internal legibility should enable the discrimination between characters (in particular, the problematic X and K, T and Y, I and L, N, M, W, I and l, O and 0 and Q, S and 5, U and V) and the identification of single characters.

Given what we know from previous experiments and the context of the ATC displays there are two typeface and layout factors that can play a major role in letter discrimination and internal legibility:

- intercharacter spacing,
- interline spacing.

Intercharacter spacing is in fact an important component of text legibility. Typefaces that are designed for a very reduced intercharacter spacing can be immediately excluded from any experimentation because they reduce the legibility of words:

- Exclude fonts that have reduced width between characters.

- Monospaced fonts can be employed for input fields but proportional fonts must be used for all the other text types. Many items in lists are potentially input fields so this would suggest a preference for monospaced fonts. However, monospaced fonts reduce the legibility of text especially in lists or where there are rows of text. Therefore, we recommend using proportional fonts for lists.

11.2 Selecting the Relevant Layout Parameters

There is a number of variables that can interact with the typeface in the context of list displays:

- the size of type;

- the size of the interline spacing;

- the colour of the text (a legible font may become less legible if displayed in a different colour)[12];

[12] Different sizes must be used for static and dynamic objects and if colour is being used to highlight the text or the background. Furthermore, size interacts with font type. Therefore, the assessment of a font size will have to be carried out for each font type, colour and object type (dynamic or static).

- the colour of the background (a legible font may become less legible when the foreground background contrast is modified).

Of these interline spacing is a very important parameter. In the context of list displays interline spacing is a factor that interacts significantly with typeface and size. It is possibly one of the main factors to test alongside the font type. Single spacing and reduced **interline spacing** has been found to be less legible and reduce reading time. Suggested values are three to four points greater than the size of the type (8-point type needs point 11 or point 12 of interline spacing). However, if the typeface has a large x height as in Sans Serif fonts and if the video reverse is being used, interline spacing must be increased.

A very high x height such as (n h) reduces legibility because it reduces the differences between characters that have the same lower section, for instance, ‘n’ and ’h’. A low x height such as (k h) is also less legible because it reduces the lower part of the character and puts in evidence the straight ascenders which again reduces the distinctiveness of each character.

The examples below illustrate how interline spacing affects the readability of a character, and how different x heights require different sizes of interline spacing.

Three characters with different x heights and three different intercharacter spacing

	Low x height	Intermediate x height	High x height
12-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af h2222 af h2222 af h2222
11-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af h2222 af h2222 af h2222
14-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af h2222 af h2222 af h2222

Verdana in three sizes and three different intercharacter spacing

	11-point character	12-point character	14-point character
12-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af2222 af2222 af2222
14-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af2222 af2222 af2222
16-point intercharacter spacing	af2222 af2222 af2222 af2222	af2222 af2222 af2222 af2222	af2222 af2222 af2222

11.2.1 Contrast

The colour of text and background must be tested to ensure that the selected typeface allows for good reading performance in conditions in which the legibility is reduced by the modification of foreground-background contrast.

11.3 Conclusions

In conclusion, the variables that must be included in the experiments with list are the following:

- **typeface** (two or three different typefaces);
- **interline spacing** (two or three sizes);
- **character size** (two or three sizes);
- **background foreground contrast** (various levels in function of existing combinations of characters colour versus background highlighting).

Page intentionally left blank

12. SELECTING THE DEPENDENT VARIABLE

12.1 Controllers' Text Processing of Lists

12.1.1 First reading

At a first reading the controllers may read the whole line from left to right, progressively obtaining more information about the aircraft and flight plan. Most of the codes and abbreviations will be rapidly recognized, as they are well known and predictable. For instance, the flight plan will indicate the desired route by displaying a set of waypoints. The waypoint codes are obviously extremely familiar and the controller will recognise them very rapidly as a global visual configuration. The same process will also be applied to the aircraft type, which will trigger a set of associated data on the aircraft performance stored in long-term memory. Similarly, entry and exit levels will also be associated to a known configuration of the traffic.

However, controllers may only read sections of the line, concentrating on destination and flight levels, ignoring other sections or leaving them for further scrutiny. Skimming the line implies focusing only on sections of it. They are helped in this process by the familiarity with the organisation of the text in which items always are displayed in the same order and position (e.g. the call sign on the left followed by the expected time over the entry waypoint, etc.). Although it is envisaged that controllers may in the future reconfigure lists to their needs and liking, we can make the hypothesis that they will always create a stable environment where the order of the items is known and can be predicted.

12.1.2 Searching and further reading of a list

In other cases when the simple presence within the list is conveying some important information about the flight (the fact that there is a communication that has been exchanged, the fact that the flight will be arriving in the sector within a few minutes), the controllers will want to occasionally monitor the list to see if there are new items that have been added to the list. This means that the list will be rapidly skimmed to identify a new or unknown item or may be analysed in a single fixation, especially of there is nothing there.

12.2 Operational Measures of Text Comprehension

In the case of lists comprehension is the main characteristic of text processing that we will want to implement as an experimental variable. In a list the text is conveying information about a flight either through a set of descriptors associated to the call sign or by the insertion in a relational network with other flights. Controllers therefore treat the series of words and codes in sequence, which enriches the description of the aircraft, and deal with the sequential

nature of the presentation by trying to interpret each successive word or code as soon as they encounter it, integrating the new information they have obtained with what they already know about the aircraft.

In the existing literature on reading the main dependent variables in studies of text comprehension have been reading time and comprehension, as measured by:

- error detection,
- recall,
- reading speed times.

* Effective reading rate: percent correct on a comprehension test.

In the context of this sort of activity where the text is short we suggest that, rather than asking subjects to read an item and to recall or recount what he has read, subjects should be given the task to recover information from the list to answer questions regarding an aircraft or a set of aircraft. Measures will be taken of capacity to retrieve the correct information from the textual display:

- accuracy or number of correct answers;

- time to answer the question: measured as the time from the display of the query on the screen to the controller's click on the correct item.

13. CONTROLLING VARIABLES

There is a number of variables to be controlled during this experiment:

- the screen environment and traffic,
- the questions,
- the targets,
- other text items.

13.1 The Screen Environment

The composition of the work position and screen environment to be simulated during the experiment can include the radar screen or not depending also on what lists are being simulated:

1. In one case the screen only displays the list and, therefore, the controller only focuses on the part task.

2. In the second case the controller is in a more realistic environment and the lists are displayed next to, or within, the radar image.

13.2 The Traffic

In order to control for the effects of familiarity (controllers importing their extensive knowledge of traffic in a sector with which they are familiar), it will be necessary to define a synthetic airspace and an arbitrary traffic on which subjects will be trained before starting the experiment.

Even when only the lists are displayed the experiment must provide a realistic set of aircraft and flight plans. This is particularly important when experimenting with electronic strips, where the aircraft information must be coherent.

13.3 The Questions

For these experiments it will be necessary to maintain a set of homogeneous questions across conditions. While questions between trials can differ even significantly, questions between different conditions must be equivalent. Homogeneity means ensuring that the structure of the questions is similar and the complexity equivalent. Experiments in ATC have shown that traffic samples can be reused if call signs are changed.

We may therefore imagine the creation of ten types of question (which allow us to have ten trials per condition) and then simply modify the call sign between conditions:

- what entry level,
- what exit level,
- what waypoints,
- what destination,
- what origin,
- what type of aircraft,
- etc.

13.4 Other Text Items

All the elements on the label will be the same in terms of font type and layout conditions (they will however follow the general presentation principles in the generic HMI specification document (Jackson & Pichancourt, 1995). All the text items on the screen should have the same font as the font being tested for the lists. As discussed in the section on labels this will allow testing the font in the most stringent condition.

14. DESIGNING THE EXPERIMENT

14.1 The Tasks

Subjects will be asked to answer a series of questions by finding the information on the displayed list.

The type of questions will be of different sorts depending also on the type of list (whether, for instance, it is a list of strips or aircraft in conflict, etc.).

14.1.1 When the list displays electronic strips

There should be at least two kinds of questions:

- those regarding a single aircraft (the flight number is given the controller must search for the flight in the list and answer a question regarding that aircraft);
- those regarding a set of aircraft (the target information is given and the subject must find all aircraft that have that property).

Examples of questions on single flights

- What is the exit level of flight AF4567?
- What type of aircraft is flight AF7890?
- At what level will flight AF1234 be entering?

Examples of questions regarding a set of flights

- Indicate all flights that will exit at waypoint XXX.
- Highlight all aircraft at level 280.

14.1.2 When the list displays relational information (e.g. conflict or approach)

- With which aircraft is AF4567 in conflict?
- What is the approach order?
- In what position is AF45678?

14.1.3 When the list displays messages

Message lists display the most syntactically complex texts of the whole controller interface. The text partially resembles natural language; however, it is automatically generated and has a very predictable structure.

The questions here can resemble:

- For what aircraft has a new entry level been requested?
- What change of route has been requested for AF12345?

14.2 The Experimental Design

The combination of the two font types with various size conditions (12 point, 14 point), two or more colour conditions of the character, two contrast conditions (where two extreme cases should be chosen among the most and least contrasted foreground/background combinations) and three interline spacing conditions (point 10, point 8) make once again, for a relatively complex design.

Each condition should be tested with at least ten trials in order to provide reliable data. We recommend to opt again for a mixed design in which a group of subjects will be presented for instance with one type of background and another group with another.

For instance, in the case described here we would envisage two groups of different subjects (for example, grouped by background colour). The advantage of having one group performing all trials with a neutral background and one group with a highlighted background is that we would have a set of consistent results within each background configuration.

14.3 Selecting the Subjects

Subjects should definitely be air traffic controllers. Given the nature of the task competence in planning and control in general is required. No particular conditions are necessary relatively to the profile of the sample, given that all the tasks are part tasks in which different levels of experience should not affect performance. Furthermore, if a synthetic airspace is used all subjects will be equally unfamiliar with the sector and traffic.

14.4 Controlling for Order

The experimenter must control learning effects that could appear if the conditions were always presented in the same order. The solution we suggest in this study is to counterbalance for order using incomplete counterbalancing (see 9.2 for a presentation).

15. EXPERIMENTAL PLAN

The experimental session is generally composed of phases, starting with a training phase, followed by the experiment proper and closed by a debriefing interview.

15.1 Experimental Procedure

15.1.1 Training

A brief training session will be necessary, including a specific presentation of the different tools, of the airspace of the functions available and an explanation of the task controllers are to perform. The controller then gets a 'hands-on' experience of the equipment and the experimental task. All controllers are given the same sequence of six conditions.

15.1.2 Experimental phase

After the training phase is completed, the controller is given a set of twenty questions in one of the conditions depending on the group to which he belongs. According to the number of conditions that each subject is presented with, a new set of questions is employed.

15.1.3 The debriefing

A short debriefing session should be envisaged at the end of the experiment.

Page intentionally left blank

SECTION D: AN EXPERIMENTAL METHODOLOGY FOR MENUS

Many menu items are in fixed positions on the screen. However, in the case of radar label menus the menus follow the call sign, although they are retrievable always from the same position within the label. Therefore, the call sign functions as a menu title opening on a menu but also as a reference position point.

There are three main types of menus:

- function menus,
- parameter menus,
- toolbox menus.

An example of a function menu is the call sign menu which includes accept, reject, transfer functions. These menus provide control options and coordination options on the aircraft.

Parameter menus include menus that allow controllers to change a flight parameter such as flight level, waypoints, speed, direction, etc. These are generally accessed from the labels of individual aircraft or from list items.

Finally, there are menus which include toolboxes that give access to a set of control tools.

The content of the menus is either a list of numerals (as in the case of flight level menus), a list of codes (as in the case of waypoint menu) or a list of single words (as in call sign menus). The critical issue in reading menu text is discriminating between items in the list.

Page intentionally left blank

16. SELECTING THE INDEPENDENT VARIABLE: THE FONTS

Because of the fixed vertical structure of menus the navigation within a menu can be directed both by semantic and motor processes. For instance, in a menu containing flight levels controllers use their knowledge of the sequence of flight levels and of the sequential structure of the menu to calibrate their movements when they perceive the numerals 280 and are trying to reach flight level 310. The processing of text in this context, therefore, is not just pattern matching between a target numeral and the numerals that scroll by; the text becomes a sort of waypoint for the controllers' trajectories in the menu, guiding the direction of their movements.

Legible fonts in this situation are those that not only support rapid identification of the target, but also an effective motor response.

Menu lists often contain items that are very similar because they belong to the same category (e.g. a list of numerals that follow each other 270, 280, 290). It is important that the characters support the discrimination between strings, discriminating clearly between strings such as 280 and 230. This implies that the single characters must be very distinct.

Another issue is that lists are the only items that are not only monitored visually, but also involve an action and a modification on the part of the controller. Candidate fonts will ensure that the motor action of selecting the item on the menu list can be done automatically without needing a continuing and focused perceptual control.

This again means that fonts must also have a good level of internal legibility, which supports the discrimination of single letters and avoids the confusion between similar characters such as 1 and l or I.

Candidate fonts should be:

- existing fonts used on menus (as benchmarks),
- standard UI interface fonts used on PC menus (e.g. Helvetica),
- fonts identified as satisfactory candidates for screen display.

16.1 Selecting the Relevant Layout Parameters

There is a number of variables that can interact with the typeface in the context of list displays:

- the size of type,
- the interline spacing,
- the background foreground contrast level.

The size of the font is a particularly important parameter. In menus, when items are selected using a mouse, selection time is a function of the distance

of the cursor to the target and the size of the target according to Fitts' Law (Jackson & Pichancourt, 1995). The time to acquire a target is a function of both the distance to target and the size of the target. Consequently, rapid access may be achieved by making items closer to the initial cursor position and larger in physical size. However, pull-down menus have the disadvantage in that the larger the target the greater the distance to all but the first item in the list. It is worth while experimenting on various sizes in order to find the right balance between screen real estate (Fitts' Law) and avoiding an excessive length of the menu.

Interline spacing can greatly improve or deteriorate readability. Text that is too tightly spaced becomes indistinct, and users are unable to see where the descenders end and where the ascenders of the following item start.

In conclusion, the main independent variables will be:

- font type (two or more types),
- size (two or more sizes),
- interline spacing (two or more sizes),
- contrast (two or more foreground background combinations)[13].

[13] Menus also include horizontal separators, usually lines between menu items. The visual strength (in terms of contrast) of the separators may have an effect on the readability of the characters and on the global foreground/background contrast.

17. SELECTING THE DEPENDENT VARIABLE

17.1 Controllers' Text Processing of Menus

Significant aspects of the processing of menus include recognising text items, discriminating between items, and selecting a target. Consequently, the main criteria which fonts must satisfy in this context is to effectively support the identification, discrimination and the perceptual-motor coordination to select the item.

17.1.1 Searching and recognising a known menu item

In many cases the controller will be opening a menu and scrolling within it to search a numeral or a word before selecting it. The items within a menu are generally well known; the user has seen them many times before. Furthermore, the label on the menu provides a first categorization of the items and therefore restricts the number of expected words or codes. As he opens a menu, the controller already makes an hypothesis of what the items in the menu list will be because he knows what are the set of potential items that are contained in the menu he has selected.

17.1.2 Discriminating between items in a menu

The items in a menu often belong to a same category or family (e.g. a list of possible flight levels), a list of waypoints, a list of numerals indicating headings or speed. The discrimination between items of the same category or numerals, which are on a continuum, can be very difficult. There are many studies showing that similar categories are more difficult to discriminate than similar words of totally different categories. Particular care must therefore be put in evaluating if a font type supports the discrimination between items within a menu.

17.2 Operational Measures

The dependent variables or measures taken will be the following:

- **Time to target** (latency measured between the moment a query to modify the flight plan is formulated or visualised on the screen and the moment the controller clicks on the menu item).

This measure can be decomposed in two measures:

- **Time to identify a target** (latency measured between the moment the query is uttered or visualised on a display and the moment the controller clicks on the menu title).

- **The time to select item from the moment in which the cursor is positioned on the menu 'title'** (this measures the search and selection time within the menu).

Finally, there must be a measure of accuracy:

- **Accuracy (measured as the number of correct selections in a series of repeated trials)**. However, we must be aware that in ATC as with any system in which operators are highly skilled and trained it is difficult to observe major errors in handling the system.

18. CONTROLLING VARIABLES

There is a number of variables to be controlled:

- the screen environment and traffic,
- the tasks,
- the measure of time,
- other text items.

18.1 The Traffic

Again, in order to control for the effects of familiarity (controllers importing their extensive knowledge of traffic in a sector with which they are familiar), it will be necessary to define a synthetic airspace and an arbitrary traffic, on which subjects will be trained before starting the experiment.

18.2 The Type of Task

For this group of experiments it will be necessary to maintain a set of homogeneous tasks across conditions. While tasks between trials can differ significantly, the requests (of flight plan modification) between different conditions must be equivalent. Homogeneity means ensuring that the structure of the tasks are similar and the complexity is equivalent. As in previous experiments we can probably rely on the strategy of simply modifying the call signs between conditions while maintaining an identical traffic. We may therefore imagine creating different types of tasks (which allow us to have ten trials per condition) and then simply modify the call sign between conditions. The order of presentation of the tasks will be randomised.

18.3 Measures of Latency

As we have discussed above one of the principal measures to compare fonts will be the latency between a request and the selection of a menu item by the controller. This interval will presumably be very short and measured in milliseconds. It is very important, therefore, to be able to time-stamp very precisely the start and end of the interval. The system log can allow a precise time-stamp of the moment in which a menu item is selected. Time-stamping the production of the request for change, however, is more difficult, especially if it is done verbally. Visual presentation of a request on a small pop up window can be logged precisely. The start of verbal utterances is difficult to control unless it has been pre-recorded and presentation is controlled by the computer which time-stamps the start or end of the recording.

This problem does not apply to the measure of the time taken from the opening of the menu to the selection of an item in the menu. It only applies to

the first phase in which the controller must find the label, extend it if necessary and position the cursor on the menu.

18.4 Other Text Items

All the elements on the label will be the same in terms of font type and layout conditions (they will however follow all specific principles present in the generic HMI specification document [Jackson & Pichancourt, 1995]). As discussed previously all the text items on the screen should have the same font as the font being tested for the lists.

19. DESIGNING THE EXPERIMENT

19.1 The Tasks

Subjects will be asked to carry out a series of tasks involving the selection of a menu item.

Subjects are either presented with an auditory input or a visual display which presents a call sign and a request for a modification in the flight plan.

The type of tasks will differ according to the menu content.

19.1.1 When the menu displays information about an aircraft

- Tasks regarding the modification of a parameter of the aircraft (the call sign is given; the controller must search for the label, extend it if necessary, position the mouse on the relevant menu, open it and select another value):

  - modify entry level of XXXXXX,
  - modify exit level of XXXXXX,
  - modify waypoints of XXXXXX,
  - modify speed of XXXXXX.

- Tasks regarding the modification of the state of the aircraft (the call sign is given; the controller must search for the label, position the mouse on the relevant menu, open it and select another value):

  - accept XXXXXX,
  - release XXXXXX,
  - request XXXXXX on frequency,
  - coordinate exit level of XXXXXX.

19.1.2 When the menus displays a selection of tools in a toolbox

The task in this case can be of finding, selecting and applying the tool to an interface object:

- select tool x and apply it to object x.

The measures will be instead taken only on the latency between request and selection of the tool in the menu.

19.2 The Experimental Design

The variables that we have suggested are the following:

- two font types;
- two size conditions (point 12, point 14);
- two or more contrast conditions (highest and lowest foreground/background contrast conditions);
- two interline spacing conditions (10 point, 12 point).

The four variables and various conditions make once again for a relatively complex design.

Each condition and combination should be tested with at least ten trials in order to provide reliable data. We recommend to opt again for a mixed design in which a group of subjects will be presented, for instance, with one size of interline spacing and another group with another, and all the other conditions are instead presented to all the subjects in a repeated measures protocol.

Obviously, the experiment must again control learning effects that could appear if the conditions were presented always in the same order. The solution we suggest in this study is to counterbalance for order by incomplete counterbalancing (see 9.2 for a presentation).

19.3 Selecting the Subjects

Subjects are air traffic controllers. No particular conditions are necessary relatively to the profile of the sample if the traffic used is synthetic or in any case unfamiliar to all the subjects.

19.4 Experimental Procedure

As for the previous evaluations the experimental session will be composed of phases, starting with a training phase, followed by the experiment proper and closed by a debriefing interview.

There are however two experimental procedures that could be used:

⇨ In the first case a more ‘classical’ experimental protocol is proposed:

- between each trial the screen goes blank;
- a different set of aircraft is present on the screen at each trial;
- there is always the same number of aircraft on the screen (e.g. twenty aircraft).

⇨ The second case could be constructed more as a mini simulation, in which the synthetic traffic evolves over a period of time (e.g. twenty minutes) and all the menu selection trials are done on the same traffic.

In this case the radar screen does not go blank between trials but the controller can monitor the evolution of the traffic for a few minutes so as to become familiar with the traffic. Only at that point the requests are presented and the controller must find and modify the flight plan. The number of requests corresponds to the number of trials, of which there should be at least ten for each condition.

This second case may however introduce an effect due to the fact that, as time goes by, the latency in the actual search of the label will be reduced because the controllers will become more and more familiar with the traffic and therefore reduce their search time.

Page intentionally left blank

SECTION E: PUTTING IT ALL TOGETHER IN A SIMULATION

After having carried out the part-task experiments discussed in the previous sections it may be necessary to envisage a test that puts together the various text types and control activities in a more global simulation.

By simulation we mean a real-time simulation of at least one control position in which the controller is presented with an environment that allows him to carry out the greatest possible variety of control activities on a realistic traffic.

There is a number of reasons that justify the need to test fonts within a simulation at this point of the experimental process of font selection:

1. To ensure the ecological validity of the results in a more realistic task environment.

2. To verify the interactions between chosen solutions if the fonts are not the same for all items (for instance, if the best candidate for menus is Helvetica and the best candidate for labels is Verdana).

3. To verify if there are no strong deviations from the results found in the part-task experiments when the part task is executed in a larger context of activities.

4. To measure controllers' subjective evaluations of the fonts in an environment that resembles more closely their normal operational conditions in terms of task complexity, time pressure and interface layout.

Page intentionally left blank

20. SELECTING THE INDEPENDENT VARIABLE: THE FONTS

The process of font selection should at this point have been carried out for the individual text types. The issue here is to evaluate the effect of their combination on the same screen. All the fonts and layout conditions that have been chosen for the individual text items (menus, labels, messages, etc.) should be combined and tested jointly. If there are too many different typefaces present on the screen once they are all combined, it may be useful to compare different combinations.

There are various test solutions:

- all the ‘best candidates’ are tested jointly and compared to existing fonts and display conditions;
- each ‘best candidate’ is tested in isolation with all other text items displayed in the same font;
- combinations of candidates are displayed together and compared to existing displays.

The main independent variables will therefore be:

- each ‘best candidate’ resulting from previous tests,
- combination of all ‘best candidates’,
- existing fonts on radar screen or fonts provided by screen manufacturer.

Page intentionally left blank

21. SELECTING THE DEPENDENT VARIABLE

21.1 Controllers' Text Processing Overall

During the global activity of ATC controllers are engaged in numerous processes of text recognition and treatment. If the simulation environment allows the controllers to carry out all the main control activities, even during a small fraction of the simulation, we should be able to observe the use and processing of all the items previously examined (search and identification of labels and list items, selection of menu items, reading of messages, etc.).

The issue here is to find the correct operational evaluation techniques to measure the different processes in the context of the multiple activities that are being performed.

21.2 Operational Measures

All of the measures taken during the part task to compare and evaluate different solutions for each text type should be replicated here.

In Sections 21.2.1 to 21.2.4 the dependent variables or measures taken are described.

21.2.1 For the labels

We will measure each time there is a request from a pilot which leads the controller to position his mouse on the label in order to respond to a request to modify in some way the status of the aircraft or the flight plan.

- **Time to identify a target** (latency can be measured in a number of ways):
  - between the moment the pilot calls to announce his entry to the moment in which the controller puts the cursor on the label to assume;
  - if eye tracking equipment is available, time from the moment the eye moves away from the last fixation point during the auditory presentation to the moment in which the call sign is fixated.
- **Accuracy** (measured as the number of correct identifications of the labels at first attempt).
- **Eye movements** (points of most frequent fixation and eye movement paths during search).

21.2.2 For the menus

- **Time to target** (latency measured between the moment the menu is opened and the moment the controller clicks on the menu item). Every time a menu is opened a log will be taken of the time to reach and select a menu item.

- **Accuracy** (measured as the number of correct selections).

21.2.3 For the lists

It is particularly difficult to envisage a way of measuring text processing in lists during a simulation, especially for measures of search and retrieval time. In fact, it is most often the case that controllers consult lists when they need to check an information or in the case of strips when they want to examine a flight plan. Controllers usually do not rely on an external request to consult the list but do so at their own need. This means that there are no external events to mark the onset of the search and processing of the text item.

However, if eye tracking is used data could be collected on the points of fixation. This data could give useful information on what elements of the list are being consulted and for how long. Fixation time could then be compared across conditions thus providing an indication of the level of legibility of the font types being tested.

21.2.4 Subjective evaluation

Another important measure taken in this type of global test is controllers' subjective assessment of the fonts. In a real-time simulation measuring user's satisfaction makes more sense than in part-task experiments, because the controllers can evaluate the readability and legibility of the font in the global experience and can better assess the impact of the font on their activity as a whole.

Typical measures of acceptability and satisfaction can be taken; such as:

- debriefing interview sessions,
- satisfaction questionnaire at the end or during of the session,
- subjective assessment of visual fatigue and strain.

22. CONTROLLING VARIABLES

There is a number of variables to be controlled namely the traffic and control environment, and how measures are taken.

22.1 The Traffic

As in all simulations the really difficult issue is to ensure a relatively homogeneous control session across different subjects, knowing that after a few minutes the control decisions taken by each controller can lead to a significantly different traffic configuration. However, given that the measures taken will regard specific sub-tasks concerning text processing, the effect of major differences of traffic configuration should not be felt. It will be important to decide how many and which events will be included in the comparison and which will be excluded.

22.2 The Measures of Time

As we have discussed previously, the latency between a request and the selection of a menu item by the controller will be very short and measured in milliseconds. It is very important, therefore, to be able to time-stamp very precisely the start and end of the interval. The system log can allow a precise time-stamp of the moment in which a menu item is selected. Time-stamping the production of the request for change, however, is more difficult, especially if it is done verbally. In this simulation scenario most inputs will be provided by 'pilots'. It will be particularly difficult to ensure that their requests are uttered at a planned time. It is therefore more realistic to ask them to click on a list which contains all of the expected requests, or in a window while they utter their request.

As discussed previously we suggest use of eye tracking equipment to measure very precisely the on-start of the search process as the moment in which the eye first moves away from the point fixated during the auditory request from the pilot.

22.3 The Experimental Environment

The radar screen

Real displays and screens must be used to ensure that the fonts appear as they will on the real equipment.

The human-machine interaction

For all conditions the display should present at least the basic HMI features and functionalities to be found, for instance, in the proposed target HMIs for future ATM systems (Jackson & Pichancourt, 1995).

Simulated traffic

In the simulation there are two options in terms of traffic: a synthetic airspace or a real airspace. A synthetic airspace allows a better control for the effects of familiarity (controllers importing their extensive knowledge of traffic in a sector they are familiar with) on which subjects will be shortly trained before starting the simulation. A real airspace requires a more stringent selection of subjects to ensure similar levels of familiarity.

23. DESIGNING THE EXPERIMENT

23.1 The Tasks

In a simulation the task is simply to control the traffic on the sector for a certain period of time.

23.2 The Experimental Design

The design will be a repeated measures design. In a repeated measures each subject performs all the tasks for the different conditions and comparisons are made between results at each condition. In this context, controllers will be performing two or more simulations with the different conditions (combination of font candidates, one best candidate for all the text items, existing font types) but will also be repeating each sub-task (such as selecting a menu item or finding a label) numerous times during the simulation on different items of the interface.

There will be at least two simulations 'played' by each subject. One simulation with, for instance, the full combination of candidates and another with a benchmark condition given by the fonts provided by the manufacturers or existing font sets.

Within each simulation a set of measures will be taken for each text type. Decisions should be taken before the experiment on how many measures per text type will be used for the comparison and if all exemplars of a certain interaction with a text type will be considered acceptable. For example, will all menu selections be considered as trials of the repeated measures regarding menu selection?

Addressing these questions before the evaluation is important for two main reasons:

1. Data must be coherent with those collected during the part-task experiments, in order to be able to compare the results of the simulation with the part-task experiments. One of the objectives is in fact to establish that putting the font candidate in a global context does not create deviations from the results found in the part experiments.

2. Data must be coherent between simulations in order to compare the two main conditions. In fact, for instance if, in one simulation the majority of the menu selections consist in a change of flight level and in the other simulation the majority of menu selections concern lists or toolboxes, the data will be difficult to compare.

23.3 Selecting the Subjects

Subjects are air traffic controllers. A homogeneous group in terms of number of years of experience is preferable. However, it is possible to envisage a range of different levels of experience, in order to verify if the solutions chosen are adapted to controllers with a different spectrum of experience. This option increases the variance of the results and is acceptable only if the experimental design is a within subjects design (where each subject performs all the tasks and comparisons are made between results of each subject across conditions). No particular supplementary conditions are necessary relatively to the profile of the sample if the traffic used is synthetic or in any case unfamiliar to all the subjects. If real sectors are used it is imperative that all subjects should have similar level of experience with the chosen sector.

23.4 Controlling for Order

In the simulation it will be possible to control the order of presentation of the main conditions, letting some subjects run the simulation, first in one condition and then in the other, and letting another group start by second condition first. Within the simulation the order in which single text items will be processed will not be controlled but will depend on the control strategy adopted by the single controllers. Obviously, there will be some common sequencing due to the timing with which some aircraft will appear on the radar screen. If the traffic is maintained the same across simulations all the controllers will attend to and assume a certain aircraft at the same time when the pilot announces himself.

23.5 Experimental Plan

A simulation has a similar temporal structure as other experiments and includes a phase of training or familiarisation starting with a training phase, followed by the simulation proper and closed by a debriefing interview.

A simulation often lasts a minimum of thirty minutes to allow for the initial incoming aircraft to cross and exit the sector.

The simulation phase has an estimated duration, which is not only determined by the duration of the incoming traffic provided, but also by the duration imposed by the controllers in their activity.

ANNEX: VISUAL DISPLAY STANDARDS

BS EN 29241-3:1993 (ISO 9241) Part 3 Visual Display Requirements

The purpose of this standard is to specify the ergonomics requirements for display screens which ensure that they can be read comfortably, safely and efficiently to perform office tasks. (Source System Concepts Ltd.).

Design Requirements and Recommendations	Measurement
Design viewing distance	minimum 400 mm for office tasks
Maximum line of sight angle	less than 60o below horizontal
Angle of view within which display is legible	at least 40-deg from normal to display surface
Character height	16' minimum, 20' to 22' preferred
Stroke width	1/6 to 1/12 of character height
Character width-to-height ratio	between 0.5:1 and 1:1 is required but between 0.7:1 to 0.9:1 is recommended
Raster modulation (for Cathode Ray Tube [CRT] displays)	cm not to exceed 0.4 for monochrome, 0.7 for colour (0.2 preferred for either)
Fill factor (for non-CRT matrix displays)	at least 0.3
Character format	minimum 5x7 for numeric and upper case; minimum 7x9 where legibility is important
Extension of matrix for diacritics/ descenders	2 pixels
Subscripts and superscripts	minimum 4x5 matrix
Character size uniformity	not vary by more than 5% anywhere
Intercharacter spacing	minimum one pixel or stroke width
Interword spacing	minimum space equivalent to capital N

Design Requirements and Recommendations	Measurement
Interline spacing	minimum of one clear pixel
Linearity	less than 2% variation in row/column length; less than 5% of character height displacement
Orthogonality	less than 0.02 difference of mean height or width of addressable area (0.04 for diagonals)
Display luminance	minimum of 35cd/m2, higher preferred
Luminance contrast	minimum 0.5 contrast modulation cm
Luminance balance	average luminance ratio less than 10:1 for frequently viewed areas
Glare	should be avoided without jeopardising luminance or contrast requirements
Image polarity	either polarity is acceptable
Luminance uniformity	not to exceed 1.7:1 for display or 1.5:1 for individual character element
Luminance coding	at least 1.5:1 to be distinguishable
Blink coding	for attention, 1 to 5 Hz, 50% duty cycle; for reading, 1/3 to 1 Hz, 70% duty cycle
Temporal instability (flicker)	flicker free to at least 90% of population
Spatial instability (jitter)	location within 0.0002mm per mm of design; viewing distance in range 0.5 Hz to 30 Hz

The Display Screen Regulations require displays to be clear, legible and stable under normal working conditions. Displays which meet BS EN 29241-3: 1993 satisfy this design requirements.

REFERENCES

Broadbent, S. (1999) *Font Requirements for Next Generation ATM Systems*. EEC Report. France: EEC.

Broadbent, S. (1993) *Harmonisation of Man-Machine Interface Experiments in the Context of PHARE Advanced Tools*. EEC Report 256. France: EEC.

EATMP Human Resources Team (2000) *Font Requirements for Next Generation Air Traffic Management Systems*. HRS/HSP-006-REP-01. Released Issue. Ed. 1.0. Brussels: EUROCONTROL.

Jackson, A. & Pichancourt, I. (1995) *A Human-Machine Interface Reference System for En-route Air Traffic Control*. EEC Report n° 292. France: EEC.

Makins, N. & Broadbent, S. (1993) *An Experimental Evaluation of Traffic Filtering*. EEC Report N° 265. France: EEC.

Kirk, R. (1982) *Experimental Design and Statistics*. Second Edition, Belmont CA: Brooks Cole Publisher.

Siegel, S. (1988) *Non-parametric Statistics for the Behavioural Sciences*. Second Edition. New York: McGraw Hill.

Page intentionally left blank

ABBREVIATIONS AND ACRONYMS

ATC	Air Traffic Control
ATM	Air Traffic Management
ATM R&D CoE	ATM R&D Centre of Expertise *(EEC)*
CRT	Cathode Ray Tube
CWP	Controller Working Position
D	Distance
DIS	Director(ate) Infrastructure, ATC Systems & Support *(EUROCONTROL Headquarters, SDE)*
DIS/HUM	*See HUM Unit*
DMD	Digital Micromirror Device
Doc	Document
EATCHIP	European Air Traffic Control Harmonisation and Integration Programme *(now EATMP)*
EATMP	European Air Traffic Management Programme *(formerly EATCHIP*)
ECAC	European Civil Aviation Conference
EEC	EUROCONTROL Experimental Centre *(France)*
EWP	EATCHIP/EATMP Work Programme
HFSG	Human Factors Sub-Group *(EATCHIP/EATMP, HRT)*
HMI	Human-Machine Interaction
HRS	Human Resources Programme *(EATMP, HUM)*
HRT	Human Resources Team *(EATCHIP/EATMP)*
HSP	Human Factors Sub-Programme *(EATMP, HUM, HRS)*
HUM	Human Resources (Domain)*(EATCHIP/EATMP)*
HUM (Unit)	Human Factors and Manpower Unit *(EUROCONTROL Headquarters, SDE, DIS; also known as DIS/HUM; formerly stood for 'ATM Human Resources Unit')*

ICAO	International Civil Aviation Organization
MT	Movement Time
PHARE	Programme for Harmonised ATM Research in EUROCONTROL
R&D	Research and Development
REP	Report *(EATCHIP/EATMP)*
SDE	Senior Director, Principal EATMP Directorate *or, in short,* Senior Director(ate) EATMP*(EUROCONTROL Headquarters)*
STCA	Short-Term Conflict Alert *(ICAO)*
W	Width

CONTRIBUTORS

Reviewers

Alexandra DORBES	EUROCONTROL Experimental Centre, ATM R&D Centre of Expertise
Alistair JACKSON	EUROCONTROL Experimental Centre, ATM R&D Centre of Expertise

The Members of the HRT Human Factors Sub-Group (HFSG)

Document Configuration

Carine HELLINCKX	EUROCONTROL Headquarters, DIS/HUM

Page intentionally left blank