E.02.24-MOTA-D3.1-Experimental Protocol

Document information	
Project Title	Project Title
Project Number	E.02.24
Project Manager	ENAC
Deliverable Name	Experimental Protocol
Deliverable ID	D3.1
Edition	1.0
Template Version	03.00.00
Task contributors	
ISAE, ENAC	

Abstract

This document details the experiment plan for the validation of Project MoTa. There are three human-in-the-loop experiments planned: baseline (June-Oct 2014), HMI0 (Oct-Dec 2014), and HMI1 (Sept-Oct 2015). The design of experiments, independent and dependent variables, simulation, scenarios, and planned analyses are presented. Each of the participants (likely from the ATCO program at ENAC) will undergo a 2h testing session. All will be conducted at ENAC. Analysis of three possible baseline data collection methods is also provided.

Authoring & Approval

Prepared By - *Authors of the document.*		
Name & Company	Position & Title	Date
Zarrin CHUA / ISAE	Project member	12/03/2014

Reviewed By - *Reviewers internal to the project.*		
Name & Company	Position & Title	Date
Mathieu COUSY / ENAC	Project Coordinator	04/07/2014
Fabien ANDRE / ENAC		

Reviewed By - *Other SESAR projects, Airspace Users, staff association, military, Industrial Support, other organisations.*		
Name & Company	Position & Title	Date
Railane BENHACENE / ENAC	Head of Aeronautical HMI program	07/07/2014

Approved for submission to the SJU By - *Representatives of the company involved in the project.*		
Name & Company	Position & Title	Date

Rejected By - *Representatives of the company involved in the project.*		
Name & Company	Position & Title	Date

Rational for rejection
None.

Document History

Edition	Date	Status	Author	Justification
V 0.1	12/03/2014	Draft	ZK Chua	New Document
V 0.2	02/04/2014	Draft	ZK Chua	Modifications after project members reviews
V 1.0	3/07/2014	Draft	ZK Chua	Modifications after Eurocontrol review

Table of Contents

List of tables

List of figures

Executive summary

This document details the experimental protocol for the validation of Project MoTa. The validation program consists of three planned experiments. We present the individual experimental protocols, a discussion of the scenario definition, and include our analysis and rational for the inclusion of a tertiary experiment. The first experiment (evaluation of current air traffic control ground operations) will occur between June and September 2014. The second experiment (evaluation of the HMI0 platform) is planned for October 2014. The third and last experiment (evaluation of the HMI1 platform) is planned for October 2015. Between each of these three experiments, smaller studies will be conducted to evaluate individual aspects of the HMI platform or to solicit performance measurements from current and former air traffic controllers from Roissy airport.

This document is split into two halves. The first half discusses the validation campaign as a whole, beginning with the design of experiments, the independent variables, the dependent measures, and the anticipated participants. We also discuss the simulation platform and the experimental scenarios. The second half discusses each of the three experiments in greater detail, including a planned timeline of experiment procedure events. The baseline data collection section, in particular, details the plans for three data collection methods: using aircraft logs, a visit to Roissy airport, and an experiment. The analysis and decision process is included. This decision to conduct a third experiment was made in April 2014.

The project validation experiment plan is a 2 × 3 mixed level factorial design, with six possible scenarios. Scenario complexity is set at two levels: Average and Hard. Taxiing technique is set at three levels: baseline, interface with no automation, and interface with automation. Taxiing technique is varied between subjects and scenario complexity is the within subjects variable. The Average scenario represents the majority of working days at Roissy and is primarily defined as clear visibility, accessibility to all taxiways, no change in runway orientation, and no precipitation. The Hard scenario represents the most difficult type of working days: winter conditions require the use of de-icing stations, wind forces a change in runway orientation several minutes into the scenario, low visibility procedures are in place, and certain taxiways are closed due to the use of de-icing stations. The taxiing techniques consist of a mix of path suggestion and conflict detection (from a multi-agent system), an interactive interface, and management of automated taxiing vehicles such as TaxiBots and eTaxis.

Overall, there are 17 key dependent variables of interest, which can be separated into four categories: operational, behavioral, subjective, and neurophysiological. Operational measures are metrics of taxiing effectiveness. Behavioral measures are observable actions taken by the participants. Subjective measures are quantitative evaluations of qualitative, participant recalled, psychological phenomena. Neurophysiological measures are quantitative measures of the changes in cerebral usage, heart rate, and eye movements. The changes in these dependent measures will be analyzed using various statistical analyses as appropriate. Fundamentally, we are trying to prove that the taxiing techniques have significant favorable impact on each of the three groups of dependent measures, across the two scenario complexities.

The three experiments will each be about 2 hours long and be conducted with air traffic control students and instructors at Ecole Nationale de l'Aviation Civile (ENAC). These students are trained for large airports such as Roissy and have received previous training on managing ground operations. We aim to have at least twenty students in every experiment, for a total of sixty individual participants. Each participant will receive training specific to Roissy airport and perform two scenarios. Afterwards, they will be debriefed on the experiment.

1 Introduction

1.1 Purpose of the document

This document details the experimental protocol for the validation of Project MoTa. The validation program consists of three experiments. We present each of the three experimental protocols and a discussion of the scenario definition. The rationale for the inclusion of the third experiment is also included. The first experiment (evaluation of current air traffic control ground operations) will occur between June and September 2014. The second experiment (evaluation of the HMI0 platform) is planned for October 2014. The third and last experiment (evaluation of the HMI1 platform) is planned for October 2015.

1.2 Structure of the Document

Section 2 provides a detailed overview of the experimental plan for the validation of Project MoTa, as conducted over three human-subject experiments. This overview covers the overarching research questions, the data collection plan, and the dependent variables which will be measured in every experiment.

Section 3 presents the design of the scenarios of interest, including hypotheses, rationale, and validation.

Sections 4, 5, and 6 discuss the experimental plan for testing the current ground operations, the MoTa platform HMI0, and HMI1, respectively. Each section presents hypotheses, the experimental procedure, and planned statistical analysis.

Appendix A concerns the collection of baseline data for use in the Project MoTa validation. All three methods for data collection are detailed, including identification of potential risks. This section ends with an analysis of advantages and disadvantages and associated workload demands.

1.3 Acronyms and Terminology

Term	Definition
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
AWOS	Automated Weather Observing System
E-ATMS	European Air Traffic Management System
ECG	Electrocardiography
EGTS	Electric Green Taxiing System. Using the Auxiliary Power Unit (APU) generator to power motors on the main wheels, the electric green taxiing system allows aircraft to push back without a tug and then taxi without requiring the use of the main aircraft engines. One wheel on each main gear is equipped with an electric motor
GND	GROUND: ATC controlling position in charge of all the a/c from the block or gate to the runway and backwards
HF	Human Factors
HR	Heart Rate

Term	Definition
HRV	Heart Rate Variability
HMI	Human Machine Interface
ICAO	International Civil Aviation Organization
LOC	LOCAL: Local Control (known to pilots as "Tower" or "Tower Control") is responsible for the active runway surfaces. Local Control clears aircraft for takeoff or landing, ensuring that prescribed runway separation will exist at all times.
MAS	Multi Agent System
MoTa	Modern Taxiing
SA	Situation Awareness
SART	Situation Awareness Rating Technique
SJU	SESAR Joint Undertaking (Agency of the European Commission)
SJU Work Programme	The programme which addresses all activities of the SESAR Joint Undertaking Agency.
SESAR	Single European Sky ATM Research Programme
SESAR Programme	The programme which defines the Research and Development activities and Projects for the SJU.
TaxiBot	A aircraft tractor controlled by the pilot from the cockpit or fully automated that pulls aircraft on ground without using aircraft's engine power.
NASA TLX	NASA Task Load Index

2 Project MoTa Validation

2.1 Validation Objectives

The success of Project MoTa is defined by the achievement of several main goals. The tool must demonstrate improvement in operations in current and future operations of ground traffic control through the use of a new human machine interface and automated taxiing techniques. As with any new human-machine interface system, the tool must appeal to ATCOs for use in current and future ground operations and support a sufficient range of operational environments and ATCO cognitive complexities.

These aims provide insight into the independent and dependent variables of interest, relevant hypotheses, and how we plan to statistically support the tool validation. These topics are discussed in Sections 2.2, 2.3, and 2.4, respectively.

2.2 Design of Experiments

The project validation experiment plan is a 2 × 3 mixed level factorial design, with six possible scenarios. Scenario complexity is set at two levels: Average and Hard. Taxiing technique is set at three levels: baseline, interface with no automation, and interface with automation. Taxiing technique is varied between subjects and scenario complexity is the within subjects variable. Table 1 presents the six possible scenarios and the expected testing period. All runs within each of the three testing sessions will randomized and balanced to reduce noise and learning effects.

Table 1: Design of Experiments

Taxiing Technique	Scenario Complexity	Significance	Testing session
Baseline	Average	Current operations	Baseline period (June 2014)
Baseline	Hard	High traffic loads	
Interface, no auto	Average	Impact of improved interface on current operations (immediate impact)	Concept Validation HMI0 (October 2014)
Interface, no auto	Hard		
Interface, auto	Average	Impact of automated taxiing techniques, short- and long-term future	Concept Validation HMI1 (October 2015)
Interface, auto	Hard		

The Average and Hard scenarios are defined in Section 3.3. Essentially, these scenarios reflect the complexity of taxiing operations and also encompass the full span of working conditions experienced and anticipated at Roissy Airport. The taxiing techniques are defined based on the usage of a redesigned interface system, aircraft demographic mix, communication technology, and automated tools available at the airport. The baseline technique is effectively current operations with current technology: paper strips are in use, there are no TaxiBots or eTaxis, DataLink is only available for preflight, and the aircraft size mix is based on current ratios and current aircraft in use (e.g. Airbus 320, Boeing 747). The interface with no automation is the same as the baseline technique, except instead of paper strips, a new interface is used and provides path suggestion, conflict detection and other information to the controller. Lastly, the interface with automation, i.e. integration of the management of the automated TaxiBots with the controller interface, uses the new interface and a mix of modern and futuristic fully automated aircraft and current aircraft are in operation at the airport. Table 2 (modified from [1]) summarizes these three taxiing techniques, including proportions of current, modern, and futuristic aircraft.

Table 2: Taxiing Technique Definitions.

Taxiing Technique	Definition
Baseline	• Paper strips and current ATC technology • Current A/C : Airbus and Boeing Light / Heavy / Medium (A320, A330 etc) • Current taxiing (no TaxiBots, no EGTS) • DataLink for PREFLIGHT
Interface with no automation	• New interface • Current aircraft • Current taxiing operations (no TaxiBots, no EGTS)

Interface with automation	• New interface
	• Mixed aircraft fleet
	  ○ Modern and Futuristic aircraft that are TaxiBots compliant
	  ○ Current aircraft that are TaxiBots compatible
	  ○ Current aircraft that are not TaxiBots compatible

2.3 Dependent Variables

The efficacy of the HMI platform will be measured across 17 dependent variables (i.e. key metrics of interest). For purposes of comparison, these variables will be collected in each of the three levels of human-machine interaction (baseline, HMI0, HMI1). There are four major categories of dependent variables, each supporting an aspect of the validation goals: operational, behavioral, subjective, and neurophysiological. Section 2.4 discusses the hypothesized change in trend value as a result of each of the HMI platforms.

2.3.1 Operational measures

Operational measures quantitatively evaluate the change in system performance due to the two HMI platforms. They provide insight on the efficacy of the tool. These types of variables are the most important within the dependent variable set and cannot be replaced or substituted. There are ten variables, many of which are correlated with each other.

2.3.1.1 Fuel consumption

The rising price of fuel and the impact on the environment are major driving factors for the design of automated taxiing techniques such as eTaxi or TaxiBot. The ATCO can also improve on fuel consumption by providing clear, coherent, and shortest path instructions to the pilot. This value is the summation of the fuel consumed by each aircraft during the scenario and is calculated once per scenario. It is obtained from the aircraft dynamic models provided by Airbus and calculated by multiplying the engine run time by the fuel consumption rate associated with each aircraft size (heavy, medium, large). The fuel consumption is influenced by the taxing technique: manual (no eTaxi or TaxiBot), TaxiBot/eTaxi with autopilot, eTaxi. Additionally, there is likely a correlation between the number of stops and restarts and the fuel consumed. An appropriate factor will be used to account for the improved efficiency associated with automated technology use.

This variable is correlated with noise and emissions. Operating aircraft engines are the primary source of noise and emissions pollution during ground operations. It can be reasonably assumed that these values are proportional to fuel consumption. The specific percentage of reduction, however, depends on the size of the aircraft. The HMI1 intends to reduce fuel consumption by 10%, per goals established through the SESAR program.

2.3.1.2 Average Taxiing time

Current literature states that the average taxi time around Roissy airport is 15-20 minutes, with instances of two hours reported. The average taxiing time is calculated by averaging the taxi times of all aircraft at the end of each scenario. A reduction in average taxiing time not only minimizes financial and environmental costs, but also improves overall commercial air transportation for passengers. This value is positively correlated with fuel consumption, and subsequently, noise levels, and missions. It is the only value that will be directly measured by the simulation.

The average taxiing time will also be evaluated as two individual components: taxi-in and taxi-out. The former refers to the taxiing time experienced by arrivals (from landing to taxiing into gate) and the latter refers to the taxiing time experienced by departures (from gate out to the moment of takeoff). This distinction allows us to evaluate the gains achievable by both eTaxi (arrivals and departures) and Taxibot (departures only).

2.3.1.3 Number of Bottlenecks/Collisions

A limited number of resources are available to the GND ATCO, such as taxiway usage and the number of TaxiBots. A measure of ground operation effectiveness is the number of bottlenecks due to an over-demand of airport resources. A bottleneck usually results in a reduction in speed and sudden stops. Those indicators, which can be measured directly, will be used to determine the number of bottlenecks. This value is positively correlated with average taxiing time and fuel consumption. Bottlenecks may be observed in real-time and determined post-hoc.

The stop-and-go profile recorded for bottlenecks will also serve as an indicator of collisions. Additionally, the source of each collision will be reviewed post-hoc. Literature notes three sources of runway incursion: pilot deviation, when the pilot deviates from appropriate and correct commands of the controller; operational error, where the controller offers erroneous commands to the pilot, who correctly follows them; and vehicle/pedestrian deviation, where another vehicle or pedestrian crosses into the path of an aircraft [3]. Pilot deviation will not be simulated in this experiment, but controllers will be told that the possibility exists, at the same rate as real-life operations. They will experience pilot deviation during the practice sessions, but this deviation will not occur during the actual experiment.

2.3.1.4 Throughput

A primary measure of airport efficiency is the number of ground movement of aircraft per hour [2]. Since this utility is focused on the responsibilities of the GND ATCO, it is assumed that an aircraft leaving the associated jurisdiction will successfully reach the final gate or runway for takeoff. This value is considered a count variable and will be normalized to the standardized value of ground movements per hour. There is a negative correlation between throughput and average taxiing time.

2.3.2 Behavioral measures

Behavioral measures are observable and measurable physical actions taken by the controller. These measures provide insight on the human-machine interaction and indicate any potential problems with the ergonomic design and data presentation of the MoTa interface. Additionally, these quantitative measures can support operational (collision, source of error), subjective (workload, situation awareness), and neurophysiological measures (workload, situation awareness).

2.3.2.1 Number of Actions (total, profile)

The efficacy of a new interface can be calculated by the number and types of operator actions. Project MoTa does not aim to reduce the overall number of actions, but to use this variable to determine how the interface supports or hinders overall ATCO workload. We are particularly interested in how the GND ATCO manages the flow of and the multiple types of information. Stylus interactions will be automatically time stamped and recorded on the WACOM tablet for each scenario, in addition to a summation of all actions. A human factors analysis performed post-hoc will catalogue the overall actions (e.g. data search, communication, rerouting). This profile will be aligned with the pre-programmed scenario events, including the aircraft trajectories and used to support analysis of the source of error in collisions or bottlenecks. An action will be counted as an observable micro-action where the controller interacts with either the machine or the pilot.

2.3.2.2 Deviation from Ideal

The overall goal of taxiing operations is to determine the shortest and safest trajectory for a given vehicle, with respect to the other ground movements and airport properties. An optimal trajectory can be determined for every given aircraft assuming the current and predicted future state of all other vehicles. The trajectory determined by the GND ATCOs will be compared in real-time against the optimal trajectory calculated by a path-planning algorithm. This deviation from ideal, in seconds, will be summed and reported at the end of each scenario. This path-planning algorithm currently exists with other projects at the Aeronautical Computer-Human Interaction Laboratory at ENAC and will be adapted for use in Project MoTa. If path planning is not available by the start of the baseline experiment, the ideal trajectories will be calculated post-hoc and compared to participant data.

2.3.2.3 Participant Behavior

The behavior of the participant provides insight regarding the interaction with the interface, scenario, and automation. In particular, this variable measures non-recorded behavior such as observations of participant difficulty with the interface ergonomics (e.g. holding the stylus upside down), reaction or lack thereof to unfolding scenario events (e.g. failing to notice an impending collision), and surprise or confirmation to automation behavior (e.g. a deviation of TaxiBot trajectory from the participant's expectations). Such behavior will be noted by the researcher during the experiment, with follow-up questions occurring in the debriefing session, if necessary. All sessions will be videotaped and saved for further review.

2.3.2.4 Clarity of instructions

The clarity of the instructions given to the pilot from the ATCO is a strong indicator of behavioral performance. Clear, concise, and straightforward directions demonstrate situation awareness and good operational performance. Ambiguous, lengthy, and inefficient directions are a mark of the opposite. We will be recording the duration, the number of directions, comparing against the optimal, and consulting with subject matter experts regarding the clarity of instructions.

2.3.3 Subjective measures

Subjective measures provide insight to the changes in operator psychology due to the scenario or the interface. These measures are non-observable and based on participants reporting real-time or post-hoc. They are primarily qualitative, but are collected using validated quantitative techniques.

2.3.3.1 Workload

Literature suggests that operator performance is highly correlated with high workload. While a minimization in workload is not necessarily the aim of a HMI, as attentional problems may result, the goal of Project MoTa is to support a suitable level for the GND ATCO. This measure can be measured in real-time or at the end of each run. We have opted for a composite, singular measure of workload for each scenario. While validated real-time workload measurement methods have been frequently used (e.g. Workload Assessment Keypad ratings, [4]), these methods are not compatible with the neurophysiological sensors. We have considered two different workload questionnaires for this study and have opted for the NASA Task Load Index (TLX, [5]), which calculates workload across six dimensions (mental, physical, temporal, performance, effort, frustration) on a 20-point range. A final weighted score is used for the overall task workload. Since workload is such a critical measure and is inherently tied in with scenario complexity, we will also evaluate the changes in workload across the six dimensions.

2.3.3.2 Situation Awareness (SA)

Situation awareness, as the term suggests, is defined as perception, comprehension, and projection of a scenario ([6]). The MoTa HMIs aim to provide sufficient situation awareness to the controller, which is known to be positively correlated with operational performance. This measure can be measured in real-time or at the end of each run. We have opted for a composite, singular measure of situation awareness for each scenario. While validated real-time situation awareness measurement methods have been frequently used (e.g. the Situation Awareness Global Assessment Technique (SAGAT, [6]), these methods are not compatible with the neurophysiological sensors. We have considered two different workload questionnaires for this study and have opted for the Situation Awareness Rating Technique (SART, [7]), which calculates situation awareness across 10 dimensions (instability, complexity, variability, excitement, concentration, division, spare mental capacity, information quantity and quality, familiarity) on a 7-point range. A final weighted score is used for the overall situation awareness.

2.3.3.3 Trust in Automation and Automation Surprise

An important aspect of the MoTa interfaces is the acceptance of the utility by the ATCO. This trust in automation is particularly important in the incorporation of the automated taxiing systems, such as path suggestion, eTaxi, and TaxiBot. This variable is measured at the very end of each experiment and serves as a composite evaluation of the automation across the scenarios. We will be using the SHAPE Automation Trust Index (SATI, [8]) to quantify the trust in automation. This questionnaire provides a mean score out of six points.

Additionally, participants will be asked to describe any instances of surprising automation behavior. This subjective qualitative measurement will allow for improvements in automation logic and representation on the HMI platform. This open-ended question is posed once at the end of each experiment and may be followed up with specific observations of possible automation surprise (cf. §2.3.2.3)

2.3.3.4 Technology Feedback

The human-centered development approach to the HMI platforms relies on input from subject matter experts, including the participants. At the end of each experimental session, participants will be asked to opine on the respective platform, including likes and dislikes. A Human Factors (HF) specialist will review all comments, including observed Participant Behavior, and determine appropriate functionalities that should be incorporated into the design. These new features will be presented in the following experiment. Participants may be asked to comment directly on specific features.

2.3.4 Neurophysiological measures

Neurophysiological measures are the objective complements to subjective measures such as workload and situation awareness. They allow from real-time monitoring of physiological and cerebral phenomena. These sensors are attached to the participant.

2.3.4.1 Cardiocirculatory

Heart Rate (HR) is measured using an electrocardiogram (ECG) which is a sensor attached to the skin of the participants in key areas such as ear lobes and the chest. This measure is generally strongly correlated with workload and is non-invasive. Kaber et al [9] showed that HR varied with respect to the level of automation and Vogt [10], in his detailed review of the literature, identified several studies that indicated an increase in HR with respect to workload. The sensors will be applied at the start of the experiment and a short calibration will be conducted, to determine the baseline HR. In general, we will be evaluating the change in HR across scenarios, automation usage, and during the progression of the events themselves.

2.3.4.2 Electrical cerebral activity

We will be using electroencephalography (EEG) to measure the electrical cerebral activity, to understand the cognitive workload. We will be using one electrode at node Cz, directly on top of the head. A three minute baseline will be taken at the start of the experiment for each participant and the changes in potential will be evaluated. Additionally, the changes with respect to specific events during the scenario will be noted and evaluated, for both the workload progression and the usage of specific key areas of the mind.

2.3.4.3 Oculometric

Oculometric measures of eye movement (such as blink rate, blink duration, and areas of interest) provide two measures of information. First, they act as a behavioral measurement in terms of a record of the ATCO's eye gaze. One can determine the efficacy of the information display (and interface maneuverability) through these measurements. For example, the frequency and duration of eye gaze on a certain aspect of the display general is positively correlated with usage. Second, oculometric measures have been proven through similar studies ([12]) to be an accurate indicator of workload. Ahlstrom and Friedman-Berg showed that as workload increased, blink duration decreased, saccade distance decreased, and pupil diameter increased. The same result was also determined in similar studies, such as ([13]).

The use of several sensors to capture such a key element is critical in ensuring that the HMI meets the project goal of reducing workload. Additionally, eye tracker measurements can be used to provide supporting evidence of situation awareness. For example, if the ATCO fails to perceive a vehicle that has entered his area of jurisdiction, or Area of Interest (AOI), it is an indicator that he lacks situation awareness at that moment. Despite the inability of neurophysiological measurements to provide conclusive and absolute evidence of situation awareness, they can suggest certain behaviors. We plan to measure this by outlining areas of interest within the workspace and within the tablet itself.

2.4 Data Analysis

Because the data for each of the taxiing techniques is collected separately, multiple data analyses will occur and answers about the different effects of the interface and taxiing techniques will be known mid-Project. This analysis approach allows us to modify the interface and the taxiing technologies as appropriate. There are four main analyses: 1) Understanding the impact of the scenario complexity levels; 2) Impact of the new user interface on current and forecasted operations; 3) Impact of the taxiing techniques (interface and automation utilities); 4) Impact of the taxiing utilities such as eTaxi, TaxiBot. Table 3 summarizes the overall dataset comparisons, the analysis, and when such an analysis can be completed.

Table 3: Planned Analyses

Dataset comparison	Analysis	When
(Baseline, average) v. (Baseline, hard)	Quantification of scenario complexity	After baseline
Baseline v. Interface, no auto	Impact of new user interface with automated tasks on current and possible operations	After HMI0
All data	Impact of taxiing techniques	After HMI1
Interface, no auto v. Interface, light auto v. Interface, full auto	Impact of TaxiBot and eTaxi	After HMI1

The effects of the HMI platform on the individual dependent variables can be determined for each of the dependent variables. It is assumed that two scenarios are distinguishable in their complexity as to induce significant changes in each of the dependent measures. Table 6 lists these hypotheses in greater detail, with the respective sections (i.e. Sections 0, 5.2, 0) postulating possible outcomes due to the effects of taxiing technology and scenario complexity. The data analysis will consist of several parts: data preparation, qualitative analysis, quantitative analysis.

2.4.1.1 Data Preparation

The data preparation is the post-hoc categorization of raw data. For example, the source of error in every collision cannot be determined in real-time and must be calculated after the test. The significance of the number of actions will not be clear until it is compared to observations made of participant behavior. Additionally, the neurophysiological data will likely require the application of some filters, to remove some of the noise. The data will need to be synchronized with a timeline of scenario events and what is being displayed on the interface. Table 4 lists the required preparation of the relevant variables.

Table 4: Data Preparation

Variable	Preparation
O: Number of bottlenecks	Counting the number of stop-and-go occurrences for each trial
O: Collisions	Evaluating each stop-and-go occurrence and determining the source
B: Number of actions	Timelines of number of screen interactions, along with events
B: Deviation from Ideal	Calculating time loss due to changes in route
B: Participant Behavior	Isolating interesting changes in behavior through video evaluation
S: Automation Surprise	Compilation of participant comments of surprising automation behavior
S: Technology feedback	Compilation of participant comments and grouping them into categories
N: Cardiocirculatory	Filtering the data; Evaluating the change in HR from baseline, separating into groups based on events within the scenario
N: EEG	Filtering the data; Evaluating the change in electrical activity in node CZ
N: Eyetracker	Accounting for head displacements; isolating specific events during the scenario; counting frequency of scans in Areas of Interest (global); counting whether specific events/alerts were scanned; average pupil diameter;

2.4.1.2 Qualitative Analysis

The qualitative analysis will naturally follow the data preparation, elaborating on the previous categorizations and reviewing effects such as automation surprise, participant behavior, and technology feedback. A list of these tasks is listed in Table 5.

Table 5: Qualitative Analysis

Variable	Preparation
O: Number of bottlenecks	Percentage of source of bottlenecks, suggestions for avoiding bottleneck (pie chart)
O: Collisions	Percentage of source of collisions, suggestions for preventing collision (pie chart)
B: Number of actions	Determining the correlation between screen interactions and events; Percentage of action type (pie chart, graph)
B: Clarity of instructions	Evaluation with expert controller to determine clarity of commands for each participant and provide and overall grade (graph)
S: Automation Surprise	Understanding the source of automation surprise and determining a resolution (bar chart, list)

S: Technology feedback	Converting participant remarks into cohesive design requirements (list)
N: Neurophysiological	Isolating specific events and describing, in physiological matters, the participant's reaction, providing supporting evidence for situation awareness (list)
N: EEG	Determining supporting evidence for behavior based on changes in electrical activity
Artifact analysis	Human-Artifact interaction and task analysis
Automation roles	Distribution of tasks between human and automation

2.4.1.3 Quantitative Analysis

The quantitative analysis will be completed with statistical tests of all relevant data. All tests will be run at $\alpha = 0.05$. Depending on the results, it may be necessary to iterate on the qualitative and quantitative analyses. In Table 6, HMI refers to both the interface and the automation. The same hypotheses also apply to the effect of just the interface.

Table 6: Dependent Variable Hypotheses

Type (O/B/S/N): Variable	Hypothesis	Test
O: Fuel consumption	HMI will reduce overall fuel consumption in all scenarios by 10%	Mixed Models ANOVA (IVs, Fuel, Time, all Numbers, Throughput, Deviations)
O: Average Taxi Time	HMI will reduce the average taxi time.	
O: Number of Bottlenecks/Collisions	HMI will reduce bottlenecks and Collisions	
O: Throughput	HMI will increase throughput	
B: Number of Actions	HMI will not cause the occurrence of spikes in the number of actions, which generally indicate confusion with the interface. On average it will reduce the duration of actions	
B: Number of Deviations from Ideal	HMI will reduce the number of deviations from the ideal trajectory	
B: Participant Behavior	HMI will be appealing to participants	Qualitative analysis: observations and participant feedback
S: Workload - mental demand	HMI will maintain or reduce mental demand from high to lower levels	Wilcoxon signed rank test (IVs, Workload)
S: Workload - physical demand	HMI will maintain or improve physical demand from high to lower levels	
S: Workload - temporal	HMI will maintain or improve temporal demand from high to lower	

demand	levels	
S: Workload - Performance	HMI will maintain or improve performance to 7	
S: Workload - Effort	HMI will maintain or improve effort from high to lower levels	
S: Workload - Frustration Level	HMI will maintain or improve performance to 7	
S: Situation Awareness	HMI will improve situation awareness	Wilcoxon signed rank test (IVs, Situation Awareness)
S: Trust in Automation	HMI will maintain or improve trust in automation to 6	Wilcoxon signed rank test (IVs, Trust in Automation)
S: Technology Feedback	HMI will not receive any major complaints regarding ergonomics, information presentation, tool functionality, only concerns regarding human-computer interaction	Qualitative analysis: observations and participant feedback
N: EEG	HMI will reduce the overall workload	Mixed Models ANOVA (IVs, change in electrical potential, average pupil diameter, average change in HR)
N: Eyetracker		
N: ECG		
Effect of participant demographics	There will be no differences due to participant demographics	ANCOVA (IVs, participant demographics)
Correlation between dependent variables	How strong is the correlation between certain variables?	Pearson's correlation coefficient, Spearman's correlation

2.5 Participants

The participant pool will primarily consist of students at ENAC enrolled in the ATCO Ingénieur du Contrôle Aerien (ICNA) program. These students will be recruited several weeks prior to the start of each experiment. No direct financial compensation will be provided, but their names will be entered in a pool to win prizes such as movie tickets. The expected individual time requirement for each participant will be about two hours, with the option to spend more time. Breaks will be provided to the participant as necessary. The student's year of study will be collected, his or her gender, age, and whether they participated in previous experiments with the Project. These covariates will be used to determine if there are any significant non-anticipated effects or required data blocking for the statistical analysis. We will be working with the instructors to determine the appropriate participant requirements, necessary training given the time limitations, and adaptations to the scenario and environment for accuracy representation of Roissy ATCOs.

Additionally, there is the possibility of recruiting active or retired ATCOs at Roissy to come to ENAC for a day of training. We would specifically recruit GND ATCOs. The years of experience, age, gender, and previous participation with the Project would be collected for covariate analysis, and for equivalency across different levels of the between subjects variable. These ATCOs would also participate in smaller sessions with the simulation, either to evaluate the efficacy of a specific element or to provide data samples for MoTa efficacy.

3 Simulation Design

3.1 Hardware, Software, and Pseudo-Pilot/Controller Design

All experiments will be conducted at ENAC in the Aeronautical Computer-Human Interaction Laboratory. The simulator consists of a desk, a projector system, a radio, DataLink interface, a paper strip printer, the DISCUS managers, and the HMI platform (Wacom tablet and pen). The participant will be seated at the desk with the platform (if in use for the test) placed in front of him, pen in his or her hand. The projector presents the view from the tower interior of the south end of Roissy airport.

The simulated view of the south end of Roissy is being developed by ENAC based on FlightGear, open-source software ([15]). We are simulating runways 26R and 26L in the West configuration, the aircraft (including size and designation), relevant vehicles (e.g. emergency services and tractors), and weather (precipitation, visibility). The viewing angle is 180 degrees (projected on screens covering a 120 degree viewing angle and a screen behind the participant) with the height of the tower included in the projected image.

The taxiing simulation is being developed by ENAC, Airbus, and EADS. The simulation includes the models of aircraft dynamics (Airbus) and the automation/management of the TaxiBots (EADS). The aircraft dynamics include changes in velocity and acceleration with respect to technology use, duration of pilot checks, and fuel consumption approximations. The simulation also includes communication via radio or DataLink between the participant (GND) and the pilots. These pilots will be simulated through the use of instructors at ENAC, who will provide the voice communication or DataLink responses. The simulation will redirect the aircraft in response to the participant's commands on the HMI interface. Appropriate communication transmission lags will be simulated for realism and to allow instructor/pilot confirmation.

There are five human roles for each experiment: the participant, the human factors specialist, the simulation manager, and two pilots. The human factors specialist (ISAE) is primarily concerned with the collection of the behavioral, subjective, and neurophysiological measures. She is in charge with the installation and calibration of the EEG, ECG, and eyetracker. The simulation manager (ENAC) is in charge of starting the simulation, ensuring that the hardware and software are functioning and communicating. The two pilots are managing the response of the aircraft and providing voice support. Two pilots are needed in order to handle the aircraft authorization frequency – the anticipated rate is an authorization request every less than a minute. The pilots will be used for the baseline experience. There is the possibility that the pilots may be substituted by other engineers or researchers, depending on the required functions.

The expected participants do not have training specific to Roissy. Based on advice from instructors of the ATCO program at ENAC, we will simplify the taxiing operations. A map of taxiway directions, guidelines on order of departure, and restricted areas based on aircraft size are provided to the participants during all the experimental sessions.

3.2 Scenario Design

Previous studies indicate that cognitive complexity is highly correlated with workload – more complex and difficult scenarios induce greater workload for the ATCO. Cummings and Tsonis (2006) describe three categorical sources within the ATC domain that contribute to cognitive complexity: environmental, operational, and display. Environment complexity consists of traffic demographics and density; operational complexity is defined as airspace properties and air traffic management principles; display complexity primarily stems from information organization. In this validation program, environmental and organizational complexities are modulated to produce scenarios of relevant levels of cognitive complexity to allow testing of various display complexities.

However, only a small set of variables must be modulated to induce sufficient differences in cognitive complexity. Traffic load is generally cited as the largest contributor to workload [2], however, guidance provided by the International Civil Aviation Organization (ICAO) lists visibility as another critical variable for surface movement systems ([16]). Feedback from two former Roissy ATCOs also notes taxiway availability as a source of complexity during ground operations management. The effects due to traffic load are readily apparent; every aircraft must be monitored by the ATCO and minimizes the available space for trajectory optimization. The effects of visibility on ATCO performance are less direct. Pilots must taxi around the airport at slower velocities during conditions of low visibility, possibly allowing the GND ATCO to react faster to the unfolding scenario, but also increasing the potential for bottlenecks and competition for limited taxing resources. Lastly, the closure of certain taxiways increases the possibility of bottlenecks and forces the ATCO to select non-optimal taxiing routes. We have chosen to modulate these three variables to produce two different degrees of complexity for our validation program. We will not be evaluating the effect of each of these three variables, but rather the effect of two representations of scenario difficulty.

3.3 Experiment Scenario Definitions

Two experimental scenarios were designed based on the following principles:

- Represent the range of nominal environmental and operational conditions that occur at Roissy
- Represent a possible distribution controller workload at Roissy
- Represent the range of cognitive complexity experienced by ATCOs

The descriptors (Average, Hard) are related to cognitive complexity. Cohesively, these scenarios span the majority of all possible working conditions at Roissy. Table 7 summarizes the values of the five scenario variables with respect to each scenario. These scenario variables were selected based on consultation with ATCOs. Visibility is kept constant as a clear day, as consultation with the ATCOs revealed that there is little effect on the workload of the GND ATCO.

Table 7: Modulation of Scenario Variables

Scenario	Traffic Load (mvmts/hr)	Configuration Change	Events
Average	~60	No	Towed aircraft (12'), pilot error (10'), forbidden zone (3'), blocked taxiway (16')
Hard	~90	Yes, T+15min, 10 min duration	Towed aircraft (3'), pilot error (6'), forbidden zone (20'), blocked taxiway (7')

Otherwise, all scenarios maintain the same aircraft properties. Only the South end is modeled, with the GND jurisdiction covering all except the two runways (and taxiways immediately next to them) and the parking stands. This configuration also corresponds to off-peak hours. At maximum capacity, this area is usually managed by two controllers, to reduce the workload. Roissy is assumed to be in West configuration by default, due to the appearance of a "hot spot" within the GND ATCO jurisdiction. A hot spot is defined as a region that has historically had the most number of collisions. Roissy has seven hot spots within the South end, but only one hot spot is relevant to the GND ATCO. The Vigie Trafic controllers, who manage the parking stands, are assumed to be active (between the hours of 06:30 and 23:45; [17]).

An equal mix of arrivals and departures will be modeled, with both types occurring at a randomized rate. Each experimental scenario will last for 30 minutes to allow for situation awareness. This duration was selected based on the advice of subject matter experts and is considered sufficient for capturing the effects of mental fatigue.

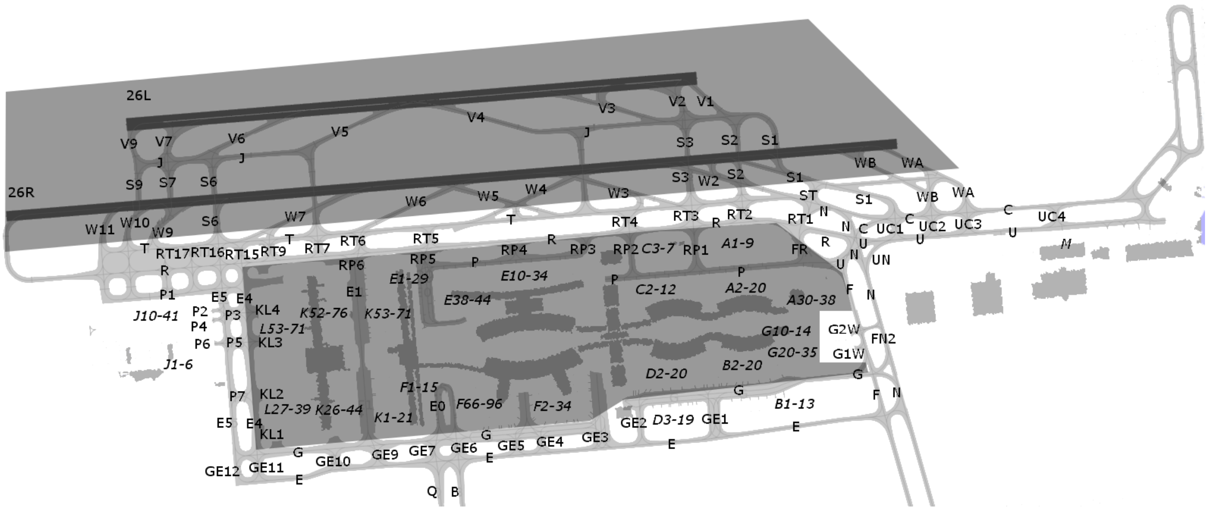

Figure 1: Roissy South, West Configuration. The non-gray areas are within the GND ATCO's responsibility.

3.3.1 Average Scenario

The average scenario represents the average workload experienced by the ATCO. None of the taxiways are closed and there are no winds necessitating a runway orientation change. It is the most common type of scenario experienced by controllers.

An aircraft request is received by the Tower approximately every minute, at a rate of 60 movements/hr (21 arrivals, 14 departures). We will be including four events that are representative of standard operations at Roissy, but are deviations of the standard taxiing operations. There is an aircraft that is towed from the North end of the airport towards Parking area J and moves at a pace of 15 km/hr, causing potential bottlenecks on taxiway E. A pilot error on taxiway F results in a head-to-head potential collision with another flight. An A380 is presented with the option of being sent through a forbidden zone (taxiway E) as it is the shortest route. A departure with a mechanical problem blocks taxiway RP15 for five minutes.

3.3.2 Hard Scenario

The hard scenario represents the most difficult type of scenario experienced by the ATCO. The traffic load is at its maximum. Normally, there would be two controllers handling managing the South end, however, we are interested in seeing if it is possible for a single controller to manage both jurisdictions. The scenario begins in the West configuration (26R, 26L). Participants receive an oral notification 5 minutes after the start of the scenario regarding a configuration change caused by significant winds. Fifteen minutes into the scenario, the arrivals are stopped as the configuration change occurs. For the next ten minutes, there are aircraft still finishing their taxing routes in the west configuration as departures are taxiing and preparing for takeoff in the east configuration. The last 10 minutes completed with the new configuration.

An aircraft request is received by the Tower about every 30 seconds, at a rate of 90 movements/h (23 arrivals, 33 departures. Since each test is thirty minutes long, the ATCO will see 45 aircraft total. As with the average scenario, there are several events built into the scenario. A towed aircraft is traversing from the North end towards Parking M, causing bottlenecks on taxiway F. A departure has a mechanical problem blocking taxiway RP16 for 7 minutes. An arrival from the north end makes an error and continues on taxiway F, instead of taking a left onto taxiway E. An A380 is presented with the option of being sent through a forbidden zone (taxiway E) as it is the shortest route.

3.3.3 Potential Conflicts

Both scenarios offer several possible conflicts that the ATCO must manage, based on the effects of the different complexity variables, the distribution of aircraft sizes, and resource utilization. We have identified 168 possible trajectories that reflect approximate standard 'shortest-path' routes from taxiways connected to the runway to taxiways leading to the different terminals (representing the boundaries of the GND jurisdiction area). The timeline of events and specific trajectories will be finalized later, but several sources of conflict have been identified.

There are six conflict categories. The frequency of these possible conflicts is modulated by the aircraft trajectory length, which is determined by its starting and end point. In each of these scenarios, the ATCO may mitigate potential conflicts by predicting future aircraft states and resource availability, or by intervening in real-time.

- **Configuration change**: participant must be aware of the changes in taxiing direction. He must manage the remaining aircraft that are still using the previous configuration and avoid collisions with departures in the new configuration. Hard scenario only.
- **Aircraft Size Separation**: Although the GND does not directly manage the order of takeoffs, he attempts to minimize the work of the LOC by respecting the required separation between Heavy and Medium aircraft. There are no Light aircraft at Roissy. Common trajectories include Heavy and Medium aircraft requiring the same taxiway at the same time. Both scenarios.
- **Blocked taxiways.** Mechanical difficulties may stop the taxiing process of an aircraft. Subsequently, a portion of the airport may not be usable for some time. The GND must be aware of this blockage and come up with alternate routes, even going against the direction of traffic on each taxiway. Both scenarios.
- **Towed aircraft**: These aircraft are moving much slower. They may cause bottlenecks on main taxiways such as R, T, F, and N. Both scenarios.
- **Restricted areas**: Certain aircraft are forbidden to access specific taxiways due to size or weight restrictions (e.g., FN2 is closed to aircraft heavier than 80t [17], and FN2 allows for easy access to and from Gate areas G2W and G1W). Both scenarios.
- **Number of aircraft**: As the number of elements grows and the specificity of each element increases, the workload of the controller is elevated. The visual distraction of such aircraft may pose difficulties to the controller in keeping track of all pieces, particularly when the GND must check up on the progress of the aircraft (e.g., transferring the aircraft to the North, LOC, or Vigie controller).

3.4 Qualitative Questions

Situation awareness, workload, and trust in automation (for the second and third experiment only) rubrics will administered to the participants. Since all participants will be native French speakers, the questions will be translated in French.

Situation Awareness Rating Technique (SART, answers on a 7-point scale):

- How changeable is the situation? Is the situation highly unstable and likely to change suddenly (high), or is it very stable and straightforward (low)?
- How complicated is the situation? Is it complex with many interrelated components (high) or is it simple and straightforward (low)?
- How many variables are changing in the situation? Is there a large number of factors varying (high) or are there very few variables changing (low)?
- How excited are you in the situation? Are you alert and ready for activity (high) or do you have a low degree of alertness (low)?
- How much are you concentrating on the situation? Are you bringing all of your thoughts to bear (high) or is your attention elsewhere (low)?
- How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (high) or focused on only one (low)?
- How much mental capacity do you have to spare in the situation? Do you have enough to attend to many variables (high) or nothing to spare at all (low)?
- How much information have you gained about the situation? Have you received and understood a great deal of knowledge (high) or very little (low)?
- How good is the information you have gained about the situation? Is the knowledge communicated very useful (high) or is it a new situation (low)?
- How familiar are you with the situation? Do you have a great deal of relevant experience (high) or is it a new situation (low)?

NASA Task Load Index (NASA-TLX, answers on a 7-point scale):

- How mentally demanding was the task?
- How physically demanding was the task?
- How hurried or rushed was the pace of the task?
- How hurried or rushed was the pace of the task?
- How successful were you in accomplishing what you were asked to do?
- How hard did you have to work to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

SHAPE Automation Trust Index (SATI, answers on a 7-point scale):

In the previous working period(s), I felt that

- … the system was useful.
- … the system was reliable.
- … the system worked accurately.
- … the system was understandable.
- … the system worked robustly (in difficult situations, with invalid inputs, etc.).
- … I was confident when working with the system.

Additional questions will include (especially as a follow up to SATI):

- Were there any instances when the automation did not perform as you expected?
- What aspects did you like/not like about the technology?
  - This question will be followed up with questions on specific elements.

3.5 Participant Training Material

The participants are presented with training materials to help situate them to the task and the regulations concerning Roissy prior to the start of each experimental session. There are several components to the training material:

- A map of Roissy airport (East, West orientations)
  - Overview of GND jurisdiction
  - Identification of forbidden taxiways
  - Principle exit and entry taxiways into each parking zone
  - Identification of parking areas in the North end
- Rules of ground operations at Roissy and a list of aircraft
  - Which parking areas must be managed by the GND ATCO
  - The directions on the taxiways
  - The aircraft fleet, also sorted by weight category
- Common aircraft and their call signs
  - Indicators and aircraft company name
- Example of strips and utilization
- A copy of the questions that will be asked after the experiment, regarding workload, situation awareness, and confidence with the technology)

4 Baseline Experiment

The baseline experiment is planned for August/September 2014. Prior to this experiment, there will be an integration session completed with a small number of subjects. The purpose of this experiment is to collect data regarding the current practice of ground control operations. There will not be any technological upgrades – strictly the strip tableau and DISCUS, the flight manager. The purpose of this experiment is to collect baseline data that will be compared to the next two studies with automated aids. Also, this baseline experiment will allow participants to provide feedback on the shortcomings and advantages of the current system. These suggestions will be integrated into the design of the interface and multi-agent system.

4.1 Procedure

This experiment is expected to last approximately two hours. Scenario difficulty is the only independent variable tested in concept validation of baseline experiment. The presentation of the two scenario levels will be randomized and counterbalanced. Operational, behavioral, and neurophysiological measures will be recorded during each run. Subjective measures will be taken at the end of each run. Table 8 presents a timeline of events.

Table 8: Experiment Schedule

Time	**Duration**	**Activity**
0:00	0:20	Practice with the equipment and different scenarios (all scenarios seen)
0:20	0:20	Eyetracker, EEG, ECG installation calibration
0:40	0:30	Run #1: Testing session
1:10	0:10	Run #1: Subjective Measures, eyetracker break/recalibration, questions
1:20	0:30	Run #2: Testing session
1:50	0:10	Run #2: Subjective Measures, remove all neurophysiological equipment, open-ended remarks by participants
2:00		**End of the experiment**

The practice session is intended to provide participants familiarity with Roissy airport, but many of the participants will have had some experience with the Roissy ground plan. Each participant will receive access to the training materials prior to the start of the experiment. These materials include a task explanation, a small manual of the simplified rules and regulations of taxiing around Roissy airport, and a copy of all the subjective tests (TLX, SART, SATI).

4.2 Statistical Analysis

The planned statistical analysis for the Baseline study will follow the generalized plan listed in Table 6, Table 5, and Table 6. However, there are a few significant changes. For example, the measure of automation surprise is not relevant to the baseline study as there is no automation tested. The AOIs associated with the eyetracker are different, as no tablet is included as part of the equipment (Figure 2: Areas of Interest for the Eyetracker). Since only one automation level is in use, a one way ANOVA can be used in lieu of a mixed models ANOVA (assuming all assumptions hold).

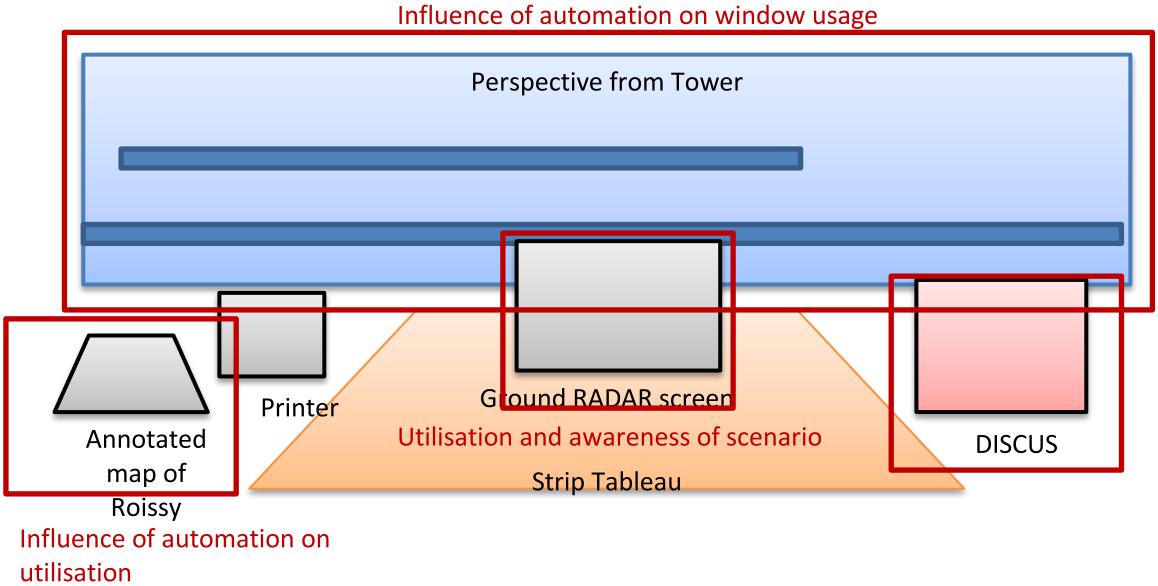

Figure 2: Areas of Interest for the Eyetracker

5 Evaluating the HMI0 platform

The Concept Validation of HMI0 is planned for between October and December 2014. Prior to this experiment, there will be an integration and pilot session completed with a small number of subjects who are not part of the participant pool, but familiar with the project and the taxiing task. The purpose of this experiment is to collect data based on the performance of the ATCOs using the new interface, but without any automated taxiing aids. The collected data will be compared to the baseline data to determine the impact of the interface. Later, this dataset will be used to quantify the impact of automated taxiing technologies such as eTaxi and TaxiBot.

5.1 Procedure

The same procedure as the baseline scenario will be used in the concept validation of HMI0 (Table 8). As with the previous experiment, a training packet will be sent to participants prior to the start of the experiment. In addition to those materials, we will include an explanation and demonstration of the MoTa platform.

5.2 Statistical Analysis

The planned statistical analysis for the Concept Validation HMI0 study will follow the generalized plan listed in Table 6, Table 5, and Table 6. However, there are a few changes. The AOIs associated with the eyetracker are different, as the tablet will be part of the equipment (Figure 3). Overall, the main goal is to show that the taxiing technology (baseline, interface) is significant. Significant interaction effects between taxiing technology and scenario difficulty, especially in the desired direction (harder scenario and more technology results in better performance) would also be useful in demonstrating the ability of the platform to support more complex scenarios. A pairwise contrast between the two levels of technology is planned.

The modification of the task from a paper-strip, near omni-directional information arrangement to a digitalized, multi-layered primary information display may cause some interesting results. First, ATCOs will now be tasked with managing the information presented. There may be some perceived incompatibilities with different information layers, resulting in information cluttering. This act of determining which pieces of information are critical and arranging them appropriately effectively acts as a secondary task. It may cause additional mental workload or minor losses in situation awareness. However, the expected gains in using such an interface outweigh these temporary shortcomings. Technology feedback should also help in correcting the interface if this situation should occur.

The interface is also designed for supervisory control, rather than replacing the original task. For example, most aircraft use the same standard taxiing routes, which are well known to ATCOs. The interface offers this standard trajectory, but also allows the controller to modify as necessary (to account for bottlenecks, taxiway closures, and other unplanned events. The trajectory trace is also visualization, or a means of remembering the aircraft position. We expect to see an improvement in the speed of giving taxiing trajectories and maintaining situation awareness of the unfolding events. For example, the participant should be able to highlight specific regions on the map and identify a closed taxiway. This visual aid allows monitoring of a critical area with a higher potential of collisions. Additionally, the interface could potentially provide a sense of feasible routes to the controller, if a specific aircraft is highlighted. This mechanism would aide during the configuration change, as the GND ATCO would be able to keep track of the direction and feasible routes for the aircraft.

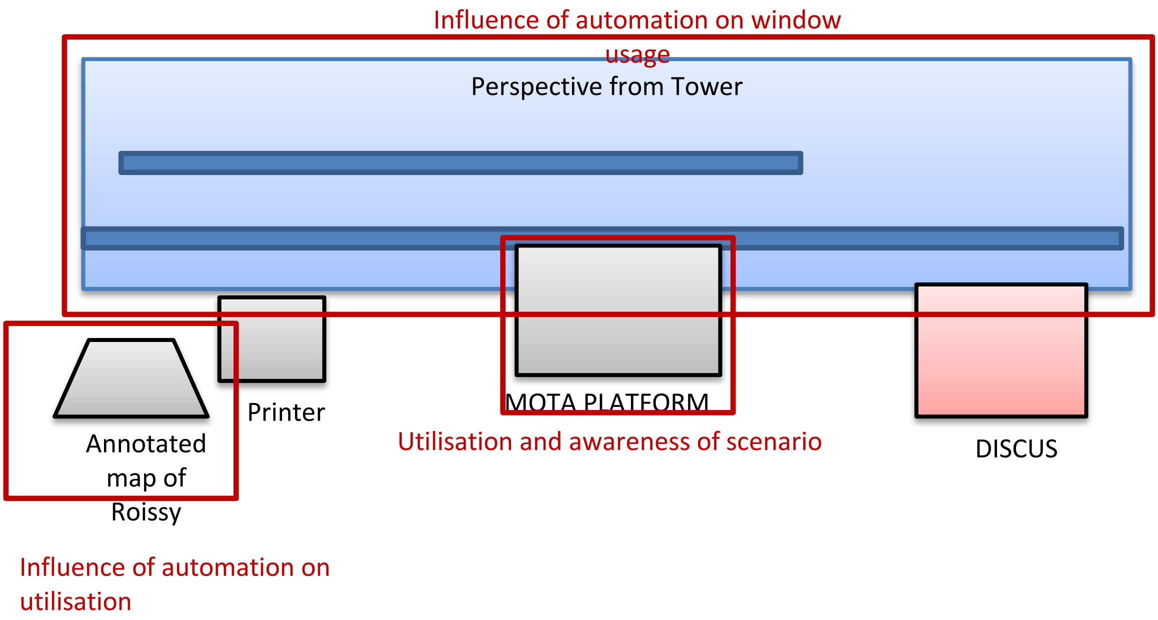

Figure 3: Areas of Interest for Eyetracker

6 Evaluating the HMI1 platform

The Concept Validation of HMI1 is planned for October 2015 and is the last formal experiment conducted during the Project schedule. Prior to this experiment, there will be an integration and pilot session completed with a small number of subjects who are not part of the participant pool, but familiar with the project and the taxiing task. The purpose of this experiment is to collect data based on the performance of the ATCOs using the new interface and automated taxiing aids, and to quantify the impact of the automated taxiing technologies themselves.

6.1 Procedure

The same procedure as the baseline scenario will be used in the concept validation of HMI1 (Table 8). As with the previous experiment, a training packet will be sent to participants prior to the start of the experiment. In addition to those materials, we will include an explanation and demonstration of the MoTa platform, and present the usage of the automation technologies. A full list of symbols (such as those denoting aircraft with eTaxi capability) will be provided to the participants.

6.2 Statistical Analysis

The planned statistical analysis for the Concept Validation HMI1 study will follow the generalized plan listed in Table 6, Table 5, and Table 6. However, there are a few changes. The AOIs associated with the eyetracker are different, as the tablet will be part of the equipment (Figure 3). Since the TaxiBot usage will be of particular interest, we will likely identify those stations as areas of interest.

The introduction of additional taxiing technology is likely to cause interesting results. The use of the interface may result in the same problems discussed in Section 5.2, namely, the potential secondary task of determining the information layout. A tertiary task is also possible, if the ATCO must directly monitor the health and availability of the TaxiBots. Otherwise, the TaxiBots acts as any other standard tug or tractor and should not create any additional effects beyond what is normally experienced.

The use of eTaxi/autopilot and TaxiBot/automated drivers is expected to create constant velocity and regular driving patterns during taxiing. This effect should reduce workload by eliminating the need to monitor pilot conformance. However, it may introduce an over-trust in the automation and the operator may make erroneous assumptions regarding the future states of the taxiing task. The introduction of the HMI algorithm will also be of interest, especially in the division of tasks. We will measure for participant perception and comprehension of alerts (oral, visual) that come from the HMI system. Eye scans in the AOI around a visual alert are generally an indicator of alert recognition, just as the inverse can indicate attentional tunneling. The MAS also can provide path suggestion based on the inputs from the participant. A closed taxiway can be considered in the algorithm, and thus, the shortest route can be determined based on this new information. Additionally, at higher counts of active aircraft where some may be forgotten, the MAS can provide alerts and reminders of time critical events.

7 Shorter Studies

Between preparations for each of the three studies, we plan to conduct smaller, shorter studies that do not require dedicated experimental campaigns. Many of these shorter studies will be planned around the availability of specialized personnel (e.g. availability of students who have just received training at Roissy; retired or active air traffic controllers who are at ENAC's campus). A portion of the experiment plan or an element of the HMI will be evaluated, with a larger focus on qualitative feedback. A subset of the measurements will be taken, based on practicality and constraints of equipment.

These shorter sessions are similar to the concept test session conducted in mid-February 2014, where an initial prototype of the interface was presented to two retired Roissy controllers. Their suggestion to focus on allowing trajectory modifications rather than generation resulted in a direct change in the functionality of the interface. There are several topics of interest to the team, but not all of these topics may be studied.

- Maximum traffic capacity: Increase frequency of workload beyond that planned for the Hard difficulty and determine the extreme limit.
- Multi-Agent System functionality: Evaluate the usability and acceptability of specific automated functions, such as the automation system automatically managing aircraft if a workload threshold is met (e.g. frequency of aircraft).
- Information management: Examining the types and arrangement of information layers, including number of layers and visualization.
- Role of TaxibBot: This technology is still new and there is no standard practice set for the taxibots. Subsequently, there are open ended questions involving the ATCO's role (does the ATCO need to manage the health of the Taxibot? Does the pilot or the ATCO request usage of the Taxibot?). We would like to solicit opinions and evaluate performance with different Taxibot operations.
- Low workload: As with most operations, the possibility exists that low workload would cause performance degradation due to the lack of stimuli to engage the ATCO. We are particularly interested to note the neurophysiological responses to low workload. Such information could be used as feedback to the automation system, which can then employ countermeasures to maintain appropriate performance.

8 Appendix: Baseline Data Collection

Baseline data is necessary in order to establish a comparison point for validation of both the interface and the taxing techniques. Three possible methods for obtaining this data were determined: using airport logs, visiting Roissy airport, and running a third experiment. Airports routinely track and keep records of aircraft arriving, departing, and traversing, for legal reasons and for improving aircraft operations. These tracks can be used to determine baseline values for operational measures such as overall fuel consumption. However, they can be difficult to access and may be inefficient in terms of achieving the desired measurements. A visit to Roissy builds on the calculation of the logs by adding observation and debriefing sessions, allowing for collection of subjective and behavioral measures. However, they have the same problems as with using the airport logs. Also, there are other possible hindrances to obtaining access to the controllers for debriefing purposes. Lastly, a third human-in-the-loop experiment offers the most amount of control over the specific taxiing scenarios of interest with respect to the other options, but also requires a substantial amount of development and execution time. It requires access to other tools (e.g. a paper strip printer) and human resources (participants, faux-pilots, specialists). Based on the answers to four key questions, we have opted for a baseline experiment that will be conducted between June and October 2014.

8.1 Using Airport Logs

Using only the airport logs to calculate baseline values for the dependent variable set requires access to the raw data and manipulating the data as appropriate. There are three tasks to this analysis: 1) Obtaining the relevant airport logs; 2) separation of the data based on the scenario variables; and 3) calculating the average values for the dependent variables of interest.

8.1.1 Obtaining the relevant airport logs

Airports routinely monitor and track ground movements, including identification of the relevant aircraft and its arrival and destination. However, this information is not publically available. We will request information such as the raw radar data (marking the position of each aircraft each second and its respective identification). Additionally, since aircraft weather and visibility is a defining factor in the experimental scenarios, it will be necessary to look at the recorded weather. Online database such as the Automated Weather Observing System (AWOS) have this information publically available for international airports.

We will request as much data as possible, as to find multiple data points related to the specific scenarios of interest.

There are several potential risks associated with this step:

- May not be possible to obtain data: Obtaining the data may not be possible due to organizational policies at Roissy.
- May not be possible to match data: If multiple sources of data must be used, it may not be possible to align this information. For example, the visibility information may only be listed for the entire day, but this value may not reflect the true weather.
- May not have data or be difficult to determine data for other ground vehicles. The movement of other vehicles such as tractors and emergency services may not have been **recorded**.

8.1.2 Separation of the data based on the scenario variables

The logs must have the following data, with the scenario variables underlined:

- **Operations**: communication (GND, LOC, Vigie, Pilot), occurrence of a controller switch, expertise of the controller, ground radar accuracy, usage of DataLink, time of data
- **Environment**: <u>visibility</u> (within the smallest time increment as possible), weather (<u>precipitation</u>), wind direction and speed
- **Airport**: runway operations (full/partial closure notification and duration, <u>orientation</u>, purpose of closure), <u>taxiway operations</u> (full/partial closure notification, duration, orientation, purpose

of closure), time log of vehicles (position with respect to taxiway, runway crossing clearance, type, arrival/depart, pushback/parking)

It is expected that traffic density, will not be included in these logs. Thus, this value will need to be calculated in order to find relevant scenarios. The data will be calculated into the two scenarios designed for the validation study. For example, data recorded during clear visibility, a traffic load of a certain rate of movements/hr and all taxiways available corresponds to the Average scenario and would be saved for additional processing.

There are several potential risks associated with this step:

- The logs do not include the variables that we are using to define the scenario (e.g., runway closure state and duration is not included). We cannot evaluate the overall performance based on this particular variable.
- May be difficult to find scenarios of the same duration as those within the experiment: It is quite possible that all three scenario levels exist, but the specific circumstances do not persist long enough to match the anticipated thirty minute tests.
- May be difficult to find matching scenarios: This risk is particularly relevant if a forecasted traffic load is simulated. It will not be possible to find a scenario of a traffic load that does not currently happen.
- May be very time consuming to go through entire dataset and attempt to define in accordance to our chosen scenarios. We would build processing scripts to read and convert the raw data, but it may not be possible to automate all aspects. The manual labor involved with data processing may exceed the project budget.
- May not be possible to obtain exact aircraft trajectories: It may be worthwhile to use the same aircraft trajectories in all taxiing technology tests, particularly if we are simulating potential collisions. However, such trajectories may not be available through the aircraft logs.

8.1.3 Calculating the average values for the dependent variables

Within each scenario block of log data, the following calculations will be taken:

- Fuel consumption: requires trajectory of each aircraft; aircraft type. Assuming an average fuel rate, one can approximate the fuel consumed for each aircraft. Additionally, we will include factors such as number of stops and restarts. A summation of all aircraft provides the fuel consumed. Values for noise and emissions will be approximated using other rates appropriate for noise and emissions.
- Average taxi time: requires time stamp of Pilot-Tower communications. For each aircraft, calculate the time elapsed between pushback request and LOC transfer.
- Bottlenecks: requires time stamp of Pilot-Tower communications, usage of the tugs or tractors; usage of the runways; utilization of the taxiways. Calculate the demand for all of the resources and determine if there is an instance where the demand is greater than the re-source.
- Operational error: requires time-stamp of Pilot-Tower communications, trajectory of air-craft, runways usage. According to [3]: 'An operational error is an action of an air traffic controller that results in: less than the required minimum separation between two or more aircraft, or between an aircraft and obstacles (e.g. vehicles, equipment, personnel on runways); an aircraft landing or departing on a runway closed to aircraft
- Throughput: requires arrival/departures list, pilot authorization requests. Difference between authorization requests and arrivals/departures
- Deviation from ideal: requires list of departures; runway usage; aircraft type; trajectory of aircraft. Same calculations are made as departure spacing efficiency, but instead for arrivals.

There are two potential risks related to this step:

- Can only comment on operational performance, does not include any information regarding subjective, behavioral, or neurophysiological measures.

- Much of this data is not easily calculated and must be repeated several times depending on the number of scenarios available, may be very time consuming.

8.2 Visiting Roissy Airport

A visit to Roissy airport consists of observing South end tower operations and meeting with ATCOs at a separate period of time. There are effectively three parts to this analysis: 1) identification of corresponding scenarios; 2) taking measurements of the controllers; and 3) obtaining the associated ground movement logs. It is not feasible or possible to take neurophysiological measures in real-time, but the other three categories of data can be measured.

8.2.1 Identification of corresponding scenarios

This step is done in conjunction with scenario design. The visit must occur during periods when the desired scenarios are most likely to occur. For example, if one of the scenarios involves a deicing event, it will be necessary to visit Roissy during the winter. It is likely that a full analysis of observed and desired scenario correspondence will occur after the entire visit, as measures such as traffic load cannot be calculated in real-time. Time start and end will be recorded during the visit, along with notes of any deviation in events.

There are several potential risks associated with this step:

- May be difficult or impossible to match real-life scenarios with the planned experiment scenarios: For example, it would be impossible to observe predicted traffic loads, as it exceeds the current maximum ground movements. Additionally, the project timeline may not allow for certain observations. Ideally, the observations should occur prior to the first concept validation (October 2014) as to avoid potential researcher bias and to permit incorporation of controller feedback and interface design ideas. This deadline eliminates the ability to observe effects due to winter.
- May be difficult to find scenarios of the same duration as those within the experiment: We may be able to find matching scenarios (e.g. traffic load, visibility, and taxiway availability), but the duration is less than the experimental duration. Extrapolation of this observed subset is possible, but effects such as mental fatigue would be lost.
- May be difficult to have the same controller during the designed duration: Working conditions, ideally, do not change due to ATCO personnel shifts. Subsequently, it may be difficult to find a matching scenario of the same duration with the same controller. Therefore, the subjective measures would be misleading.
- May not be possible to videotape the working session: This potential risk is not particularly egregious with respect to experimental validity, but it does increase the workload associated for the researchers and the possibility of missing an event.

8.2.2 Measurements of the Controllers

Only a subset of the dependent variable set can be collected during the observations, in real-time: number of bottlenecks, clarity of instructions, participant behavior, and automation surprise. These variables require access, visual and auditory, Visibility of the radar map (number of bottlenecks, participant behavior, automation surprise); Pilot-Tower communication, generally a headset plugin (clarity of instructions); and offline controller communication (automation surprise). The researchers will note observations as they occur. Video and audio will be recorded if possible, for later review.

Video and audio capture permit measurement of the number of actions taken by the controller. It is unrealistic for the researchers to count and categorize actions in real-time. Therefore, this measure will be noted after the observations and likely after the visit.

Part of the visit to Roissy includes a debriefing session with the controllers. This debriefing session may occur immediately, one-on-one, after the observed scenario with the relevant controller, or be held with multiple controllers in the same session. The focus of the debriefing session will be on the subjective measurements: workload, situation awareness, trust in automation and automation surprise, clarity of instructions, participant behavior, and technology feedback. Questionnaires for each of these measures will be distributed. Automation surprise, clarity of instructions, and participant behavior will be closely linked to researcher notes from the observations.

There are a few potential risks associated with this step:

- May not be able to administer the questionnaires right after the runs: Administratively, it may not be possible, realistic, or ideal to collect data immediately after the scenario. This gap in time may results in memory decay or confounding effects. The controller may not remember the specific instance in question or confuse the event with another.
- May not be able to locate the controller later that day: The controller may not be available or circumstances may arise when the participant can no longer be involved with the experiment process. As such, the data is lost for that particular participant.

8.2.3 Obtaining the Associated Ground Movement Logs

The same analysis as described in §4.1 Using Airport Logs will be used in this step. The same associated risks are relevant in this situation and an additional risk, specific to this visit:

- Inability to collect ground movement logs within a reasonable period of time: It is not clear when the logs are accessible and may hinder data analysis.

8.3 Human-Subject Experiment

The purpose of a baseline experiment is to collect data based on the performance of ATCOs using current taxiing technology. The collected data will be used as a reference point for both determining shortcomings of current taxiing technology and for quantifying the impact of the HMI0 and HMI1 platforms. A baseline experiment also provides the team and opportunity to validate and test the desired three levels of scenario complexity envisioned for the validation program.

There are three steps to this option: 1) adapting the simulator and validating that the system works properly; 2) conducting the human-subject experiment; and 3) analyzing the collected data.

8.3.1 Adapting the Simulator

This step is led by ENAC and consists of determining which elements are required for a realistic simulation of the ATC tower, finding all of the necessary automated and human components, and integrating the system as a whole and ensuring that it is accurate. Section 3.1 discusses this experiment set up in greater detail. The only change for the baseline experiment is the types of hardware devices (no HMI platform, taxiing technology, only paper strips and existing tools such as the flight list). However, a software/hardware integration check must be conducted to ensure that all is working properly.

8.3.2 Conducting the human-subject experiment

This experiment is expected to last approximately 2.5 hours. The two experimental scenarios will be tested. Scenario difficulty is the only independent variable tested in the baseline experiment. The presentation of the four levels will be randomized and counterbalanced. Operational, behavioral, and neurophysiological measures will be recorded during each run. Subjective measures will be taken at the end of each run. A maximum of two participants will be tested each day. A maximum of 40 participants can be tested in one month, although the testing rate will likely be much lower than this maximum capacity.

Table 9: Baseline Experiment Procedure Schedule

Time	Duration	Activity
0:00	0:20	Practice with the equipment and different scenarios (all scenarios seen)
0:20	0:20	Eyetracker, EEG, ECG installation calibration

0:40	0:30	Run #1: Testing session
1:10	0:10	Run #1: Subjective Measures, eyetracker break/recalibration, questions
1:20	0:30	Run #2: Testing session
1:50	0:10	Run #2: Subjective Measures, remove all neurophysiological equipment, open-ended remarks by participants
2:00		**End of the experiment**

8.3.3 Analyzing the Data

The planned statistical analysis for this experiment will follow the generalized plan listed in Table 6, with a few minor differences. First, since only one automation level is in use, a one way ANOVA can be used in lieu of a mixed models ANOVA (assuming all assumptions hold). Second, all hypotheses assume that the scenarios levels have a significant impact on the dependent variables. We are planning six pairwise contrasts between each of the scenario difficulty levels. We will conduct post-hoc contrasts (using Scheffé's method) for any non-pairwise contrasts following the initial data analysis.

At the end of the test, we should have well defined descriptions of the dependent measures with respect to each of the scenario levels. Additionally, we will be able to determine three scenarios with suitable differences in complexity.

8.4 Comparative Analysis

Each of the three baseline data collection methods offers advantages and disadvantages. There are four main categories of interest: administrative, scenario, debriefing, and data. Administrative concerns time and financial costs. Scenario relates to degree of fidelity and resource usage. Debriefing is primarily concerned with accuracy in data collection and access to participants. Data is a summary of which of the four categories of dependent variables can be collected. Table 10 summarizes these advantages and disadvantages. Green smiling faces ("**:)**") are advantages; red frowning faces ("**:-(**") are disadvantages; and question marks ("**?**") indicate unknown information.

Table 10: Advantages and Disadvantages of the Three Baseline Data Collection Methods.

		Roissy	Logs	Exp
Administrative	**No additional work for MoTa** Required development work to get Simulator ready in style of Current Operations	:)	:)	:'-(
	No additional work for ISAE Initial evaluation for data logs already planned	:'-(	:)	:'-(
	Does not require controllers to act as volunteers	:'-(	:)	:'-(
	No additional travel (and associated time, €) Assuming ENAC would be present at the Baseline study. A minimum of two researchers are needed for Roissy	:'-(	:)	:)
Scenario	**Ability to control scenario type** Uncertain if desired scenario will appear	:'-(	:'-(	:)
	Ability to control scenario duration Uncertain if desired scenario will match duration of tests	:'-(	:'-(	:)
	Ability to account for individual behavior Uncertain if it will be same controller in desired scenario/duration	:'-(	:'-(	:)
	Realism of scenario Ability to replicate all details and eliminate the simulation bias	:)	:)	:'-(
	Fidelity of participants to reality	:)	:)	?

	Experienced controllers who have worked the GND ATCO position at Roissy			
	Ability to video/audio record session test Uncertain if can take video during Roissy	?	:'-(	:)
	Ability to test/validate the scenarios before HMI0 and HMI1 We can run extra scenarios in the Baseline test & check if truly cognitively challenging to the controller	:'-(	:'-(	:)
	Availability of data Logs may not capture desired scenario variables	?/:)	?	:)
Debriefing	**Immediacy of tests** May not be possible to get a hold of controllers immediately after the desired scenario	?	:'-(	:)
	Controller ability to discern scenario Controller may confuse scenarios and events, especially with delay in administrative affairs	?	:'-(	:)
	Availability of data Logs may not capture desired scenario variables	?/:)	?	:)
	Ease of calculation No screen capture, no keystroke logger, requires counting	:'-(	:'-(	:)
Data	**Ability to collect Operational measures**	:)	:)	:)
	Ability to collect Behavioral measures	:)	:'-(	:)
	Ability to collect Subjective measures	:)	:'-(	:)
	Ability to collect Neurophysiological measures	:'-(	:'-(	:)

A trip to Roissy requires less work for team Project MoTa but reduces the amount of control over scenario generation. There are several unknowns regarding the ability to sustain accuracy of the subjective measurements or data collection. All data except for neurophysiological measures can be collected. Similarly, using the aircraft logs requires less work for the team on a whole but the data analysis is labor intensive for a small portion of the teal. It potentially allows a substantial amount of control over the scenario. However, there is no interaction with the controllers and there are a number of large risks, including the ability to properly define and evaluate the scenario in accordance to the experimental scenarios. Only operational measures can be taken, although subjective measures can be latently collected during the debriefing sessions of Validation study HMI0 and HMI1. Furthermore, the use of logs can be paired with a trip to Roissy. An additional experiment solely for the baseline scenario demands substantial work for the entire team but provides the most amount of control over the scenarios, debriefings, and the data collection. The data advantages may be worth the development cost, however, especially if the aircraft logs prove to be cost-ineffective.

The stark contrast between the methods with respect to the amount of data measurable prompted an analysis of the significance of each category. Since all methods are guaranteed to collect the operational data, we can comment on the effect of taxiing techniques (including using just the new interface, without TaxiBots or eTaxis) and the effect of the TaxiBot and eTaxis. Having behavioral data for all taxiing technique conditions supports the operational results (such as identifying whether cognitive complexity stems from environmental/operational or display) and permits thorough investigation into the potential shortcomings and problems with current technology. However, not having baseline behavioral data eliminates the latter, but this shortcoming may be negated through concept tuning sessions. With all subjective measures, one can comment, in a quantitative manner, changes in phenomena such as workload, situation awareness, and trust. Without this data, however, there is no means of quantifying the changes due to the interface. Lastly, a full set of neurophysiological measures provides dynamic mapping of changes due to the interface design. Additionally, collecting neurophysiological measures during the baseline experiment would allow for possible integration a half year in advance, compared to October 2014 for the first dataset of neurophysiological matters.

Four key questions have been identified through this comparative analysis and concern the MAS, reaction to a specific event, forecasted traffic loads, scenario complexity validation (Table 11). These questions can help determine whether a baseline experiment is truly needed or not – only one must be relevant. Incorporating neurophysiological measures in the MAS has been determined to be beyond the scope of Project MoTa. Determining the specific variables and quantifying the effect on performance is not within the scope of the Project.

Table 11: Critical Questions in Baseline Experiment Decision

Question	Possible Response and Action
Multi-Agent System Are we trying to incorporate neurophysiological measures and how much development time is needed to incorporate this information? (Answer comes from Project Management)	**< 1 year: No need for a baseline experiment**
	1 year: Need an experiment
Reaction to Specific Event Are we trying to measure the neurophysiological reaction to an event that occurs in the middle of the scenarios, e.g. closure of a taxiway, imminent collision? (Answer comes from Scenario Design)	**Yes: Need baseline experiment**
	No: Do not need an experiment
Forecasted Traffic Loads Are we trying to evaluate whether our tool can handle forecasted traffic loads, something that cannot be observable in current operations? (Answer comes from Scenario Design)	**Yes: Need baseline experiment**
	No: Do not need an experiment
Complexity Tests Are we also going to try to determine which variables contribute to scenario complexity? (Answer comes from Project Management)	Yes: Need baseline experiment
	No: Do not need an experiment

9 References

[1] Project Modern Taxiing. E.02.24-MOTA-D1.1-Definition Phase Report. 06 January 2014.

[2] M.L. Cummings, C. G. Tsonis, "Partitioning Complexity in Air Traffic Management Tasks", The International Journal of Aviation Psychology. Vol 13, No 3, 277-295. 2006.

[3] K. Cardosi, S. Chase, and D. Eon, "Runway Safety", Air Traffic Control Quarterly, Vol. 18, No. 3. 303-328. 2010.

[4] S. Verma, T. Kozon, S. Lozito, L. Martin, D. Ballinger, and V. Cheng, "Human Factors of Precision Taxiing under two levels of Automation". Vol. 19, No. 2. 113-141. 2010.

[5] S.G. Hart, S.E. Lowell, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." Advances in Psychology, Vol 52, 139-183. 1988.

[6] M.R. Endsley ,"Towards a Theory of Situation Awareness in Dynamic Systems", Human Factors. Vol. 37, No. 1. 32-64. 1995.

[7] R.M. Taylor, "Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design." AGARD, Situational Awareness in Aerospace Operations, 1990.

[8] D.M. Dehn, "Assessing the Impact of Automation on the Air Traffic Controller: The SHAPE Questionnaires", Air Traffic Control Quarterly. Vol. 16, No. 2. 127-146. 2008

[9] D.B. Kaber, C.M. Perry, N. Segall, and M.A. Sheik-Nainar, "Workload State Classification with Automation during Simulated Air Traffic Control", The Intl Journal of Aviation Psychology; Vol 17, No 4, 371-390. 2007.

[10] J. Vogt, T. Hagemann, and M. Kastner, "The Impact of Workload on Heart Rate and Blood Pressure in En-Route and Tower Air Traffic Control", Journal of Psychophysiology, Vol 20. No 4. 297-314, 2006

[11] H. Ayaz, P.A. Shewokis, S. Bruce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment", Neuroimage, Vol. 59. 36-47. 2012.

[12] U. Ahlstrom, FJ. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload", International Journal of Industrial Ergonomics. Vol 36, No 7: 623-636. 2006.

[13] J.B. Brookings, G.F. Wilson, and C.R. Swain. "Psychophysiological Responses to Changes in Workload during Simulated Air Traffic Control", Biological Psychology. Vol 42. 361-377. 1996.

[14] F. Faul, E. Erdfelder, E., A.-G. Lang, , A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences". Behavior Research Methods, Vol 39, 175-191. 2007.

[15] "FlightGear Flight Simulator". http://www.flightgear.org. Open-source software.

[16] International Civil Aviation Organization. "Manual of Surface Movement Guidance and Control Systems (SMGCS)." First Edition. 9476-AN/927. 1986.

[17] Directorate General for Civil Aviation, Direction des Services de la navigation aérienne. « Manuel d'Exploitation TWR/APP", DO/SNA-RP/CDG/SE. 2008-2009.

-END OF DOCUMENT-