The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Facial Thermography for Assessment of Workload in Safety Critical Environments

Adrian Cornelius Marinescu

Supervisor: Prof. Sarah Sharples
Dr. Alastair Campbell Ritchie
Prof. Hervé Morvan

Faculty of Engineering
University of Nottingham

This dissertation is submitted for the degree of
Doctor of Philosophy

November 2017

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 50 000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 110 figures.

Adrian Cornelius Marinescu
November 2017

Acknowledgements

I would like to thank the European Union and the University of Nottingham for making this research possible. The project was founded through the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under the REA Grant Agreement No. 608322. The University of Nottingham offered a great environment and access to amazing resources and infrastructure without which this project would have been very difficult to implement.

I am extremely grateful for the continuous support that I have received from my supervisors Prof. Sarah Sharples and Dr. Alastair Campbell Ritchie. Thank you for your kind words and patience, I will always remember going out of our meetings feeling more motivated and enthusiastic about my work. This meant so much to me.

My examiners, Dr. Glyn Lawson and Dr. Dick de Waard, have reviewed this thesis and through their constructive comments have helped me improve it.

Airbus Group Innovations UK and Airbus Helicopters also played a big role in this research and I would like to thank Tomas Sánchez López, Michael McDowell, Gavin Powell, Peter Talbot-Jones and Alistair Nottle for their support during my placement with Airbus, for making the Helicopter Simulator Study possible and for the advice provided during my PhD.

I would like to thank Hervé Morvan for writing the INNOVATE project and for putting together a great team; they were to become some of my best friends. I will never forget our incredibly engaging conversations, I am thankful for all the things I have learned from all of you, including a bit of Italian. Thanks to Horia Maior for all the advice and for the contribution he brought to my last study, I hope that we will plan many more studies in the future.

I would especially like to thank Roxanne Parnham for being supportive and patient during difficult moments[1] and also the Parnham family for making me feel at home.

A special thank you to Amalia Soare and Sorin Soare who have encouraged me to pursue a PhD and who have offered their support throughout this journey.

[1] ATL

I have learned a lot about science and about the world during my PhD, but there is still so much to learn and explore. I could not imagine life without wondering about everything that there is; being allowed to learn about and explore nature is one of the greatest gifts. I am deeply grateful, first of all to my parents, Carmen and Mircea, for having received this gift since early childhood and also to all other people that have offered it along the way. I wish so much that everyone would benefit from similar opportunities.

Abstract

The fast changing modern world is placing humans in positions we have not had time to evolve and adapt to by natural means, we are thus faced with the task of understanding our abilities and limitations, at both a physical and mental levels and design the world around us with these in mind. This is in line with the aim of the discipline of ergonomics (or human factors), to "optimize human well-being and overall system performance by contributing to the design and evaluation of task, jobs, products, environments and systems" (International Ergonomics Association 2014). This is a large task, spanning multiple other disciplines.

The research presented in this thesis is in the area of workload, a concept used to describe the interaction between a task and an operator in terms of demand, perception of task and performance. Many tools and methods have been developed aiming at measuring workload, ranging from subjective measures, primary and secondary task measures, task analysis and physiological measures. The main focus of this research is on physiological methods of assessing workload in safety critical environments. Within the domain of physiological methods for workload assessment, many techniques have been explored over the years and will be presented in the thesis with their advantages and disadvantages. Despite all the efforts made to develop a reliable physiological measurement assessment method for workload, further research is needed; the research presented here focuses on facial thermography as a non-invasive, real-time assessment method for workload, coupled with other physiological measures such as heart rate, breathing rate and pupil diameter.

The human physiological response to changes in workload has been examined in three studies which also explore the use of multiple physiological measures as a means of estimating the level of workload. While two of the studies were performed in laboratory conditions having students as participants, a third study was performed in an ecologically valid helicopter simulator in order to test the physiological reactions of highly trained individuals to changes in workload. The results indicate that facial thermography, especially nose area temperatures, as well as pupil diameter respond well to changes in workload and could be used as a non-invasive, real-time method of estimating workload. The flight simulator study revealed that even highly trained individuals have similar responses to changes in demand as the general public.

This thesis contributes to the measurement and assessment of workload by using physiological measures, especially facial thermography and presenting the relative contribution of each of the measures in both laboratory and real-life scenarios.

Thesis Publications

- Adrian Cornelius Marinescu, Sarah Sharples, Alastair Campbell Ritchie, Tomas Sánchez López, Michael McDowell, Hervé P. Morvan Physiological Parameter Response to Variation of Mental Workload. In: Human Factors: The Journal of the Human Factors and Ergonomics Society, September 30, 2017

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background and Motivation

Since the 1980s, passenger air traffic has doubled every 15 years and it is expected to double again by 2034, with 70% of the traffic relying on the existent network [1]. Near future air transport challenges such as increased air traffic, the need for more efficient routes or the introduction of free flight raise new issues of relevance to human factors. The pilot of the future will have to operate in a more congested airspace, aided by more complex technology. Aircraft pilots and air traffic controllers are just an example of safety critical roles that this research project has looked at. One aspect of human factors that has potential to support the management of increased demand against available cognitive capabilities is workload.

This research explores techniques for measuring the mental workload experienced by humans by using non-invasive and minimally intrusive physiological measurements. These measurements have tremendous potential to aid the real time understanding of human workload, but present a number of challenges in the design and implementation of methodologies (see e.g. [2], [3]).

Mental workload has been suggested to have a strong relationship with human performance, the current consensus being that both excessively high and excessively low levels of mental workload influence performance negatively [4], [3]. Traditionally, some methods of workload assessment have been difficult to implement *in-situ* in a real work environment due to their invasive nature, as it has been necessary to interrupt tasks or wear uncomfortable equipment. Advances in physiological sensors and data analysis techniques mean that tools such as facial thermography [5] are now realistic candidates for non-invasive capture of workload in real time. Lehrer *et al.*, concluded in a flight simulator study that the minimum R-R intervals (time interval between heart beats extracted from ECG data) in a task significantly discriminated between high and combined moderate and low-load tasks [6]. Eye movement

activity was used by Ahlstrom and Friedman (2006) in an air traffic control study and they concluded that blink duration, blink frequency, mean saccade distance and pupil diameter can provide a sensitive measure of mental workload. They have established that an increase in experienced mental workload level is correlated with an increase in pupil diameter [7]. For such candidate measures to be deployed, new knowledge is required to establish the validity, reliability and sensitivity of such tools. Previous studies have explored whether it is possible to infer mental workload by using facial thermography. These studies have shown a high correlation of workload with the decreasing temperature of the skin covering the tip of the nose. Ora and Duffy (2007) first used a simulator driving task together with a mental arithmetic loading task to increase the mental workload while measuring nose and forehead temperature followed by performing a study in a real car driving situation. They demonstrated that there is a strong correlation between the change in nose surface temperature and the subjective ratings for mental workload while the forehead temperature remained relatively constant [5]. Another study in a ship simulator showed that nasal temperature and heart rate variability are good indices for effective navigation, and also connected the measures to the variation of mental workload [8]. The reason for the nose temperature drop proposed by Ora and Duffy is the vasoconstriction response of the autonomic nervous system to mental stress or negative emotion, mediated primarily by the sympathetic nervous system. Thermal imaging of the forehead, nose, eyes, cheeks and chin during a cognitive stress test was able to classify mental workload into three levels with 81% accuracy [9]. However these studies did not establish the ‘added value’ of facial thermography as a physiological tool over other techniques such as heart rate or pupil diameter/eye movements (both of which can require the use of more intrusive and personally worn monitoring equipment).

1.2 Thesis Research Questions and Aims

Current technology advances have enabled a growing number of researchers to examine various types of physiological data for assessment of workload, especially in safety critical environments. While some approaches have higher face validity, especially brain measures such as functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS) or electroencephalography (EEG) they are either not portable or require the user to be wearing equipment on their head, making them more suitable for use in laboratory conditions. The ideal sensing technology for these types of applications would be "invisible" to the user, would not interfere with the task and would have high validity.

The aim of this research is to explore means of objectively estimating human experienced workload levels while performing a task, by analysing physiological data recorded with non-invasive sensors in terms of their validity, reliability and sensitivity.

The aim presented above translates into the following research questions:

1. How do human physiological responses change in response to variations in task demand and task performance?
2. How are physiological responses associated with variations in subjective reports of mental workload?
3. Can multiple combined physiological parameters explain more of the variability in mental workload or performance than individual parameters?
4. How do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator?

1.3 Thesis Contributions

Answering these research questions required a multidisciplinary approach, bringing together the fields of human factors, image processing and machine learning. The main contribution is brought in the field of human factors. The contributions of this thesis can be summarized as follows:

- This thesis contributes to the assessment of facial thermography as a non-invasive real-time method for workload measurement.
- This thesis contributes also to the assessment of combined physiological measures (facial thermography, pupil diameter, cardiac measures and breathing rate) for workload measurement and their individual contribution.
- The final contribution of this thesis is showing how measures such as facial thermography, heart rate measures and breathing rate respond to changes in demand for highly trained individuals in an ecologically valid aircraft simulator.

1.4 Thesis Overview

This thesis contains 8 chapters. Table 1.1 provides an overview of the chapters together with the methodology and addressed research question.

	Description	Methodology/ Addressed research question
Chapter 2	Background on workload and methods of measuring workload	Literature review
Chapter 3	Introduction to physiological measures of workload	Literature review
Chapter 4	Introducing the methodological approach, sensor description and data analysis methods used	
Chapter 5	Examining the relationship between experienced mental workload and physiological response by non-invasive monitoring of physiological parameters	• Empirical study (10 participants) • Research questions: 1, 2, 3 • Task: Visual computer task • Physiological measures: Heart Rate, Breathing Rate, Pupil Diameter, Facial Thermography • Publication: Physiological Parameter Response to Variation of Mental Workload. In: Human Factors: The Journal of the Human Factors and Ergonomics Society, September/2017
Chapter 6	Examining the relationship between experienced mental workload and physiological response by non-invasive monitoring of physiological parameters for highly trained individuals in an ecological valid aircraft simulator	• Empirical study (8 participants) • Research questions: 1, 2, 4 • Task: Flight Simulator Scenarios • Physiological measures: Heart Rate, Breathing Rate, Facial Thermography
Chapter 7	Examining the relationship between experienced mental workload and physiological response by non-invasive monitoring of physiological parameters - Study 1 extension	• Empirical study (11 participants) • Research questions: 1, 2, 3 • Task: Visual computer task • Physiological measures: Heart Rate, Breathing Rate, Facial Thermography, fNIRs
Chapter 8	Conclusion and Discussion	

Table 1.1 Description of chapters

Chapter 2

Literature review of workload

This chapter offers an overview on workload literature in terms of proposed models and means of measuring it. Even though the construct of workload appears to be simple, there is no universally accepted definition for it: as Linton *et al.* stated in 1989: "The simple fact of the matter is that nobody seems to know what workload is. Numerous definitions have been proposed, and many of them seem complete and intuitively 'right'. Nevertheless, current definitions of workload all fail to stand the test of widespread acceptance or quantitative validation."[10] Even so, workload is a popular and useful tool in the discipline of ergonomics and human factors.

The first part of this chapter introduces the concept of workload and the main models describing it, and the second part of the chapter presents the most spread measures of workload as well as a list of criteria for the measures.

2.1 Workload

Even though many definitions of workload have been proposed over the years, the common aspect that unites them is their end use. They all aim to provide a tool that will enable the "evaluation of tasks, jobs, products, environments and systems in order to make them compatible with the needs, abilities and limitations of people" with the goal of "optimizing human well-being and overall system performance". This quote comes from the definition of ergonomics (or human factors) as stated by the International Ergonomics Association (2014).

Mental workload is one of the most widely used concepts in ergonomics and human factors [4], as jobs in the past decade have had an increasing mental component. In the view of Sharples and Megaw, the cognitive activity takes place in a social context, interacting with physical elements that should not be ignored [3]. Various definitions of workload have been proposed. Young and Stanton (2002) defined mental workload as "the level of

attentional resources required to meet both objective and subjective performance criteria, which may be mediated by task demands, external support, and past experience" [11]. In the context of multiple resource theory, Wickens (2008) considers that the concept of mental workload relates to the demand component of the theory, referring to the human limited mental resources that the demand is placed on [12], [13].

To offer a better description on workload, some researchers have proposed models and frameworks to describe it; the list below contains some of the definitions and models that are presented in this chapter providing help in understanding workload and that are used to describe how workload was manipulated during the studies presented in this thesis:

- Multiple Resource Model
- Limited Resource Model
- Framework for workload definition and evaluation

2.1.1 Multiple Resource Model

While not a workload theory, Wickens' Multiple Resource Model is based on the separation of resources in the brain and it focuses on predicting situations of overload, when multiple tasks compete for the same resources. This theory and workload assessment tools based on it consider the structural dimensions of information processing and account for the ability of the human to time share more than one process, predicting the degree of interference between different tasks. [14]

Wickens initially started with a three dimensional multiple resource model that was later extended to four dimensions. The four dimensions of the multiple resource model were represented using a cube: Fig.2.1. The stages dimension is split into two types of resources: perception together with cognition and responding. The access dimension reflects the distinction between spatial and verbal processing. The modalities dimension was updated since the initial model and now contains the visual channel (focal and ambient), the auditory channel and the tactile one. "To the extent that two tasks demand separate resources along these four dichotomous dimensions, (a) overall time sharing will improve and (b) increases in the difficulty of one task will be less likely to degrade performance of the concurrent task" [15].

Wickens' multiple resource model, has extended upon the ideas of attention capacity models [16] and while not indending to be a workload theory, it aimed at explaining which type of tasks can be performed simultaneously by a human without degrading the performance due to the sharing of resources. Johnson *et al.* (1998), based on the multiple resource theory,

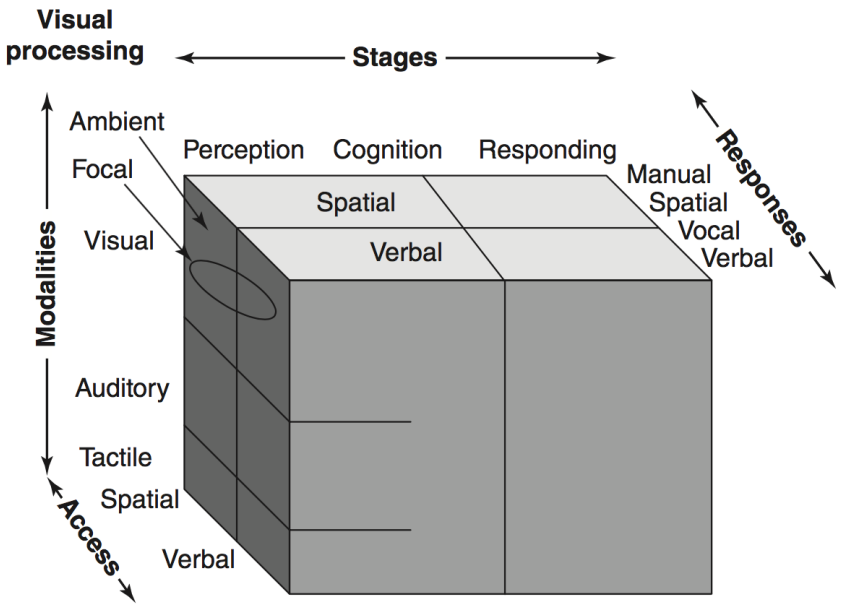

Fig. 2.1 Multiple Resource Model [15]

tested the effects of different distractors in reducing pain, predicting that the more a distractor shared processing resources with the perception of pain, the greater the interference. The results obtained were contrary to their prediction [17]. Navon (1984) warns that "attempts to measure workload, to identify resource pools, to predict task interference by performance resource functions...may prove to be disappointing as would attempts to isolate computer components inside the human mind" [18]. Other aspects that the multiple resource theory does not take into account are related to the capacity available for each of the resources at a given moment (these could be influenced by the level of arousal, mood, motivation and age), how the resources are allocated as it seems to be influenced by the strategy adopted at individual level and what part learning plays [14]. Another aspect that the multiple resource model does not take into consideration is the appraisal of the task by the operator, as the perception of workload depends not only on the demands imposed but also on how they are perceived.

2.1.2 Limited Resource Model

Kahneman (1973) was one of the first to consider the human as having a limited capacity central processor; ‘multiple task work’ places demands on the operator’s information processing resources, requiring a choice of strategy from the operator and the appropriate allocation of

resources [14]. If mental workload is to be conceived in terms of limited available resources and task demands, Fig. 2.2 would offer a simple description of the relationship. The vertical axis on the right shows the performance on the primary task while the vertical axis on the left indicates the resources being used. The maximum available resources are indicated by the horizontal dotted line. For as long as the resourced allocated to the primary task are lower than the maximum available resources, performance is maintained at a high level and there are even spare resources available, described by Wickens *et al.* as *reserve capacity* [15].

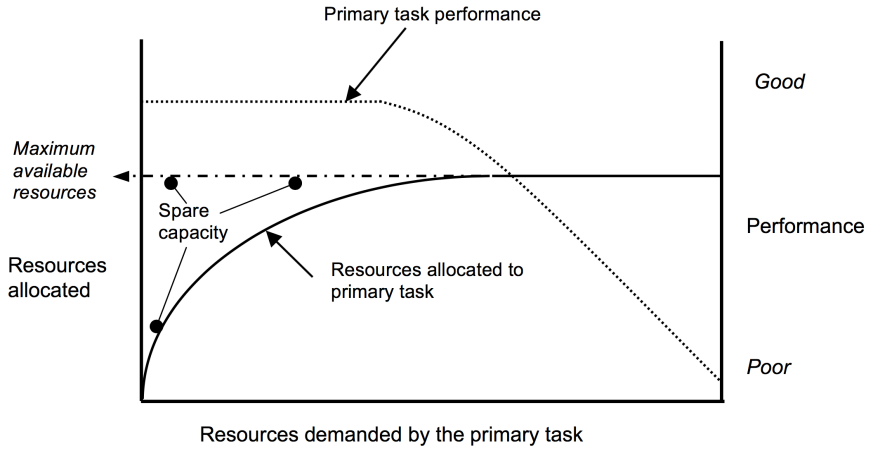

Fig. 2.2 Limited resource model [15]

One important aspect predicted by this model is that while spare resources are available, even if the demands of the task are increasing, performance will not suffer and thus not be a good indicator of workload. As primary task demands go beyond the maximum level of resources, performance starts to decline and can now be said to be inversely related to primary task performance [3].

Sharples and Megaw have considered a number of issues that the above hypothetical model has and illustrated them using a graphical representation presented in Fig. 2.3. In order to accommodate additional factors (for example stress) they have introduced a variable level of available resources. The impact of expertise is modelled by having different ascending gradients in the solid lines, representing different rates at which resources are demanded by the task. The impact of underload is also considered by having a lower than maximum performance when the demand for resources is very low. Based on the ATC (Air Traffic Control) research performed by Sharples *et al.* [19] and Edwards *et al.* [20], the decline in performance is depicted as less graceful, including what they have called a "precipice in performance". Their improved model also includes a dip in performance due to data limitation while still on the left side of the graph where spare resources are still available.

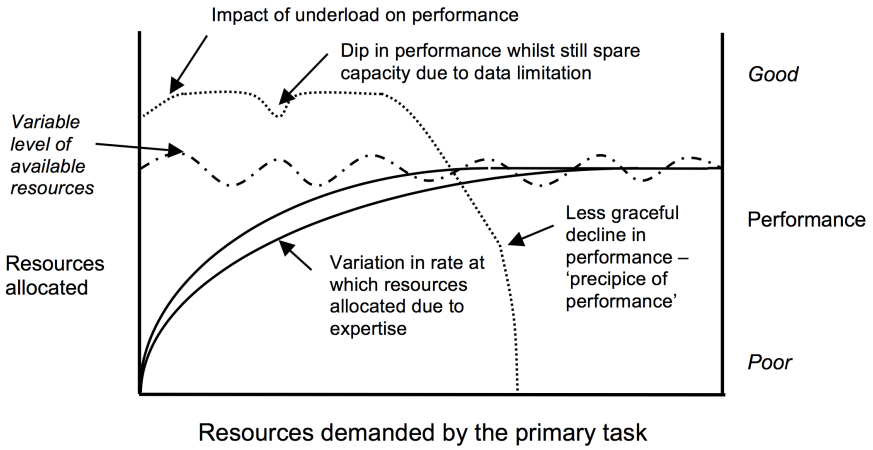

Fig. 2.3 Limited resource model [3]

2.1.3 Framework for Mental Workload

Sharples and Megaw have proposed a dynamic framework for understanding the implications for the measurement process of mental workload [3]. The framework contains three components (Figure 2.4)

- Physical and cognitive task demands: reflecting the characteristics of the task imposed on a person and noting that externally measurable demand may be different from the demand experienced by the individual;
- Operator workload: "equivalent to measuring operator strain or effort" [3]
- Performance: often described in terms of speed and errors

The relationships between the components in Figure 2.4 are described by Sharples and Megaw in the following way:

1. Operator workload is not seen as simply a function of task demand, as it is influenced by how the task is perceived. Pickup *et al.* [14] talked about effort as being a consequence of demand created by loading factors, in this case represented by the *physical and cognitive task demands* and *external and internal influences* presented at point 5.
2. Operator workload and performance are not always negatively correlated as expected, as it is difficult to detect when an individual is working harder to maintain the same level of performance; this indicates that performance alone should be used with caution in estimating the level of workload.

3. Performance feedback, either internal or external, can affect the way the task is perceived and thus influence workload.

4. Performance can in some cases modify the task itself in a way that task demands are altered.

5. External influences (e.g. job type, organisational and safety culture or team support) can influence perception of workload. Internal factors such as operator skill and motivation can influence workload directly or indirectly through the chosen operator strategies.

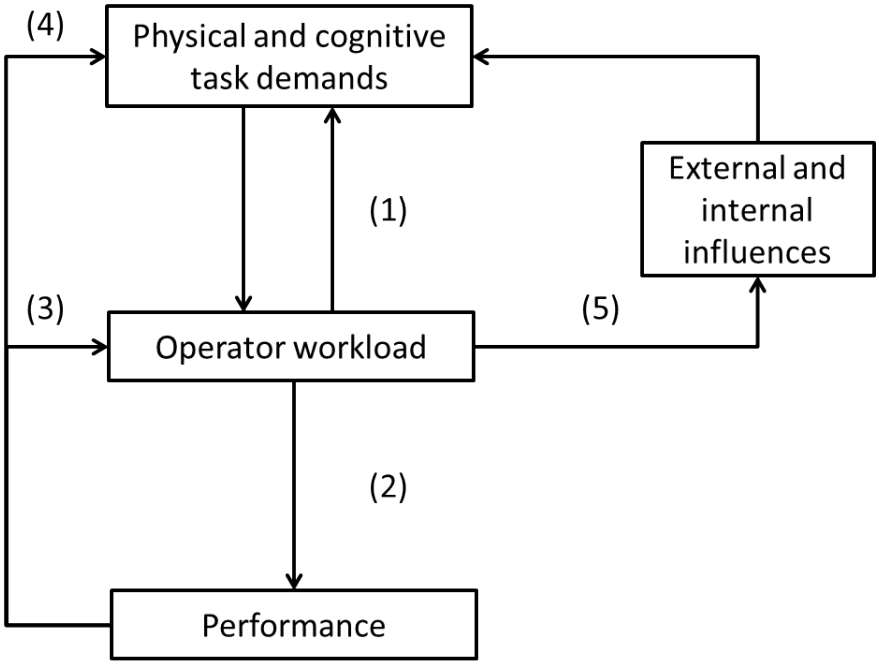

Fig. 2.4 Mental Workload Framework [3]

This framework will later be referenced when presenting the task used in the first and third study presented in this thesis.

2.2 Measures of Workload

Before the second part of this chapter introduces measurement techniques of workload, the currently accepted criteria that these measures are compared against will be presented.

2.3 Criteria for Workload Measures

This section will present a list of criteria for assessing the quality of workload measurement techniques, proposed by O'Donnell and Eggemeier [21], Wickens and updated by Sharples and Megaw in [3]. These criteria are listed below, each having attached a question that will later be used in the thesis to describe the measurement techniques that were explored.

1. **Validity:** is the method measuring workload?
    - Face Validity: would the measure be accepted as measuring workload by all involved stakeholders?
    - Construct Validity: does the method measure all aspects associated with workload?
    - Concurrent/Convergent Validity: do multiple workload measurement methods show the same trends?
2. **Reliability:** if applied multiple times, are the results provided by the method consistent?
3. **Generalisability:** can the method be applied in multiple domains?
4. **Sensitivity:** how small are the changes in task demand or performance that can be detected by the method?
5. **Interference:** does the method technique interfere with the primary task?
6. **Diagnosticity:** can the method distinguish between demands incurred by different resources (as described above in the Multiple Resource Model)?
7. **Selectivity:** can the method distinguish mental workload from other factors such as emotional stress or physical workload?
8. **Granularity/bandwidth:** what is the time resolution of the method?
9. **Feasibility of use:** how easy would it be to use the method?
10. **Acceptability and Ethics:** would the participant and those around him accept the measurement applied?
11. **Resources:** how practical is the method in terms of required financial and time resources?

2.3.1 Subjective Measures of Workload

Subjective measures of workload are some of the most popular mainly due to their ease and low cost of use. They also have high face validity as they intend to capture how the participant perceived the effort they invested into the task. Numerous measures of workload have been developed, but only the two measures chosen for this research will be discussed in this thesis.

ISA

The Instantaneous Self-Assessment (ISA) [22] is a single dimensional scale and was developed primarily as a subjective measure of mental workload for air traffic controllers; it involves the participants self-rating their workload on a scale from 1 (low) to 5 (high). One of the main advantages of this scale, as the name implies, is that it is an instantaneous judgement, allowing the researcher to probe inside the task as it is developing; it is not relying on retrospective memory. The disadvantage is that it may interfere with the task. If carefully applied, interference can be minimised.

NASA-TLX

NASA-TLX (Task Load Index) [23] is one of the most popular subjective multidimensional scales for measuring workload. The measure can only be used retrospectively and provides insight on six subscales containing the following factors: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. There are two steps in administering the measure. First the participant is asked to weigh the factors; this is done using 15 pairwise comparisons of each of the factors and the participant is asked to select one in each case that they thought was more influential on the task. This will reflect the contribution of each factor to the workload. The second step involves the participant rating each of the factors on a scale having 20 increments and a bipolar description of LOW/HIGH at each of the ends [23].

2.3.2 Primary and Secondary Task Measures of Workload

These techniques have a higher level of objectivity compared to the subjective measures. They aim at inferring the level of workload based on the performance level of either the primary task or a secondary task. The main disadvantage with this approach is that, as discussed above, in the limited resource model, task performance might not be sensitive to changes in workload while there still is spare capacity even though there is a variation of workload. Introducing a secondary task will partially solve this problem and reflect the level

of workload during the spare capacity region, but in doing so might become very intrusive to the primary task. Nevertheless these methods have the advantage of being easy to apply, primary task performance will be used later in the thesis mainly as a concurrent measure of workload.

2.3.3 Measuring Workload in Real Life Settings

Most of the studies involving the measurement of workload are performed in laboratory conditions; in the majority of the situations, this allows for a carefully controlled environment both in terms of the performed task and external influences. While this approach has its advantages, it may sometimes not reflect the complexity of the real life environment.

Field studies present challenges ranging from little control over the conditions [24], measurement of participant state [25], setting up the recording equipment, to the ethical implications it might bring. In the context of flight test evaluation, Bonner and Wilson [26], in a comparison study between simulated and real flight, mention several problems that they encountered, starting with mission length which may be between 1 and 12 hours implying the recording equipment should be mobile to allow for the crew to move around, and the timing of movements as to account for it in the processing of the data. The variable nature of the flights is also discussed as it makes the use of inferential statistics more difficult. They also mention one very interesting aspect about subjective ratings in simulator vs. real life situations, the dissociation from physiological measures; they observed little changes in heart rate in simulated emergencies while the mean subjective ratings increased. Bonner and Wilson propose that the subjective ratings might reflect the assumed demands while heart rate might be sensitive to the actual demands. During real landings they observed the opposite effect, increases in heart rate while the subjective ratings were low, proposing that heart rate reflects the actual demand while the subjective ratings were low due to the routine action of landing. The only moments when they reported a high correlation between subjective ratings and heart rate was during extremely high workload situations like touch and go on an icy runway. They conclude that heart rate adds valuable information to the test and evaluation missions.

In the study of workload, a lot of the interest revolves around safety critical environments (e.g.: air-traffic control, railways, driving or flying). These environments present great challenges in performing field studies firstly due to the risks involved and secondly due to the complexity and cost.

High fidelity simulators can help overcome many of the challenges, allowing for better control over the conditions, reduced risk and lower costs; all these while compromising on the risk perception experienced by the participant and possibly simulator sickness [27].

Nevertheless these type of studies are very important and they are probably the closest to replicating the real life conditions; they also require good research skills to setup [25].

In the flight domain, numerous simulator studies have been performed. Gateau *et al.* explored the use of fNIRS measuring blood oxygenation in the prefrontal cortex as an online measure for pilot state estimation. They were successful in classifying low vs. high working memory load with 80% accuracy in a flight simulator task [28]. More studies exploring physiological measures will be presented in the next chapter.

The main disadvantage with physiological measures is that they are not influenced solely by variations in workload, there may be many other reasons for a response. One of the aims of this research is to explore which physiological measures show a good response to variations in workload and to also test if multiple measures used together can provide a better indication of variations in workload.

2.4 Chapter Summary

While the concept of workload is appealing, mainly due to its apparent simplicity, confusion can appear many time between researchers [29]. Over time, the definitions and models attempting to clarify what workload really is have improved, but still there is no widespread accepted model or definition. This chapter has offered a brief overview of the main definitions and models of workload and presented some of the subjective methods used to measure it. As it represents the main focus of this thesis, a separate chapter will be dedicated to the literature review of physiological measures of workload, presenting some of the most used techniques.

Chapter 3

Physiological Measures of Workload

This chapter aims to introduce the most used physiological measures applied in workload assessment and also the possible underlying physiological phenomena responsible for the physiological changes. It begins by justifying the use of these type of measures to infer the level of workload and then goes on to introduce the physiological measures used in the studies presented in this thesis as well as a few others that have not been used but show great promise. Physiological measures of workload aim at solving the shortcomings of the other measuring techniques. By assuming that the perceived workload experienced by someone will be reflected in their physiological changes; they seem to be offering a way of circumventing the subjectivity or intrusiveness of other measures. The models described above refer to a pool of resources into which the person taps in order to complete a task. These resources most likely reside inside the human brain where most of the processing and fusion of information takes place. Physiological measures rely on the assumption that as phenomena occur inside the brain in order to address the demands placed on the person, various resources are mobilized. As this is happening, physiological responses occur, which may include changes in heart rate, breathing rate, pupil diameter, and skin temperature; it could be said that these changes are symptoms of workload. Most of the physiological parameters that are monitored are under the control of the autonomic nervous system, which is not under conscious control conferring an advantage as it avoids the subjectivity factor.

3.1 Reasons to use Physiological Measures

As presented in Chapter 2, the framework for mental workload offers an understanding of the implications for the measurement process of mental workload. Directly or indirectly, operator workload is influenced by physical and cognitive task demands, by performance and by external and internal influences. The relationships between these four elements can

be complex for more elaborate tasks and measuring workload can prove difficult. One type of measurement technique that has good face validity is subjective measurements; discussed in section 2.3. A subjective measurement probes how the operator perceives the task, which is very valuable. Some drawbacks of these methods are: they can prove to be intrusive to the primary task; the operator may be biased; some of them are retrospective, relying on memory; and finally they cannot be deployed as continuous real time measurements. Another measurement approach that aims to probe "inside" the operator to reveal how they are perceiving the task relies on physiological measures. The assumption is that as physiological responses address the demands imposed on the operator, there have to be some measurable changes. The advantages are mainly that some of the parameters are under the control of the autonomic nervous system, and not under conscious control, so these measures can be continuously recorded without interfering with the primary task and they have the potential to offer high face validity, provided that the right parameters are chosen. Over the years, researchers have attempted to look at various physiological measures to infer the level of workload. Cardiac and respiratory measures, eye response measures, facial thermography and brain activity measures will be reviewed in this chapter.

3.2 Cardiac and Respiratory Measures

This section will offer an overview on previous research looking at exploring the potential of cardiac and respiratory measures for workload estimation.

The rhythm of the heart is modulated by the sinoatrial node, which is influenced by both the sympathetic and parasympathetic branches of the autonomic nervous system (ANS). There is a continuous balance between the two branches of the ANS; the sympathetic activity increases the heart rate while the parasympathetic branch decreases it. Because it is controlled by the ANS, cardiac activity has been considered a good candidate measure for workload.

Casali and Wierwille (1983) conducted a study to investigate sixteen potential metrics of pilot workload. One of the tested measures was heart rate and heart rate standard deviation, acquired by plethysmography. The study was conducted in a flight simulator and presented 30 participants with cross-country flights. The variation of demand was manipulated in three levels by means of communication call signs; these were transmitted over the radio at various time intervals and the participant was supposed to depress the push-to-talk button when hearing a target call sign. Their findings revealed that mean heart rate, heart rate standard deviation and breathing rate computed over seven minutes time intervals were not able to discriminate between the three levels of demand [30].

Brookings el al (1996) assessed multiple physiological measures in air traffic control tasks to determine if they would be sensitive to changes in workload. Eight air traffic controllers performed three types of scenarios in an air traffic control simulator. They concluded that heart rate did not demonstrate significant differences to the demand manipulations. In the same air traffic control study they also assessed breathing rate and breathing amplitude for its sensitivity to changed in workload. They concluded that the respiration amplitude was not affected by the change in air traffic manipulation whereas breathing rate was higher as the complexity of the scenario increased but was not affected by the increase in traffic [31].

Verwey and Veltman (1996) conducted a driving study in an instrumented car, assessing the mental workload experienced by 12 participants using various measurement techniques. Among the monitored parameters were heart rate variability and inter beat intervals. Both have proven not to be very sensitive to changes in workload; inter beat intervals marginally distinguished between two conditions of different workload and heart rate variability was lower in the no-loading task driving periods compared to the baseline [32].

Bonner and Wilson (2009) have recorded heart rate data from pilots, copilots and loadmasters during test and evaluation flights, including aircraft handling and normal flight. Even though during the variable nature of the flights inferential statistics could not be used, the points that were made are that there was substantial increase in heart rate for the pilot in control which could play a role in identifying sharing of workload problems inside the cockpit; another point that was made is that increases in heart rate similar to the ones during high workload were reported when the crew members were moving through the aircraft, removal of these artefacts should be considered [26].

Svensson and Wilson (2009) in a flight simulator study recorded heart rate during 35 simulated combat missions. Heart rate data was averaged over two minutes time intervals and significant differences in heart rate were found between approach and intercept phases [33].

Wilson (2002) studied the heart rate of ten general aviation pilots in a scenario divided into three parts meant to induce a variation in workload. One interesting aspect of this study is that each pilot flew the scenario twice within a few weeks. The results showed that the physiological measures recorded were consistent between the two flights. Heart rate increases were reported during takeoffs and landings, also differences in heart rate were reported between scenario conditions [34].

In an exploratory study P. Lehrer *et al.* (2010) have assessed whether cardiac measures could be used to measure workload. Seven professional pilots took part in their flight simulator study involving 18 flight tasks. The mental workload for the tasks was rated by both experienced test pilots as well as the pilots performing the task using NASA-TLX.

Flight performance was evaluated by experts on a five point scale. They found that SDNN (standard deviation of normal R-R inter-beat intervals) was associated with expert ratings of mental workload even when the NASA-TLX results of the participants were not, suggesting that the cardiac measures assess something that the NASA-TLX does not [6].

Haapalainen *et al.* (2010) found that the electrocardiogram median absolute deviation and median heat flux were the most accurate at distinguishing between low and high levels of cognitive load. They used a computer based task focused on visual perception and cognitive speed to vary the level of demand while multiple sensors collected physiological data, including heart rate [35].

Brookhuis and de Waard (2010), in a driving simulator study, have reported that on average, the heart rate of 20 participants when crossing a junction and when accepting a gap between oncoming cars in order to turn left has increased with 5 beats per minute [25].

Stuiver *el al.* (2014) conducted a driving simulator study investigating if the sensitivity of cardiovascular measures computed on short time segments (30-40 s) is enough to detect short-lasting changes in mental effort. Fifteen drivers completed six one hour driving sessions in two different traffic density conditions (low and high) each with or without fog. During the study, heart measures and blood pressure were recorded. They have reported a decrease in heart rate variability and systolic blood pressure variability during the increases of workload caused by fog and also that the decrease was stronger in the low traffic condition. During the high traffic condition, blood pressure was observed to be higher compared to the low traffic condition. The largest cardiovascular effects caused by fog were found in the low traffic condition; this was attributed to a possible ceiling effect of variability measures [36].

A laboratory conditions tracking task (pitch and roll), displayed on a small screen was used by Spyker *et al.* (1971) explored the use of physiological measures to determine the level of pilot workload; heart and respiratory measures were used as well as skin impedance and EEG. They found the breathing rate data to be encouraging as it showed visual changes in some of the participants. Respiration seemed to increase in amplitude and decrease in regularity as the participants transitioned from the easy to the hard scenario, suggesting that they are affected by workload; among all the features, respiration was found to provide the strongest response [37].

While not all results presented by the research studies mentioned above agree with regards to the response of cardiac and respiratory measures, it is most likely either because the measures were not sensitive to the manipulation of demand either because other external factors influenced the measures. Cardiac and respiratory measures can be easily influenced by factors not related to mental workload, as reported by [26]; while performing a study in a high fidelity simulator will increase the realism of the scenarios, the physical work invested

into flying together with the movement required to reach instrumentation as well as the radio communications can easily lead to variations in the recorded physiological data. While this is not a negative aspect and should definitely be examined if such technology is to be developed for use in real world scenarios, for the first study presented in this research, it was decided to keep the movement and verbal communications of the participants to a minimum.

3.3 Pupillary Response

Pupillary response has been studied in connection to mental effort since the 1960s, Hess and Polt (1964) recorded the pupil diameter of participants performing simple multiplications of increasing difficulty and reported the increase of pupil diameter with the increase in difficulty [38]. Kahneman and Beatty (1966) demonstrated that pupil diameter increases during a short term memory task when the participant is instructed to memorize a string of digits and constricts as the participant reports back the digits [39]. Krebs *et al.* (1977) investigated the potential relationships between eye behaviour and pilot workload in a series of flight simulator scenarios. Even though pupil diameter showed a strong relationship with task difficulty, upon re-examination of the data, it was concluded that potential limitations of the recording equipment might have influenced the measures as the pupil appeared elliptical to the oculometer used when the eye rotated away from the centre (as when monitoring various peripheral equipment) [40].

Two of the sixteen metrics used by Casali and Wierwille (1983) were pupil diameter and eye-blinks. To have a better measure of pupil diameter, given the challenges mentioned by Krebs *et al.*, pupil size was divided by iris size only when the subjects were fixated on the attitude indicator (measured approximately every 10 seconds). The pupil diameter measure was the only one that reliably discriminated low from medium loading and low from high loading; it did not discriminate medium from high loading. The authors do advise that the results should be considered with caution as the apparatus for measure was not very advanced and as the results are non-monotonic [30].

In a laboratory study, Recarte *et al.* (2008) aimed to compare three measures of workload (the NASA-TLX subjective measure, pupil diameter and blink rate). Using a combination of cognitive and visual tasks they found that pupil diameter was larger when performing a cognitive and detection task than when performing only the cognitive task; this effect was found to be smaller than the one measured by the NASA-TLX subjective scale [41].

De Greef *et al.* (2009) studied multiple eye measures under different levels of workload as potential objective measures for it. Eighteen participants performed the role of human operators in a simulated combat management scenario. Three different cognitive load

scenarios (underload, normal, overload) were designed and subjective workload ratings as well as pupil diameter, fixation time and saccade distance measures were obtained. The results revealed a significant difference in pupil diameter between the three scenarios except in distinguishing the normal scenario from the overload one. Fixation time was able to provide a way to discriminate between the three levels of workload with higher fixation times for increased workload. Saccade distance and saccade speed showed no significant differences [42].

Di Stasi *et al.*(2013) proposed the use of eye measures as a means of estimating workload. In their study, forty-four participants completed a computer based fire incident simulator where they had to minimise the spread of fire in a forest by using either a water or a control fire strategy. The fire strategy was expected to put more demand on the participant as it required a higher level of planning. Each participant performed 20 trials, the first 16 being identical while the last four would introduce and unexpected change in wind direction. It was found that the pupil diameter decreased for both groups in the last four trials which differed from previous research indicating an increase in pupil diameter with the increase in demand [43].

Most of the studies presented above, apart from the last, agree with the physiological response of increase in pupil diameter with the increase in workload. As this is a minimally intrusive measure, it was decided it was going to be used during the first study presented in this research. While pupil diameter is also sensitive to light intensity and it would potentially be difficult to use outside laboratory conditions, the task designed for the study presented in Chapter 5 had the advantage of not inducing decreases in light intensity, making it an ideal environment to test the response of pupil diameter to the variation of demand.

3.4 Facial Thermography

Facial thermography is a measure of skin temperature on the surface of the face, it can be measured by thermocouples placed on the skin or, less intrusively by a thermal camera. Skin temperature is strongly connected to the blood flow in the area, which is under sympathetic control of the nervous system.

Naemura *et al.* (1993) have studied the effect of loud noises on nose temperature. Fifty-two participants split into two groups were subjected to noises of 45 dB and 100 dB. They found that under the 100 dB noise, the nose temperature decreased by about 0.5°C, while there was no significant change for the 45 dB condition. One other observation that they

made was that the temperature drop was larger in female participants compared to male participants [44].

Genno *et al.* (1997) used facial skin temperature to infer stress. The study used two tracking task periods interrupted by an emergency in the form of an alarm that could not be cancelled unless the participant introduced a 12 character password learned at the beginning of the study. Before and after the tracking tasks, rest periods were introduced. Nose temperature was measured using thermistors attached to the nose, forehead, cheeks, chin and ears. The results revealed an increase in nose temperature during the rest periods and a drop in nose temperature during the tracking tasks; it was observed that there was a further drop in nose temperature during the emergency period, which is in agreement with the effect reported by Naemura *et al.* [44]. Forehead temperature was reported to be constant opening up the possibility of using it as a reference for the nose temperature [45].

The facial thermography studies presented above did not benefit from the cheaper, more accurate, smaller and easier to use thermal cameras available at the time of writing, and limitations in the technology may have contributed to the small number of studies in the field. In 2007, Ora and Duffy reported the results obtained in their research into developing a facial temperature based non-intrusive measure for mental workload. Their study included both a car simulator scenario as well as a real car driving scenario. The simulator task involved driving in city-like and highway-like situations, both of them with and without a mental arithmetic loading task. For each of the conditions, the temperature was measured before and after the stimulus; a Modified Cooper Harper subjective questionnaire was administered after each drive. The mental arithmetic conditions induced a significantly larger temperature drop in the nose area compared to the conditions with no secondary task; nose temperature was found to be correlated to the subjective scores. The real driving study was conducted to see if the temperature changes are similar to the ones in the simulator. The forehead temperature was almost constant in both simulator and real car. As opposed to the simulator, the car scenario showed no significant changes in nose temperature [5].

In a ship simulator study, Murai *et al.* (2008) evaluated nose temperature together with heart rate variability as a means of estimating mental workload. A professional ship captain took part in the study, the task being that of entering a ship into a port. A pair of spectacles was used as reference points for selecting the nose and forehead region of interest. The forehead temperature was used as a reference for nose temperature. It was confirmed that the nose temperature dropped when the navigation difficulty increased [8]. This study was extended by the same researchers (Murai *et al.* 2015), with professional port coordinators participating in the study. In this second study, temperature was sampled every 30 seconds from a point located in the middle of the nose; the initial results were confirmed, the nose

temperature decreased when the port coordinator was involved in decision-making. This was attributed to higher levels of demand being placed on them [46].

Kang *et al.* (2008) aimed to validate the efficacy of thermography as a measure for assessing mental workload. Twenty participants completed seven task blocks (96 questions each) of alpha-numeric tasks and had 4 seconds to respond to each question. Nose and forehead temperatures were recorded as well as subjective ratings on the Modified Harper Scale and SWAT (Subjective Workload Assessment Technique). Performance was assessed in terms of participant reaction time and task accuracy. Nose temperature was found to be significantly affected, decreasing during the first block when learning was occurring and gradually increasing up to the fifth block after which it started to slowly decrease again. Forehead temperature was not affected. Another interesting finding is that no gender differences were found. Although the stimulus was different, Naemura *et al.* (1993) reported gender differences in response to loud noises [44]. Kang *et al.* concluded that thermography is a reliable and valid method for identifying learning progression by detecting mental workload [47].

Reyes *et al.*(2009) used facial thermography to measure driver response to in-vehicle human-machine interfaces. Sixteen drivers took part in a driving simulator study and completed 60 second tasks (radio tuning, CD-switching, address entry for navigation) with two types of interfaces (jog dial and touch screen). Four facial areas were evaluated: forehead, nose, inner eye and entire face over three time windows (15, 30 and 60 seconds). Tracking was accomplished by the placement of small metal trackers above the eyes. Mean nose temperature was found to be correlated with lateral control input performance measures. The standard deviation was more sensitive for eye, forehead and face while mean temperature was more sensitive for the nose. The authors conclude that facial temperature measures correlate with driving performance as well to subjective ratings of workload and were most significant when computed over a 15 second window. In the conclusions, the authors also mention that separating different driver states (mental effort) from other dimensions (frustration or anger) is a challenging task [48].

Stemberger *et al.* (2010) proposed the use of a neural network classifier together with facial thermography data for estimating the level of workload levels. Twelve participants participated in a cognitive stress test consisting of three blocks. The low workload level asked the participants to press a button when they saw a number in the sequence of displayed numbers. For the medium level of workload, they were required to press the button when three even numbers were displayed consecutively while for the most difficult level they were required to press the button when the digit was identical with the one displayed two trials earlier. Seven areas of the face were tracked, including nose, eyes, chin, cheeks and forehead.

Subjective ratings confirmed that the task manipulation induced three different levels of demand even if performance measures did not allow for the discrimination of the low and medium conditions. The authors reported that the neural network achieved a classification accuracy of 81% for the three levels of workload when trained on data from all participants and 98.9% accuracy when trained on a random participant. This result implies that the thermal changes are not consistent across participants [9].

Most researchers seem to find the nose temperature to be the most responsive to various stimuli such as loud noises, stress or workload. Dominguez *et al.*(2015) added other stimuli to the list. In their study, they explore the changes in facial temperature as the participants experience emotions such as pain, empathy or love. The emotions were induced using pictures from the International Affective Picture System and conclude that thermography is a valid index for such stimuli. They report an increase in nose temperature with positive valence and increasing arousal. In cases of empathic response, both positive (laughter) and negative (watching other people suffer), nose temperature decreased [49]. Moline *et al.* (2017) reported a drop in nose temperature while the participants were lying [50].

The facial thermography measure was not extensively used in previous research; one of the novelties of the research presented in this thesis is the continuous tracking of facial landmarks throughout the facial thermography recording. First of all, this does not introduce the requirement of the participant to be wearing markers [48] or glasses [8] and secondly it allows for the continuous tracking and temperature extraction from multiple areas around the face.

3.5 Other Physiological Measures

Many other physiological monitoring techniques have been examined in the field of measuring workload. Some of the most promising measures being currently researched fall into the category of brain sensing techniques and will be briefly presented here. The fact that the cognitive side of workload places demand on the brain offers these types of techniques great face validity and intuitive acceptance. Parasuraman and Mehta [2] have classified the various neuroergonomics measurement techniques in terms of the portability, cost, spatial and temporal resolution. fNIRs (functional near-infrared spectroscopy) stands out as having high portability, moderate spatial resolution but low temporal resolution while EEG (electroencephalography) has moderate portability, low spatial resolution but high temporal resolution. These two measures will be briefly presented as research results indicate they offer great insights into the measurement of workload.

fNIRS operates by measuring changes in oxygenated and deoxygenated haemoglobin [51] levels in the pre frontal cortex; the haemoglobin molecule which carries the majority of the oxygen in the blood absorbs different frequencies of the near-infrared light depending of its level of oxygen saturation, a property by which fNIRS measures the level of oxygenation. Multiple studies have shown that the change in blood oxygenation is significantly sensitive to changes in task load [[52], [53], [54], [55]]. Maior *et al.* (2015) have confirmed this finding using a verbal and a spatial task that fNIRS can distinguish between cognitive and rest states in both types of tasks [56].

While the fNIRS method requires near-infrared light to be sent through the skull in order to perform measurements, EEG (electroencephalography) relies on the voltage difference between an active electrode and a reference electrode; this is induced by the activity of large clusters of neurons. Derived from the EEG signal are the ERPs (event related potential) which represent the response to a specific stimulus. Both EEG and ERP have been used by researchers in the context of workload, in the same manner as fNIRS, it provides great face validity. Unlike fNIRS, the temporal resolution is high while the spatial resolution is lower. In their 1996 study, Brookings el al assessed the response of EEG together with eye measures, heart rate and respiration to various difficulty levels in an air traffic control task. They reported that changes in traffic manipulation, complexity or volume produced significant changes only in the EEG measures [31]. Wilson (2002) also found EEG activity to be sensitive to more cognitively demanding flight segments such as takeoffs and landings [57].

Both fNIRS and EEG methods show promising results; however they were not directly investigated in the research presented in this thesis as the aim was towards methods that are completely non-intrusive. Figure 3.1 presents three examples of EEG sensors of increasing spatial resolution (from left to right) while Figure 3.2 shows the fNIRS sensor used during the study presented in Chapter 7.

Fig. 3.1 Examples of EEG sensors. From left to right, the NeuroSky [58] (1 electrode sensor - low spatial resolution), the Emotiv [59] (14 electrode sensor - higher spacial resolution) and an 128 electrose EGI NetAmps 300 sensor providing high spatial resolution [60]

Fig. 3.2 Image of the fNIRS used during the study presented in Chapter 7

fNIRs will be mentioned in Chapter 7 as it was used during the study but it does not represent my contribution. The fNIRS measurements during the study were performed by Horia Maior, a researcher on fNIRS as a measurement technique for mental workload, at the University of Nottingham [56] and are presented as a comparison to the other measures.

3.6 Thesis Overview

Chapter 2 has presented the basic underlying concepts of workload and various means of measuring it. Chapter 3 has focused on the physiological measures of workload and the studies that explored them. Following the above presented literature review, the connection to the four main research questions of this thesis becomes clearer. While the first research question explores how human physiology responds to changes in task demand has been studied previously, there still are disagreements between study results. This aim provides an opportunity to further test the most commonly used measures (cardiac, respiratory and pupil diameter) and to explore the facial thermography response, especially analysing it separately on various areas of the face. The second research question looks at the relationship between physiological responses and subjective reports of mental workload, a comparison between two ways of measuring workload.

While most studies to date have examined how individual physiological measures respond to changes in mental demand, a few have looked at multiple physiological measures (e.g.: [30],[8]). The third research question explores the use of multiple physiological measures,

including facial thermography, which has not been used in many studies, to estimate the level of demand. It is expected that not all measures will respond in the same way to changes in demand, as different measures could show senstivity to different aspects of workload or may have a different bandwidth in which they are sensitive. Combining them could potentially lead the way to creating a more powerful measure of workload.

Mental workload, as a multidimensional construct and as described by the framework [3] presented in Chapter 2, is influenced by a multitude of factors such as physical and cognitive task demands, performance and external and internal influences. Skill, training level and experience fall in the category of internal influences. In this context, measuring the physiological response of the general public to workload leads to the next question. Would someone who has undergone extensive training in dealing with extremely demanding situations show a changes in the physiological measures that were used? This is the main reason for setting the fourth research question: how do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator?

The research presented in this thesis focuses primarily on assessing the physiological changes that occur as a response to variations in task demand and how these are associated with subjective reports of workload and how multiple physiological measures can be used to explain the changes in workload. The thesis also addresses the issue of whether highly trained individuals also respond in the same way to changes in workload. The main contribution lies in the area of assessing facial thermography as a non-invasive method for workload measurement as well as in the assessment of combined physiological measures for workload level estimation. The main contribution to the assessment of facial thermography was the more accurate facial landmark tracking performed in a non-invasive manner, allowing for the analysis of temperature variation in multiple areas and the use of a specially designed task for the laboratory studies, that allowed for a better control of the variation of demand and for accurately synchronizing the physiological data streams with the task performance and subjective ratings.

The next chapter will provide an overview of the sensors used in this research, the data they provide and the processing methods. Chapter 5 will introduce the first laboratory study aimed at examining the relationship between experienced workload and physiological responses of facial skin temperature, breathing rate, pupil diameter and heart measures. Chapter 6 discusses the results obtained in a study conducted in an ecologically valid helicopter simulator, examining the physiological responses of highly trained individuals to variations in task demand. The last study is presented in Chapter 7, expanding on the data collected in the first study, investigating also the response of fNIRS to the task used in the study as well as how this compares to the other measures used.

Chapter 4

Methodological Approach

This chapter aims at introducing the sensors that were used and the data processing techniques involved. The first part of the chapter describes the specifications of the sensors while the second part shows data samples as well the data processing techniques used.

4.1 Description of Sensors

Three studies will be presented in this thesis, all involving collection and processing of physiological data from human participants. The sensors that were used are presented below:

1. Zephyr BioHarness 3
2. RED 250 Eye Tracker
3. FLIR SC7000 Thermal Camera
4. FLIR A65sc Thermal Camera
5. fNIRS300

4.1.1 Zephyr BioHarness 3

The Zephyr BioHarness 3 [61] is a physiological monitoring module attached to a chest strap Fig.4.1, recording heart activity, breathing rate and acceleration. The areas marked with "HR Sensor locations" in the picture are conductive and have to be in contact with the skin; for best results it is recommended to have them moisturised before use. The breathing rate is obtained by a pressure sensor in the strap measuring the expansion of the rib cage. The user manual mentions that the breathing rates are accurate only during sedentary periods.

Table 4.1 lists the pieces of physiological data recorded by the Zephyr BioHarness 3 that were used for the studies:

Measure	Reporting Frequency (Hz)	Range	Units
ECG Amplitude (waveform)	1000 (in developer mode)	0.25 - 15	mV
Heart Rate	1	0-240	Beats/Minute
Breathing Rate	1	0-120	Breaths/Minute
Heart Rate R-R intervals	Per R detection	250-1500	ms (equivalent to 40 - 240 beats/minute)

Table 4.1 Physiological Parameters Measured by the Zephyre BioHarness 3

The accuracy of the heart rate and breathing rate are provided based on activity levels based on the United States Army Research Institute of Environmental Medicine guidelines Table 4.2.

Activity Level	Heart rate accuracy (beats/minute)	Breathing rate accuracy (breaths/minute)
Low activity (static)	2	3
Moderate activity (walk/jog)	3	3
High Activity (run)	3	5
Talking (breathing rate in range 6-25 bpm)	-	5

Table 4.2 Heart rate and breathing rate accuracies

Apart from the heart rate variability time domain measures presented above, research has shown that frequency domain measures of heart rate variability can be informative of mental demand [32]. Frequency domain measures require a high accuracy in R-wave detection (1-2 ms) [62] as well as a more complex filtering of the ECG signal. The recommendations for frequency domain measures include a time duration of the recording for estimating lower frequency components of the spectrum of approximately 2 minutes; due to the design of the study, demand was varied every 45 seconds and thus the analyses were performed on 45 seconds time intervals. Another recommendation is to use time domain analysis for recordings shorter than 24 hours as they are equivalent and easier to perform than the frequency domain analysis [63]. Due to these reasons, as well as the automatic R-R peak

Fig. 4.1 The Zephyr Bioharness 3 Sensor [61]

detection perfomed by the Zephyr BioHarness and returned in the form of R-R intervals, it was decided not to perform a frequency domain analysis. Although not considered for this thesis, a frequency domain analysis should however be explored on longer time intervals for the collected data.

4.1.2 RED 250 Eye Tracker

The RED (Remote Eyetracking Device) 250 [64], developed by Sensomotoric Instruments is a fixed infra-red illumination based eye tracker; it can be used both in monitor integrated and stand alone configurations. For the purposes of this research it was used in stand alone mode together with a television screen Fig.4.2.

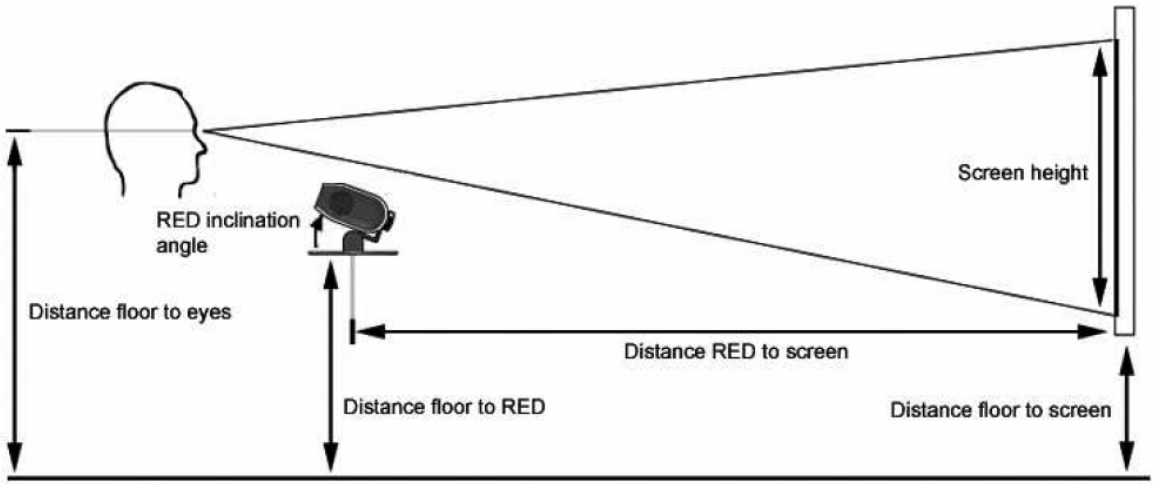

Fig. 4.2 The RED 250 standalone setup [64]

The RED 250 reports gaze position with an accuracy of 0.4°and pupil diameter at a rate of 60 Hz. The eye tracking module emits near-infrared radiation at a frequency of 870 nm. The calibration of the eye tracker and the data collection were done using the iView X™software. The post processing of the data was done using Matlab.

Blink frequency was not part of the raw data exported file by the iView X™software and was not analysed for this thesis. Blink frequency was shown to be a sensitive measure of mental workload in the driving context [65] and the data collected during this research should be further analysed to reveal blinking frequency and verify if it agrees with other research into mental workload.

4.1.3 Thermography

Thermograpy is a type of imaging achieved with thermal cameras that are sensitive to parts of the infrared radiation spectrum. These cameras capture the radiation emitted by objects in a specific range; the intensity of their emitted energy varies with temperature and wavelength and for objects cooler than 500°C, the emitted radiation lies completely within the IR wavelengths [66]. Depending on the detector material, the camera is sensitive to a certain part of the electromagnetic spectrum Fig.4.3. The camera detectors are of two types: thermal detectors and quantum detectors. One of the two cameras used for this research belongs to the MWIR (medium wavelength IR) and uses a quantum detector while the other to the LWIR (long wavelength IR) and uses a thermal detector.

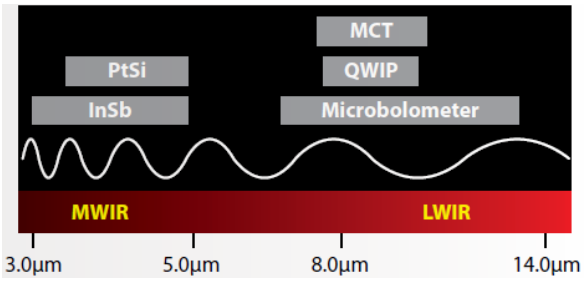

Fig. 4.3 Detector materials - spectral responses [66]

The quantum detector camera provides a higher accuracy but its sensor needs to be cooled, making it less mobile; such a camera was used for the laboratory studies. The microbolometer camera on the other hand is smaller and cheaper but does not provide the same signal to noise ratio.

FLIR SC7000 Thermal Camera

The FLIR SC7000 [67] is a cooled InSb detector thermal camera used for measuring temperatures of solid objects in a non-invasive manner. The main technical specifications of the camera are presented in Table 4.3.

IR resolution	640 x 512 pixels
Thermal sensitivity/NETD	<20 mK
Image Frequency	100 Hz
Spectral Range	1.5 - 5.1 μm

Table 4.3 FLIR SC7000 specifications

During the studies, the data was collected using the FLIR Altair software and later post processed using Matlab. The raw data saved using the FLIR Altair software consists of a digital level signal which has to be converted to temperature using a calibration curve. The calibration curve and reading of the data from the .ptw files was done using the FLIR Matlab SDK.

It has to be mentioned that this thermal camera was used together with the eye-tracker which uses near-infrared light for illumination. As the wavelength used by the eye-tracker is of 870 nm [64] and it is outside the 1.5-5.1 µm spectral range of the thermal camera, the measurements of the thermal camera were not influenced by the eye-tracker.

FLIR A65sc Thermal Camera

The FLIR A65sc [68] is a uncooled microbolometer type thermal camera used for measuring temperatures of solid objects in a non-invasive manner. The main technical specifications of the camera are presented in Table 4.4.

IR resolution	640 x 512 pixels
Thermal sensitivity/NETD	<0.05°C@ +30°C/ 50 mK
Field of View	45°x 37°
Focal Length	13 mm
Image Frequency	7.5 Hz
Spectral Range	7.5 - 13 μm

Table 4.4 FLIR A65sc specifications

During the studies, the data were collected using the FLIR ResearchIR software and later post processed using Matlab. The raw data saved using the FLIR ResearchIR software

consists of a digital signal S which has to be converted to temperature using the following equation:

$$T_{[K]} = \frac{B}{\ln(\frac{R}{S-O} + F)} \tag{4.1}$$

where ln is the base-e logarithm and S corresponds to the 14-bit pixel value. The R, B, F, O register values are generated by the camera.

4.1.4 fNIRS300

The fNIRS (Functional Near Infrared Spectroscopy) sensor uses blood oxygenation to determine the activation of areas in the brain, a higher blood flow indicating higher activation. fNIRS is based on the used of near infrared spectroscopy [51]. The headband consists of four infra-red emitters operating in the wavelength range of 700-900 nm, and ten infra-red detectors. The sensor can be seen in the bottom part of Fig. 4.4, with the four emitters in the middle row and two rows of five detectors each, one on the top part of the sensor and one on the bottom part. The top right hand side part of the same figure shows a person wearing the fNIRS sensor. During the study presented in this thesis, the fNIRS sensor was worn as shown in Figure 4.5, covered with an elastic band that would minimize the influence of outside infra red radiation on the sensor.

Fig. 4.4 The fNIRS 300 sensor [69]

Fig. 4.5 Image of the fNIRS sensor while being worn

4.2 Data Analysis

This section describes how the data recorded by the sensors was processed to extract the features that were later compared to subjective and performance results.

Fig.4.6 describes the way physiological data were reported. The top part of the figure shows a hypothetical physiological response in blue, as a time-series, while the red dots represent the instances ISA measures were collected (at the end of each sub-stage). The lower part of the figure shows a zoomed in area of the above, containing the interval between two ISA reporting times. For the study described in Chapter 7, the physiological data are reported as averaged values over the following time intervals: 45s, 30s and 15s prior to each ISA sampling point, as depicted in the figure while for the study described in Chapter 5 the data are reported as averaged over 45s time intervals

A 2 minute baseline physiological data consisting of heart rate and breathing rate was collected for each participant (during which they were asked to be still and relaxed); in the end this was not used and it was not reported. The reason for this is that in most cases the physiological signal was not stable and in the end the data was reported on participant by participant basis, without making comparisons between participants in terms of physiological data. When comparisons of physiological data were made between participants, it was in terms of number of standard deviations from the mean and not absoloute values. These two minutes, together with the time spent in the room while the task was explained represented

an acclimatization period from the point of view of the facial thermography measure; a facial thermography baseline was not recorded.

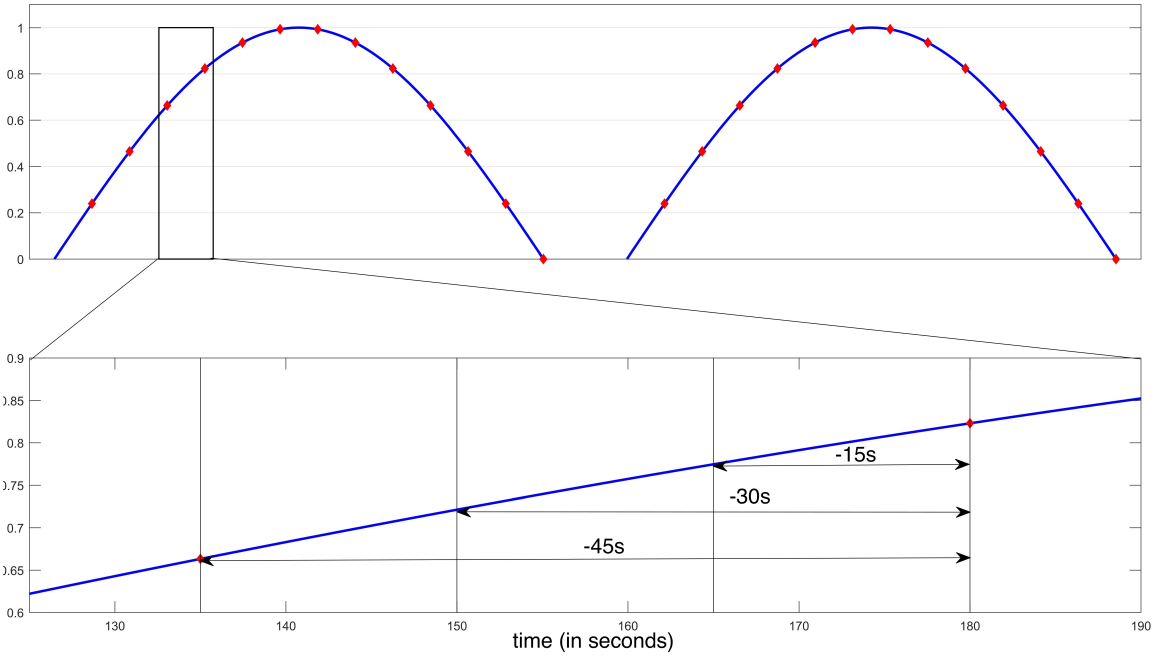

Fig. 4.6 Reporting of the physiological data

Fig.4.7 shows and example of how physiological data will be presented in the following chapters. In this case, the subjective data are presented in grey and the physiological data in blue (this figure shows the mean of the R-R time intervals computed over 45s time intervals leading up to the moment the subjective data were sampled).

4.2.1 Zephyr Data Analysis

The Zephyr sensor performs and ECG, a sample is shown in Fig.4.8 from which it computes parameters such as heart rate and heart rate variability; breathing rate is measured by a pressure sensor. The device also measures 3-axis acceleration but it was not used in this thesis. The Zephyr data presented further refers to the heart beat R-R intervals and breathing rate. As it can be seen in Fig.4.8, the R peak is the strongest and less influenced by noise, which is the reason why the R-R intervals are used as a parameter and also as a way of estimating heart rate. The Zephyr performs the R peak detection and generates the R-R data. The Zephyr data were downloaded after each session in form of .csv files which were read in Matlab and converted to time-series so that they can be synchronized with the other physiological data, subjective data and game performance. Figures 4.9, 4.10 and 4.11

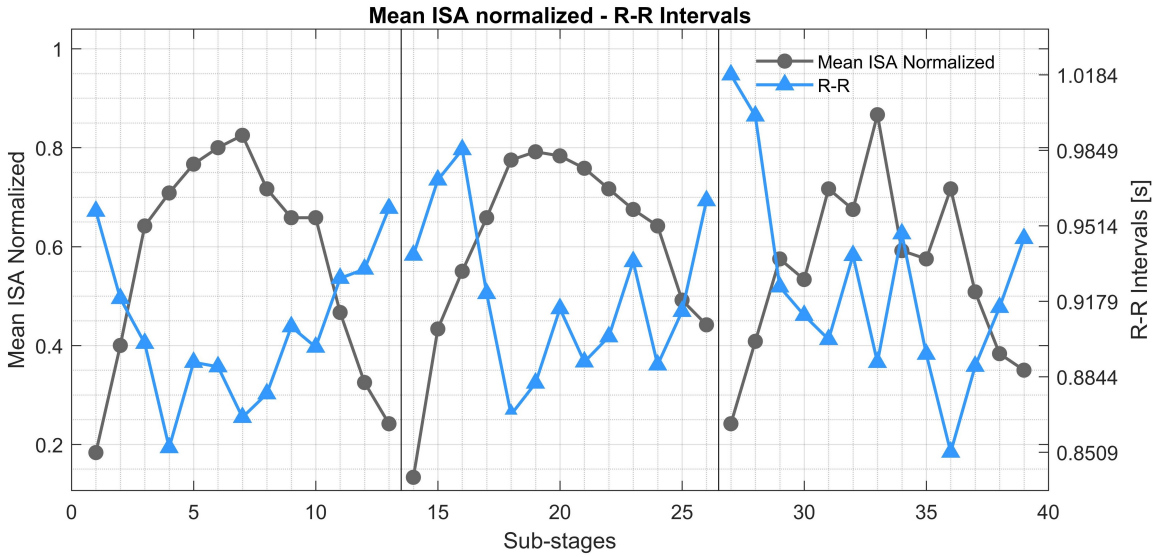

Fig. 4.7 Example of data presentation

show data samples for the R-R time intervals, heart rate and breathing rate. As there is a direct relationship between the R-R measure and the instantaneous heart rate, only the R-R measures was used. The data were synchronized with the task duration, represented in blue in these figures and also with the subjective data sampling times. Features were computed on the Zephyr data for the time intervals leading up to each subjective rating sampling time. The features were computed over 45 seconds time intervals for the first study (presented in Chapter 5) and over 45, 30 and 15 seconds time intervals for the study presented in Chapter 7. For the R-R data, the following time domain measures were derived [70], [71], [72], [73]:

1. mean R-R time: Mean inter-beat interval across the sample interval in milliseconds
2. SDNN: Standard deviation of the R-R data
3. SDSD: Standard deviation of successive differences of the R-R data
4. RMSSD: Root mean square of successive differences of the R-R data
5. NN50: Number of successive R-R differences that differ by more than 50 milliseconds
6. pNN50: Proportion of the NN50 from the total number of successive differences of the R-R data
7. NN20: Number of successive R-R differences that differ by more than 20 milliseconds
8. pNN20: Proportion of the NN20 from the total number of successive differences of the R-R data

9. SD1: The breadth of the Poincare plot of the R-R intervals across its identity line [74]

10. SD2: The length of the Poincare plot of the R-R intervals along its line of identity [74]

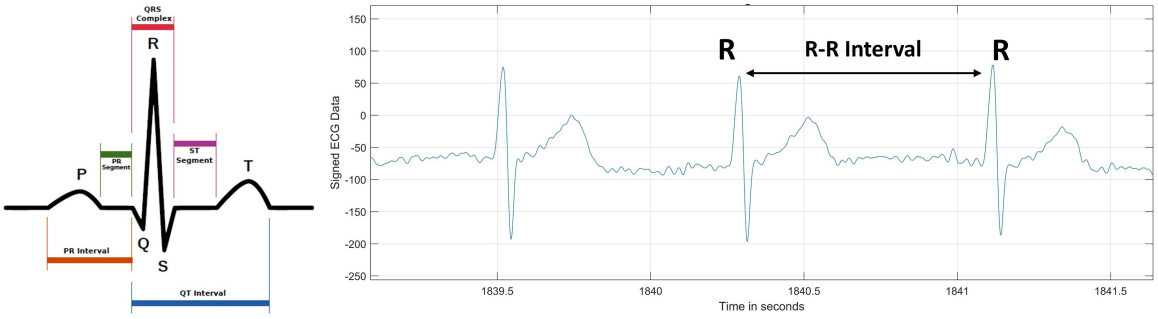

Fig. 4.8 QRS Complex (left) [72] and Zephyr ECG sample (right) recorded with the Zephyr BioHarness 3

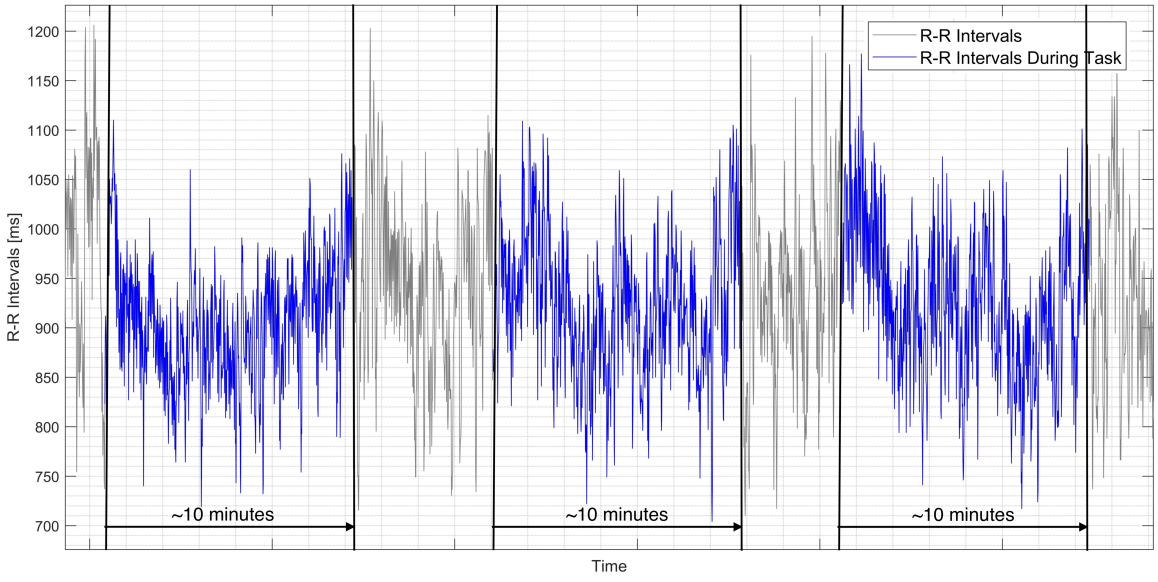

Fig. 4.9 R-R sample

4.2.2 RED 250 Eye Tracker Data Analysis

The RED 250 Eye tracker was used together with the iView X software. A calibration was performed before the start of each condition. The eye tracking data recorded by the software were exported to .csv format so that it could be read in Matlab and stored as a time-series together with all other pieces of data. For the purposes of the study, only pupil diameter

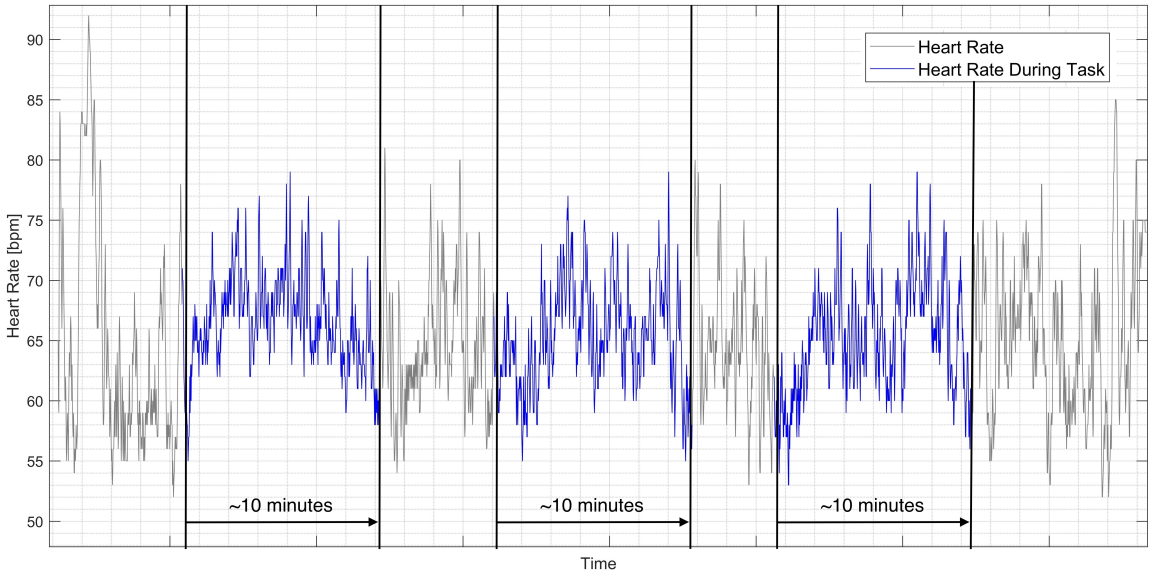

Fig. 4.10 Heart Rate sample

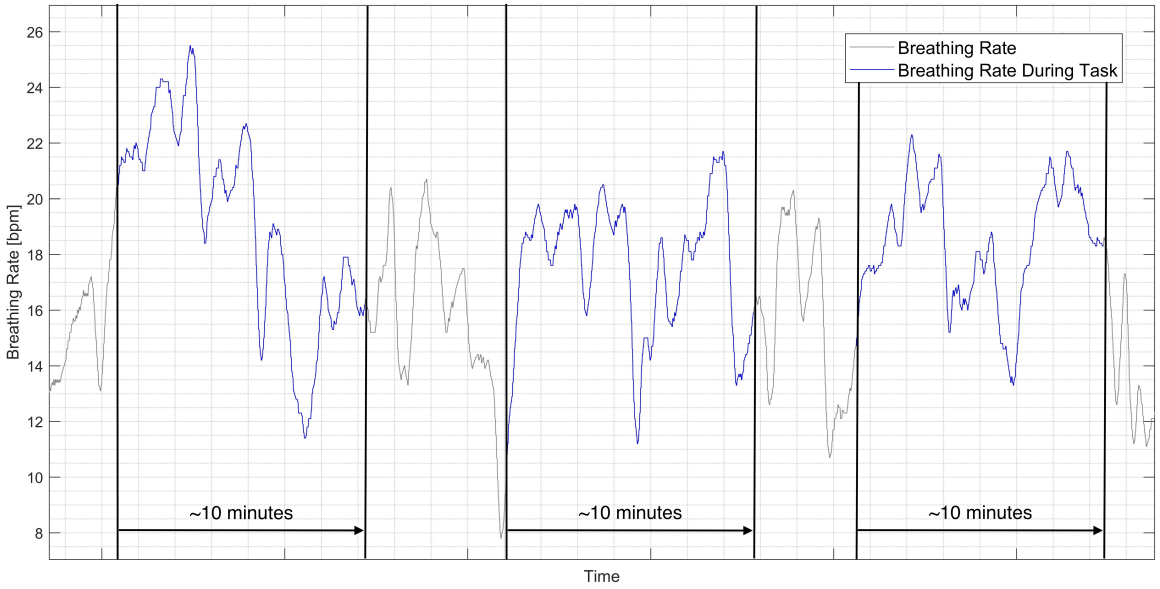

Fig. 4.11 Breathing Rate sample

was analysed. Besides pupil diameter, the eye tracker also generates confidence data, taking integer values from 0 to 3. Fig.4.12 shows an example of pupil diameter data, including the low confidence data. Fig.4.13 shows the pupil diameter data after removing the low confidence measurements. In order to generate features, the data will be averaged over 45 seconds times intervals before the subjective ratings were sampled.

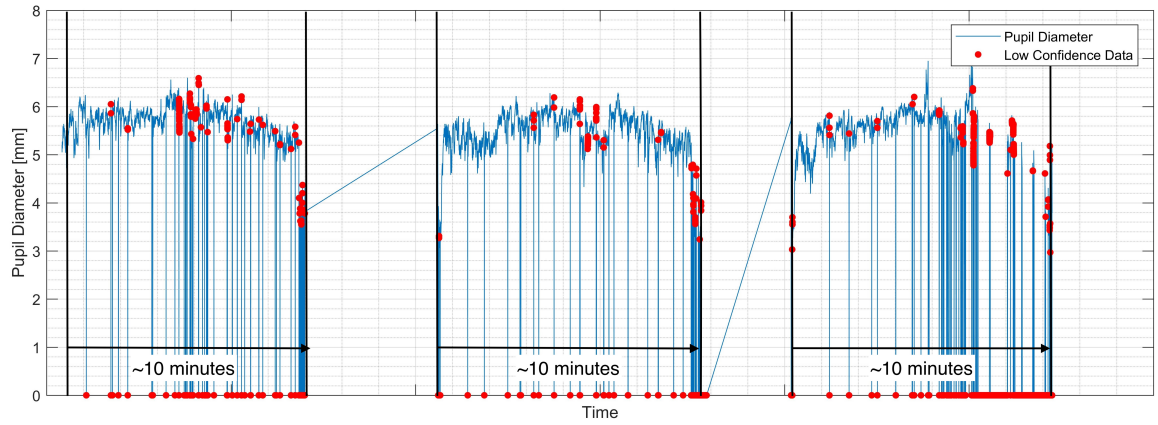

Fig. 4.12 Pupil diameter data and confidence

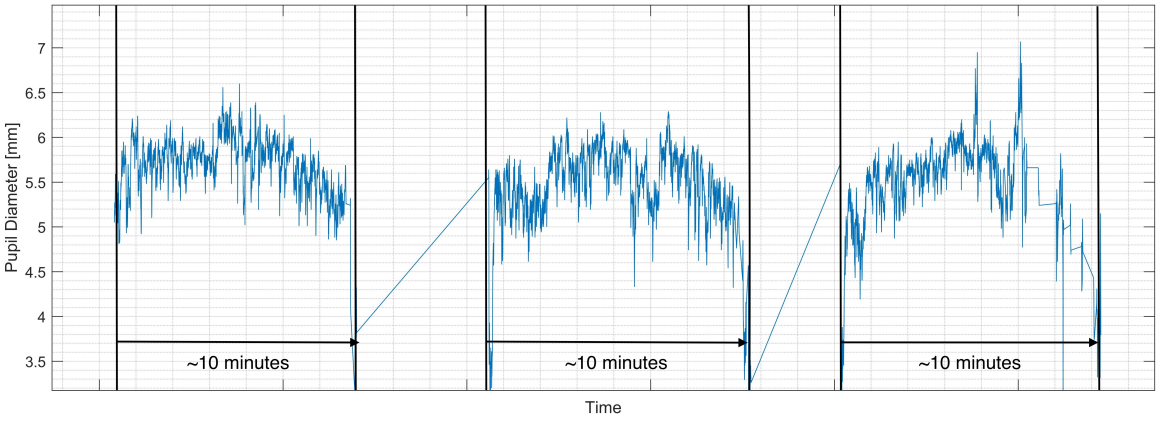

Fig. 4.13 Pupil diameter after removal of low confidence data

4.2.3 FLIR Thermal Data Analysis

One of the challenges faced during this research was extracting temperature data from the thermal images without placing markers on the faces of the participants and at the same time allowing the participants to freely move their heads. This was achieved by using two different approaches. First approach was used for the first study and it involved tracking of facial landmarks by mainly finding edges and is presented in the subjection Landmark tracking

FLIR SC7000. The second approach works better with the noisier images acquired with the microbolometer camera and was used for both the second and third studies presented in this thesis. It is based on the Robust Cascaded Pose Regression (RCPR) algorithm of Burgos et al. [75] and is presented in subsection RCPR based landmark tracking for FLIR SC7000 and A65sc.

Landmark tracking FLIR SC7000

This approach in tracking the facial landmarks was used only for the first study and it was largely improvised by using several techniques that will briefly be described here. This was most likely not the most efficient way of accomplishing the task, which is why it was not used for the later studies.

The frame rate of the camera was set at 50 Hz, resulting in about 90000 frames for each of the participants. The FLIR software records the data in a .ptw file which was read using the Matlab FLIR SDK. The .ptw files recorded for each of the participants were each around 50 GB in size.

The .ptw file stores each frame in a Digital Level format which needs to be converted to temperature by using a calibration curve. As the conversion to temperature was time consuming, the Digital Level data were used for the image processing and only the data from the regions of interest were converted to temperature. So, at this stage we have a one dimensional matrix of digital level data, which for the purposes of image processing was converted to a 3 dimensional RGB matrix using a colormap with 16384 levels.

Once the RGB image is generated (Fig.4.14), the first step is to detect the eyes. The entire algorithm relies on the successful detection of the pupils inside the eye regions. The eyes are detected by using the Viola-Jones algorithm. In order to train an eye classifier, a set of positive labelled images were generated for each participant together with a set of negative images. The images were selected every 550 frames from the original video. In order to increase the data set and make it more robust to head rotations, for each image two more were created having a random rotation between -20 and 20 degrees. After removing the pictures where the participant's eyes were not clearly visible due to head rotation, a set of rotated and not rotated pictures were used for labeling. In the case of the participant presented in this example, 294 positive images were manually labeled with eye regions using the Matlab Image Labeler. The Viola-Jones algorithm also requires a larger set of negative images, in this case images that do not contain eye instances. As there was no available set of similar thermal images that do not contain eyes, the positive samples used before had the eye areas filled with black colour and then randomly rotated 50 times each, at angles between -20 and 20 degrees in order to generate a large set of negative images. The final created set

of negative images contained in the case of the participant presented here 6154 instances of thermal images with no eyes. The positive and negative image sets were used to train a cascade classifier using the Matlab function 'trainCascadeObjectDetector' with a False alarm rate of 0.2, 15 stages and Haar feature type.

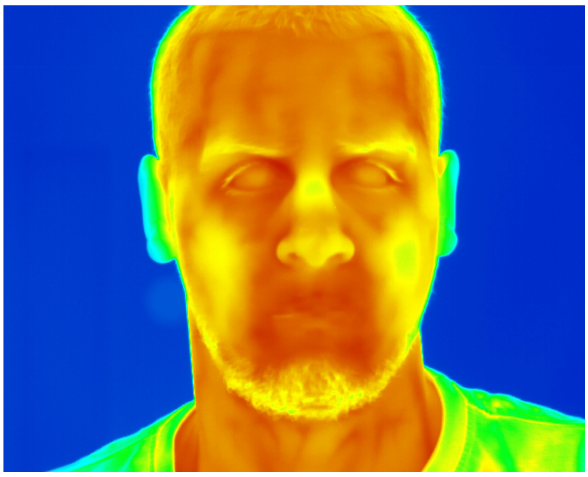

Fig. 4.14 RGB Thermal Image

The generated object detector is then used to find the eyes in the images. To increase the speed and minimize the chance of a false detection, the search was performed in a smaller area of the image, as shown in Fig.4.15. The area was delimited using colour thresholding and setting some rough face proportions for each participant.

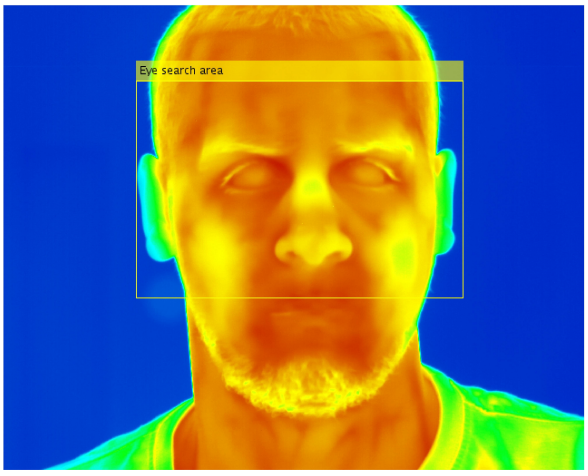

Fig. 4.15 Eye search area

Fig.4.16 shows a successful instance of eye detection, eyes were classified into right and left by the X coordinate in the image.

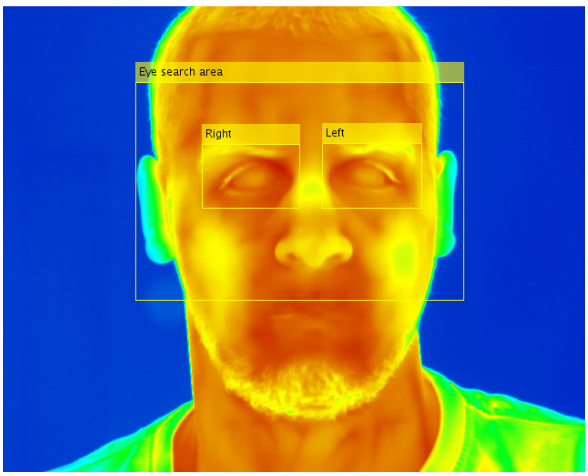

Fig. 4.16 Successful eye detection example

The right and left bounding boxes around the eyes are used as a search area for the most likely candidate of the pupil. In order to narrow the area more, a 30 by 26 pixels area around the centre of each of the eye bounding boxes was used. Looking at this area in terms of thermal data, it is obvious that the pupil is colder than the surroundings, the left side of Fig.4.17 shows a 3D plot of the area containing the pupil. The final pupil position (green circle) is chosen to be the closest local minima to the centre of the eye bounding box (black circle). Even though it does not always offer a perfect result, it is better than just the centre of the eye bounding box. The local minima function was written by [76]. The detection of the pupils also provides information with regards to the face tilting angle that is used later for the detection of the other points.

Points F and H in Fig.4.17 a quarter of the way from each pupil to the other one. The next step fixing some stable points on the forehead. The approach was to find the point on the eyebrow directly above the pupil. In order to do this, a perpendicular 'search line' was extended from the pupil upwards, perpendicular on the line joining the pupils. This 'search line' would intersect the eyebrow; as that would be a colder area, the point on the eyebrow was fixed as where the minimum temperature would occur along the search line. Fig.4.18 shows the temperature profile along the search line together with the minima point representing the intersection of the eyebrow.

To sum up, the landmarks that were fixed so far are presented in Fig.4.19. Points A and C on the forehead were computed as a function of the stable point on the eyebrow. The main purpose of points A and C is to extract the temperature profile along the line AC on the forehead.

The next landmarks to be detected are the points around the nose. This was accomplished by extending a 'search line', perpendicular on the line that joins the pupils, from the middle

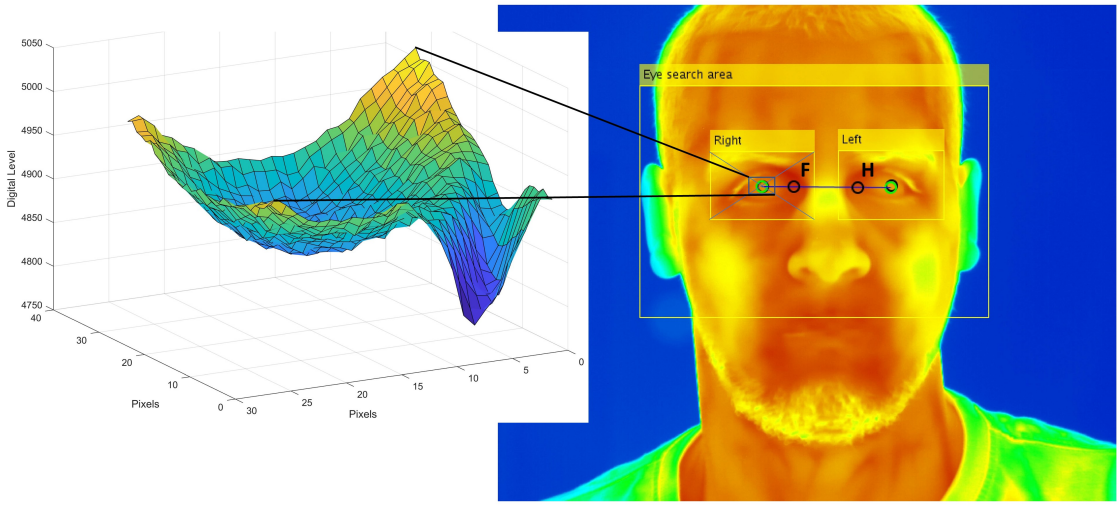

Fig. 4.17 Pupil detection example

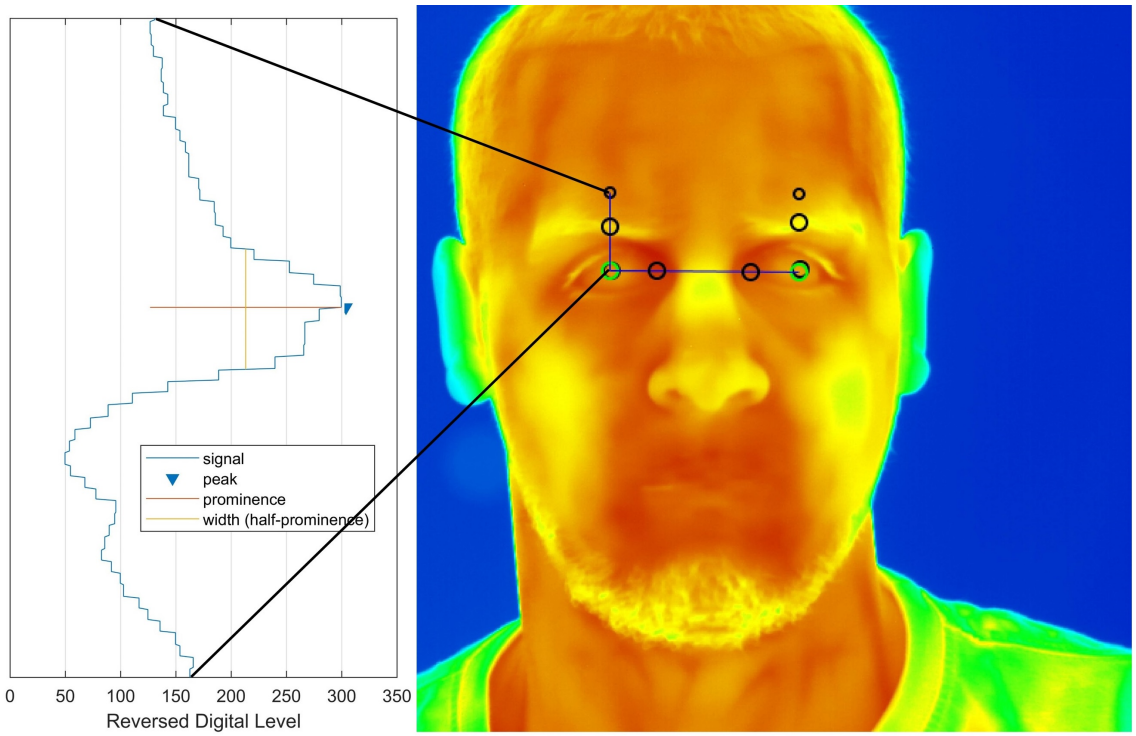

Fig. 4.18 Eyebrow detection example

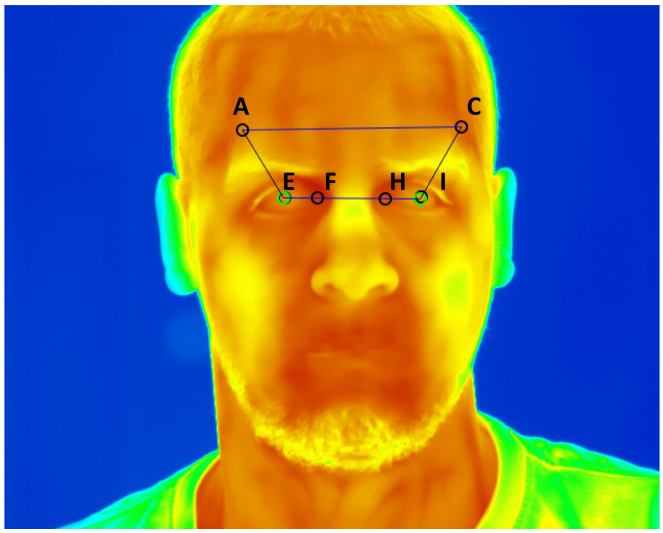

Fig. 4.19 Eye and Forehead Landmarks

of it. The length of the search line was customized for each participant depending on the face proportions. A nose bounding box was build around the line and established as a search area. Fig.4.20 shows the nose bounding box and the conversion to an intensity image on which the Prewitt edge detection method is used. The detected edges have been drawn on the thermal image along with the points on either side of the nose, below nose and the nose tip. These are not perfectly aligned with the face orientation but will later be adjusted.

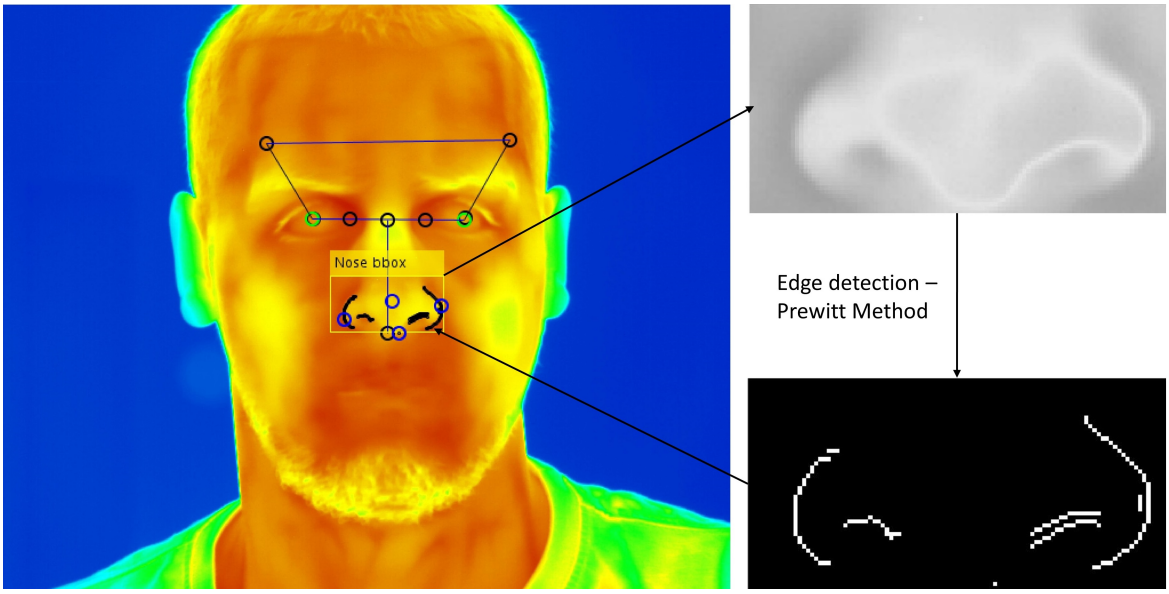

Fig. 4.20 Nose edge detection

The left side of Fig.4.21 describes how some of the side face landmarks were detected. From point 1, representing the middle from the nose tip to the line joining the pupils, 'search lines' were extended outwards on both sides. Line 1-2 on the right side helps in first detecting the ear edge (3) in a similar way to finding the eyebrow, based on the a peak in a thermal signal along line 1-2. Once this point was found, line 3-4 was extended inwards, this time point K on the edge was detected by analysing a peak along the 3-4 line in a binary image generated by using colour thresholding that removed the ears and keeping only the warmer areas Fig.4.22 right. The same approach was used for points D, K, O, J, N and Q shown in the left side of Fig.4.21. Points L and M represent the intersection between the line KN and the lines passing from the nose tip through either of the pupils. The parameters of the colour thresholding mask and peak detection along the lines had to be adjusted for each of the participants which makes this method harder to generalize without making adjustments. Once the parameters were set, the detection was allowed to run for all the images in the recording.

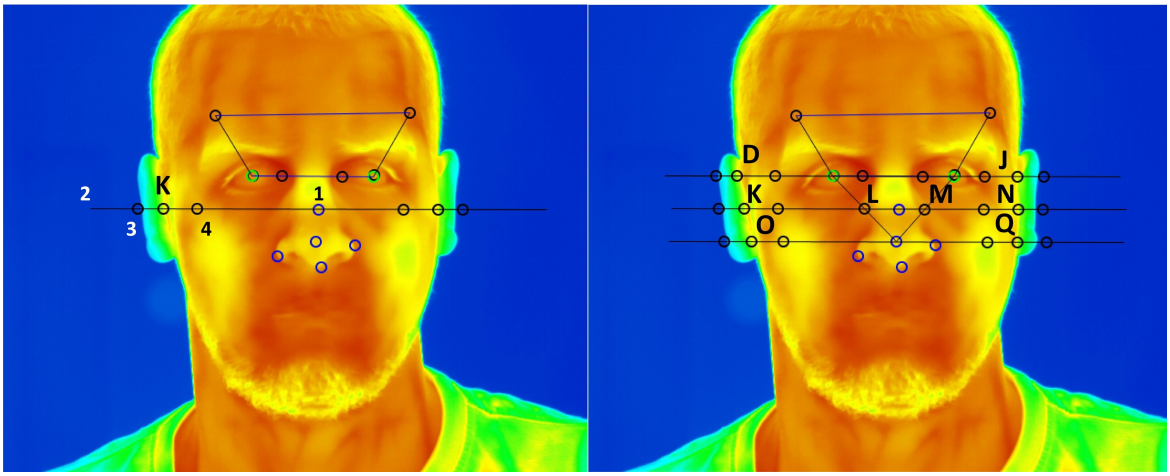

Fig. 4.21 Face side edges detection

The left side of Fig.4.22 shows the method used for detecting the edges in points R and U. For point R, the line passing through points N and P was extended to point 1 such as to be sure it goes beyond the edge of the face. As for point K described above, point R was detected based on the profile of the line crossing the edge of the face, both in terms of thermal data as well as using the mask presented in the right side of Fig.4.22 by using colour thresholding. For point U, the search line was extended from point R and is parallel to the line joining the pupils. The point of intersection with the edge of the face was computed in the same manner.

Fig.4.23 shows the final result of all the facial landmarks drawn on the thermal image. Points S and T were updated, S is the intersection between lines FL and RP, T is the

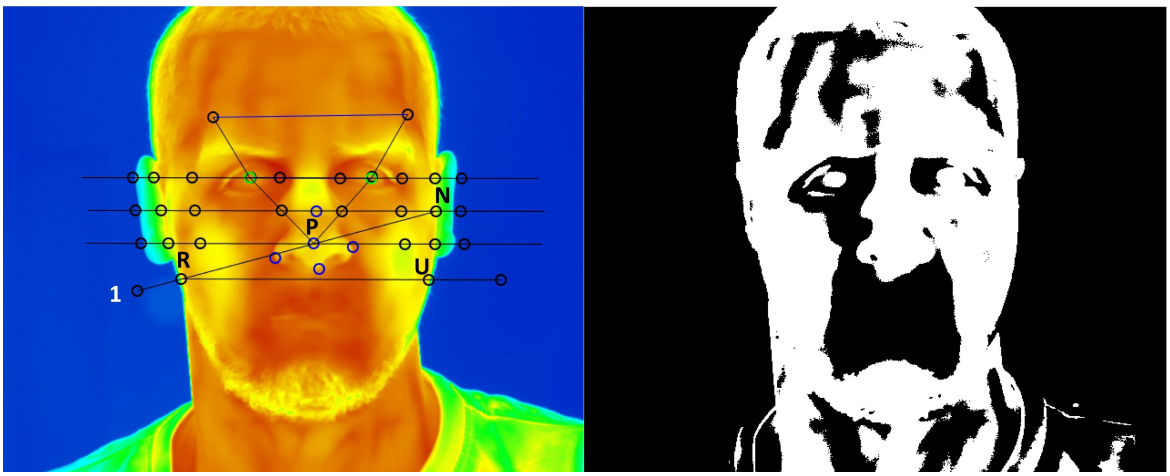

Fig. 4.22 Side edges detection and face mask

intersection between lines HM and UP; Y represents the intersection point between lines KP and FL while Z is the intersection between PN and HM. Point V has been updated as well as the point where the perpendicular line from P onto line RU intersects RU.

For each frame, the thermal data in form of Digital Level would be extracted from multiple regions of interest and converted to degrees Celsius according with the calibration curve, an example of which is shown in Fig.4.25. After the temperatures have been extracted, the data would be stored in a structure containing also the time the frame was taken. Temperature data were extracted from along lines, inside polygons and inside circles. The reason for extracting temperature from along the lines was so that peak temperatures would be examined, as a potentially more accurate way compared to extracting data from points, which might turn out not to be so accurately detected. An example of data along the line KN for the frame presented above can be seen in Fig.4.24 showing also part where it intersects the colder area of the nose. The following lines were used for temperature extraction: EI, AC, DA, AB, BC, FB, BP, BH, CJ, KE, OE, EL, IM, EG, GI, FL, GL, GM, HM, QI, NI, KX, XN, KL, KY, LP, MP, NM, NZ, GP, PV, RV, RS, UV, UT, KN, OQ, RU.

Average temperatures were also extracted from the following polygons (defined by their vertices) on the right side of the face: KEL, EFL, FGL, LGX, KLSROK, LXP, LPS, RSV, SPU, YPS, KLY, FBG and the symetrical ones on the left side of the face: MIN, HIM, GHM, GMX, MNQUTM, XMP, MPT, TVU, PTU, ZPT, NMZ, BGH.

Average temperature was extracted from points: A, B, C, E, F, G, H, I, L, M, P, S, T, V on a radius of 5 pixels (the radius of the circles presented in Fig.4.23). Other landmarks in the figure are represented by a single point, not a circle (e.g. D, K, O, R); temperature was not extracted from these areas as they were too close to the edge of the face and small errors

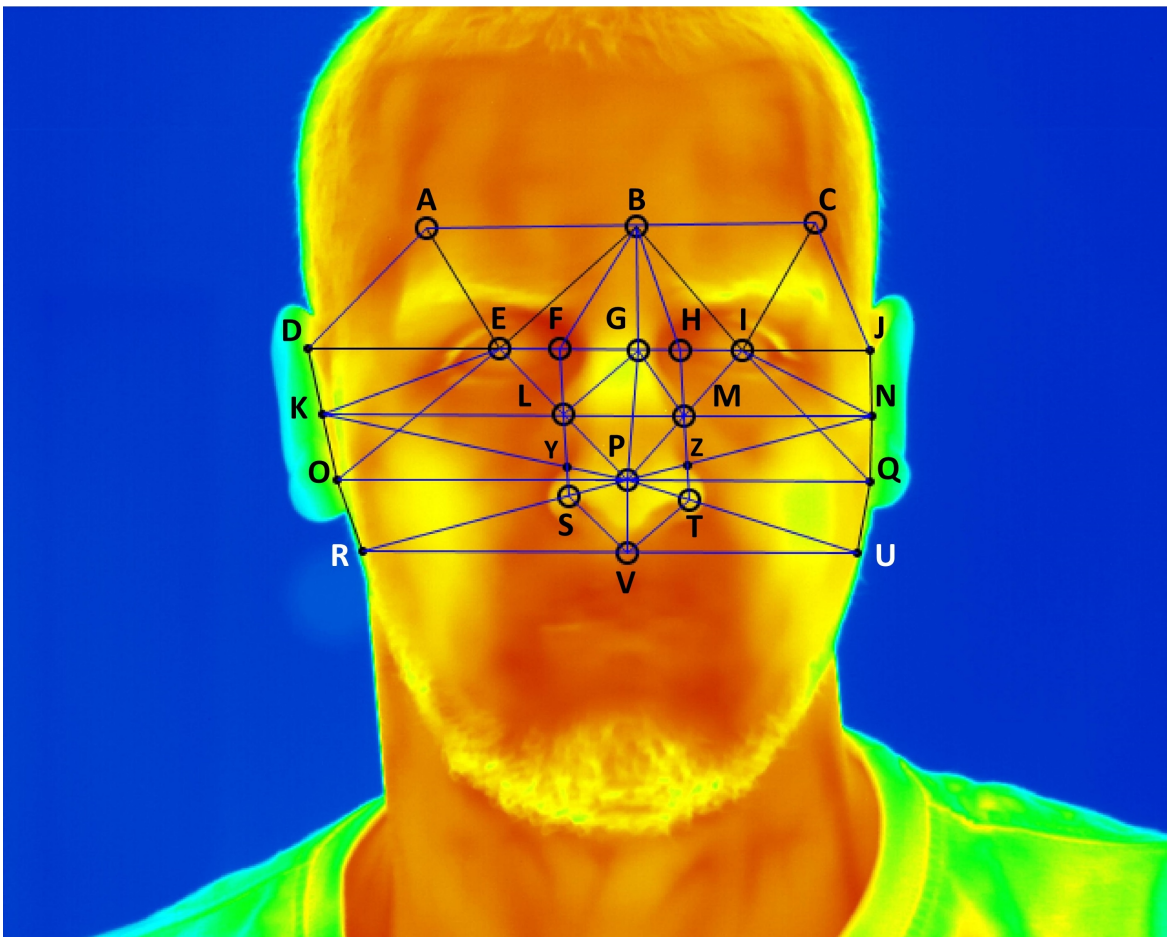

Fig. 4.23 Facial landmarks

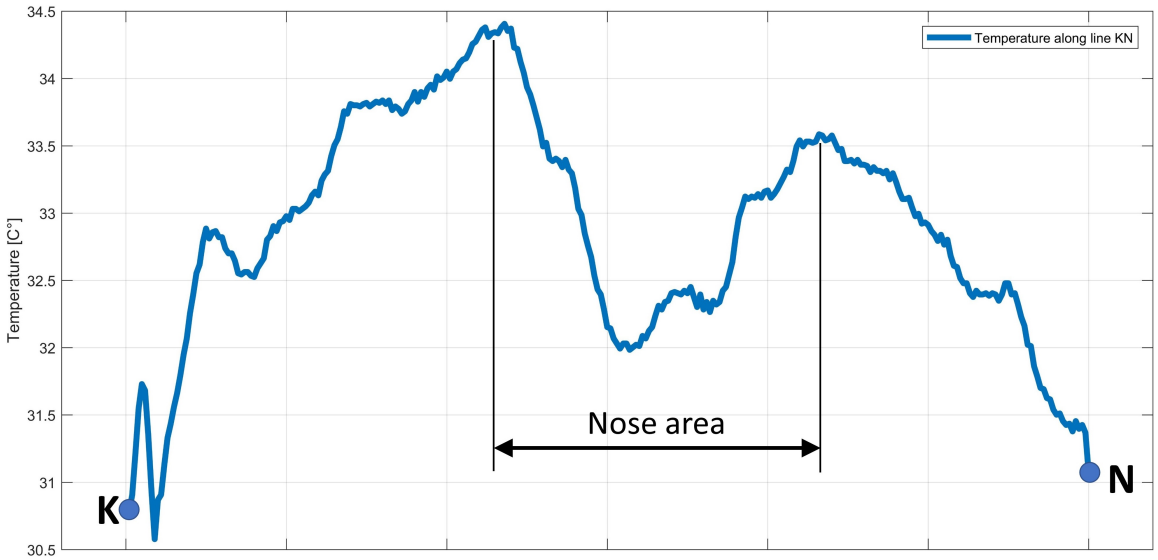

Fig. 4.24 Temperature along line KN

in tracking would influence the average temperature in the circle too much if for example the circle samples temperatures from the much colder background.

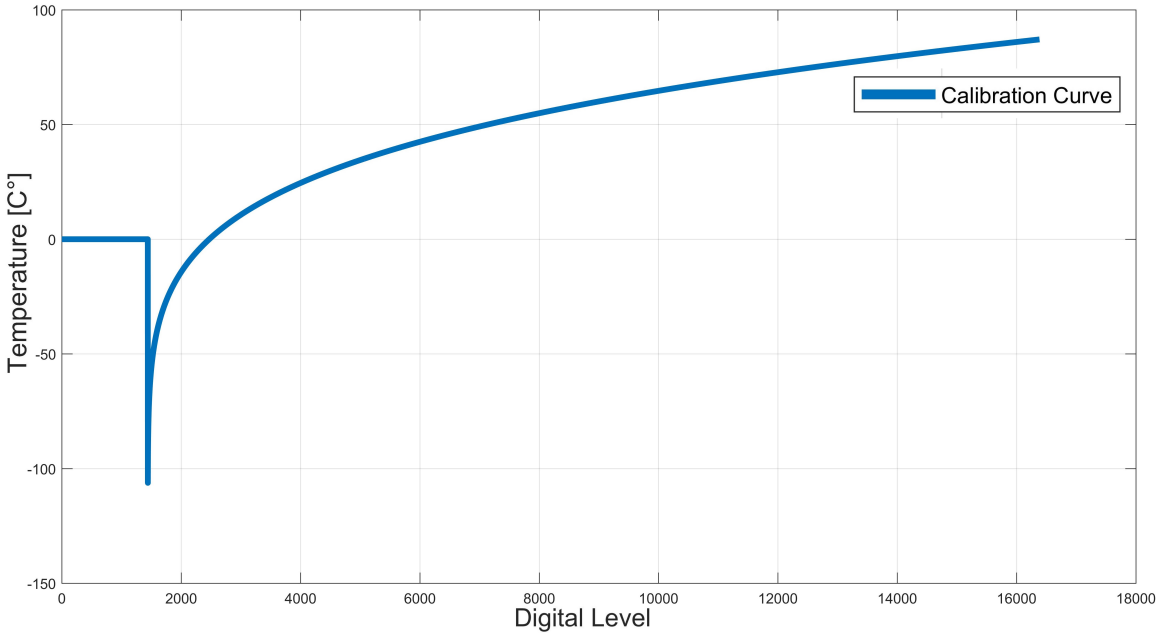

Fig. 4.25 Calibration curve example

RCPR based landmark tracking for FLIR SC7000 and A65sc

The approach in landmark tracking described above proved not to be very easy to use, as parameters had to be adjusted as well as eye classifiers had to be trained for each participant individually. Besides these disadvantages it also turned out that the method was not robust enough when using the more noisy images captured by the FLIR A65sc microbolometer camera in the real world environment (the Aberdeen helicopter flight simulator).

The Robust Cascaded Pose Regression (RCPR) algorithm of Burgos et al. [75] was used for both the Aberdeen study (images were captured using the FLIR A65sc) as well as for the third study. Even though the FLIR SC7000 thermal camera that was used in the first study was also used in the third, it was decided that the use of the RCPR algorithm would improve temperature extraction time and accuracy.

The RCPR code, made available by the authors of the paper [75] came with a few already trained models for tracking facial landmarks in visual images. One of the encountered challenges was that when the models were applied to the thermal images, the accuracy would suffer. Thus, the first step in the extraction of temperatures using this method was retraining the models with labeled thermal images. For each of the cameras (FLIR A65sc and FLIR

SC7000) two models were trained, one for the participants wearing glasses and one for the participants not wearing glasses. As glass is not transparent to the wavelength range that the cameras are sensitive to, wearing glasses would obstruct features around the eyes (one more reason for which the algorithm presented above and relying on first detecting the pupils would not work in this case).

In order to get a set of labeled thermal images, a point selection tool with a simple interface was built Fig.4.26. About 50 thermal images were selected from each participant, roughly uniformly distributed throughout the video, showing the face in various poses but at reasonable enough rotation angles so that the main features are still visible.

Fig.4.26 shows two consecutive screen-shots from the point selection tool. Initially the user is presented with a thermal image and a number of points grouped into six clusters: face perimeter points, eyebrow points (two clusters), eye points (two clusters) and nose points. To increase the speed, one cluster of points can be grabbed and moved with the mouse cursor into an approximate position on the face and then, individual points within the cluster can be adjusted. The right side of the image shows the all points moved in their location. Once the user has arranged the points, the Forward or Back buttons can be used to navigate through the selected thermal images. When pressing the Forward button, the previously point structure is maintained as in most cases it will be a picture of the same participant maintaining the same facial proportions, helping to speed up the process. There is also a Save button and an Exit one. The images present the particular case of the Aberdeen study for participants not wearing glasses; in this case 388 images were labeled and used to train the non-glasses model. For the glasses model, 71 extra pictures of participants wearing glasses were added.

After training the model and running the landmark detection, the results are presented as in Fig. 4.27 for the Aberdeen study case. The figure also shows the labels used for the facial landmarks. The right side of the figure shows an example of a participant wearing glasses. The points, lines and areas obstructed by the glasses were ignored as the temperature readings would have been highly influenced. The yellow bounding box around the faces is needed by the RCPR algorithm as a search area for the landmarks and it was computed by means of colour thresholding.

A similar process was applied in the case of the fNIRS laboratory study presented in the thesis. 432 pictures of participants not wearing glasses were labeled for training the model. Fig.4.28 shows the landmarks that were tracked in this case while. Table 4.5 presents the points, lines and polygonal areas from which temperature was extracted as average. The left side of the table presents the interest areas for the Aberdeen study whereas the left side of the table presents the interest areas for the fNIRS study. As mentioned above, some interest areas had to be ignored because the participants were wearing glasses. In these cases a subset

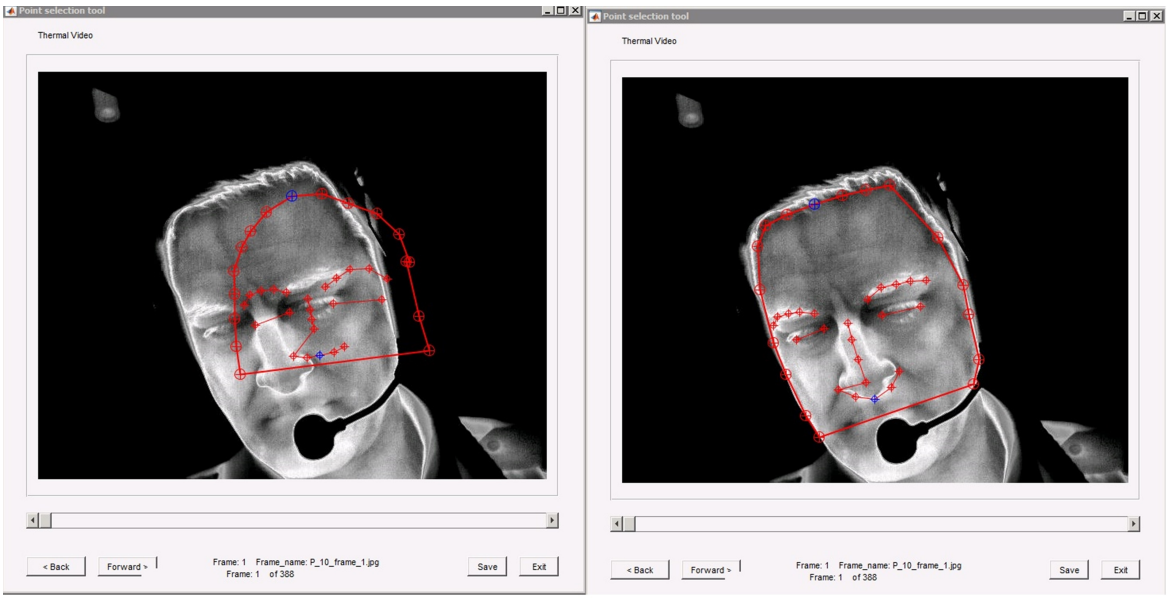

Fig. 4.26 Point selection tool screen-shots

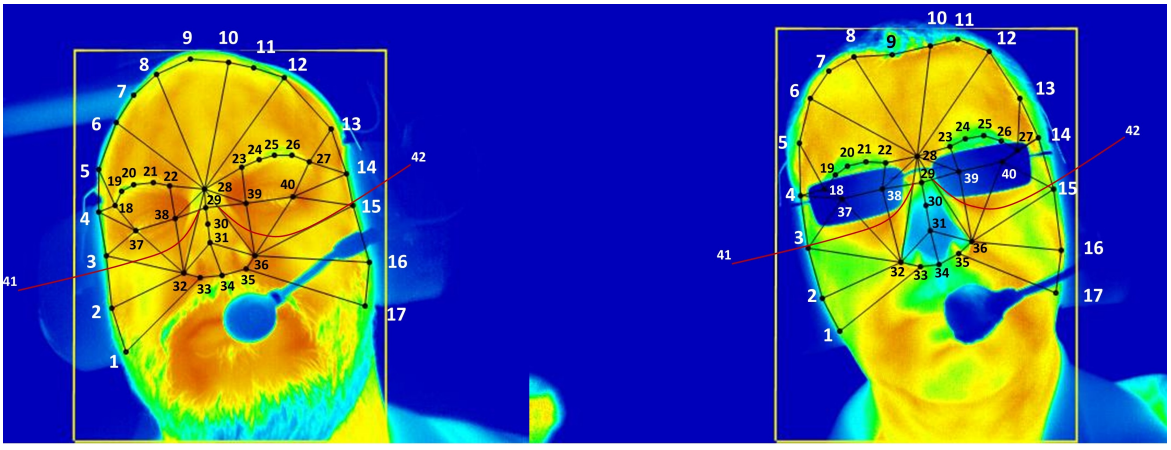

Fig. 4.27 Landmarks results and labeling for both non-glasses and glasses for the Aberdeen study

of the areas were used, namely the ones that are bolded in the table. The notations in Table 4.5 correspond to the labeled points in Fig. 4.27 for the right side of the table and Fig.4.28 for the left side of the table.

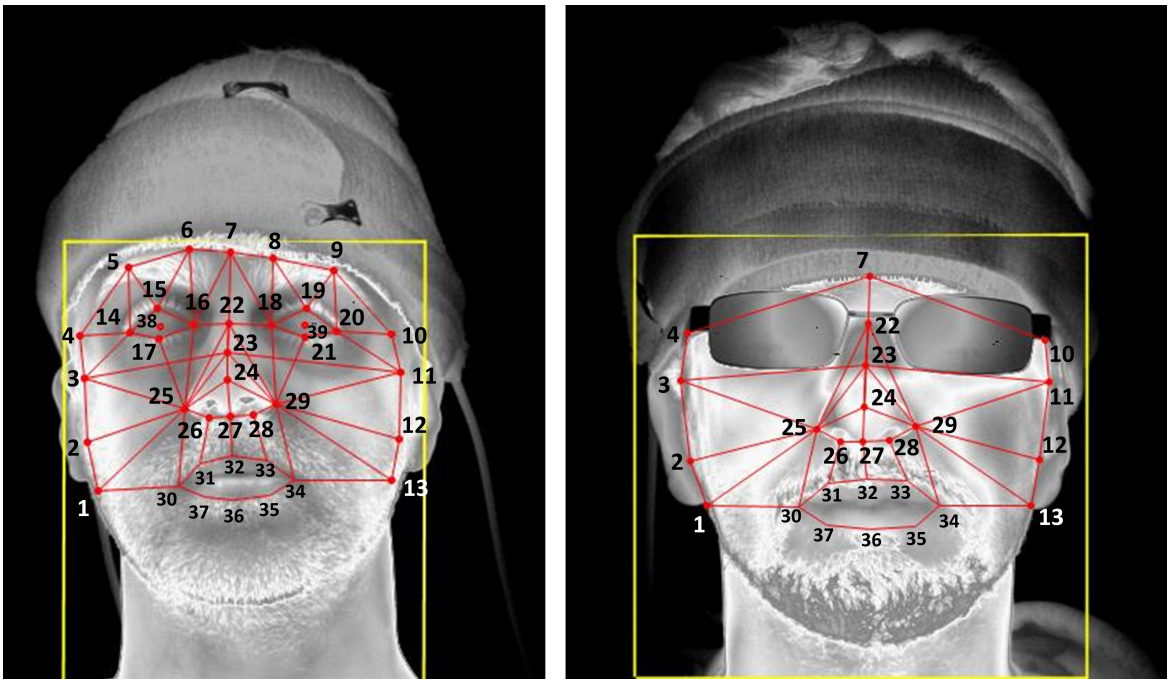

Fig. 4.28 Landmarks results and labeling for both non-glasses and glasses for the third study

The output of the tracking process consists of data time-series for each of the entries in Table 4.5 and for each participant. The next step is to extract features from the data. Due to the tracking algorithm failing to identify the right areas in various frames or due to small variations in the tracking, the time-series data will contain both outliers (due to large errors in landmark tracking) and noise (due to small variations from frame to frame).

An example of feature extraction for Study 3 presented in this thesis is given below Fig.4.29. First, the time data generated by the task are used to split the thermal data time-series into three parts (represented in blue), as there were three conditions. This generally removes erroneous data caused by ample head movements after the end of the condition (e.g. red ellipse), when the participant was no longer facing forward. After this, a moving average filter was applied on the data using the rloess method on 3% of the data which in this case amounts to about 15 seconds (represented in red).

Is is not always the case that the tracking is so accurate and the data has so little noise. An example is Fig.4.30 shows the mean temperature across line 2-25 for the same participant.

The now filtered data (in red) were used to generate the features. As subjective ratings were obtained every 45 seconds, the thermal features consisted in the average value of the

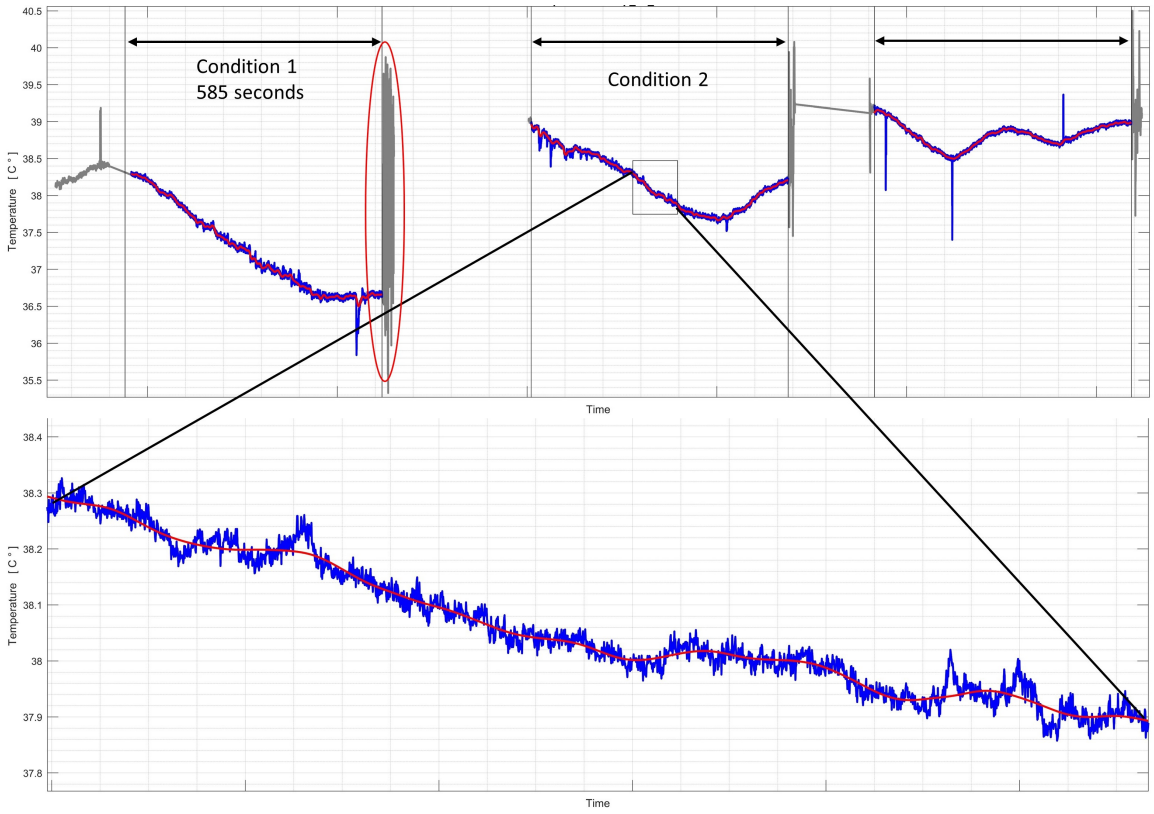

Fig. 4.29 Mean temperature for line 22-23

temperature in one of the interest areas over the 45 seconds time interval before the subjective rating was obtained. For the third study the data were averaged also over the last 30 and 15 seconds before the subjective rating.

4.3 Chapter Summary

This chapter has offered an overview of sensors used during this research as well as the data processing steps leading up to the generation of features that will later be used in the comparisons with the subjective or performance data. The next chapter will introduce the first laboratory study aimed at examining the relationship between experienced workload and physiological responses of facial skin temperature, breathing rate, pupil diameter and heart measures.

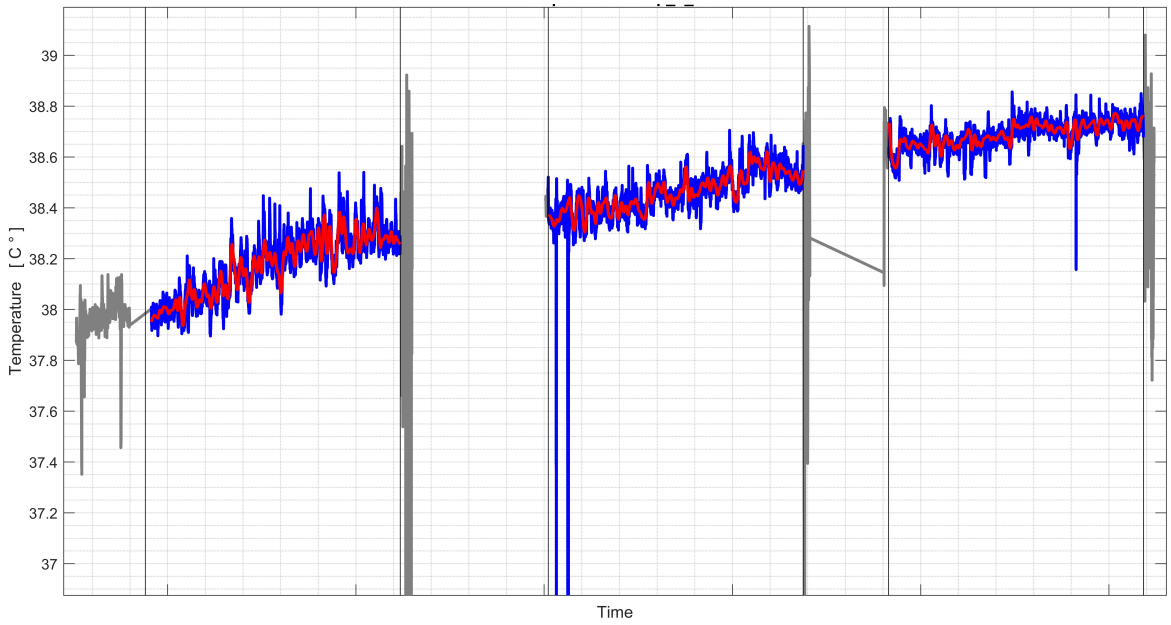

Fig. 4.30 Mean temperature for line 2-25

Aberdeen Study		
Points	Lines	Polygons
28	**1-33**	**1-2-32-33**
29	**2-32**	**2-3-32**
30	**3-32**	3-37-32
31	3-37	3-4-37
32	4-37	**6-7-8-28**
33	**8-28**	**8-9-10-28**
34	**9-28**	**10-11-12-28**
35	**10-28**	14-15-40
36	14-40	15-40-36
37	**15-36**	**15-16-36**
38	15-40	**16-17-35-36**
39	**16-36**	18-19-20-21-22-38-37
40	**17-35**	23-24-25-26-27-40-39
	22-23	37-38-32
	28-29	39-40-36
	29-30	22-28-38
	30-31	23-28-39
	31-34	28-29-38
	33-34	28-29-39
	34-35	38-41-32
		39-42-36
		28-29-30-31-32
		28-29-30-31-36
		31-32-33-34
		31-36-35-34

fNIRS Study		
Points	Lines	Polygons
16	**1-30**	**1-25-30**
18	**1-25**	**1-2-25**
22	**2-25**	**2-3-25**
23	**3-25**	3-14-17-25
24	7-16	16-17-25
25	7-18	14-15-16-17
26	**7-22**	6-7-16
27	**11-29**	7-16-22
28	**12-29**	16-22-25
29	**13-29**	**22-23-25**
32	**13-34**	**23-24-25**
36	15-17	**24-25-26-27**
38	16-22	**25-26-31-30**
39	16-25	**26-27-32-31**
	17-25	**13-29-34**
	18-22	**12-13-29**
	18-29	**11-12-29**
	19-21	11-20-21-29
	21-29	18-21-29
	22-23	18-19-20-21
	22-25	7-8-18
	22-29	7-18-22
	22-24	18-22-29
	23-24	**22-23-29**
	23-25	**23-24-29**
	23-29	**24-29-28-27**
	24-25	**28-29-34-33**
	24-29	**27-28-33-32**
	24-27	**22-25-24-29**
	26-31	**24-25-26-27-28-29**
	27-32	**25-26-27-28-29-34-33-32-31-30**
	28-33	
	30-31	
	31-32	
	32-33	
	33-34	
	34-35	
	35-36	
	36-37	
	37-30	

Table 4.5 Temperature extraction regions of interest for the Aberdeen study (left) and the fNIRS study (right)

Chapter 5

Physiological Response to Variation of Workload Demand

5.1 Introduction

The aim of this chapter is to examine the relationship between experienced workload and physiological response by non-invasive monitoring of physiological parameters including facial skin temperature, breathing rate, pupil diameter and heart rate variability. Most studies to date have examined how individual physiological measures respond to changes in mental demand while a few have looked at multiple physiological measures (e.g.: [30],[8]). This study explores the response of multiple physiological parameters to workload by using a task that varies demand in a gradual and well controlled manner across all participants. Besides the carefully controlled variation of demand, one novelty lies in the fact that multiple areas, covering most of the face were examined for changes in temperature as a response to the variation of demand; another novel aspect is related to quantifying the added value of each of the measures when estimating the level of demand.

Methods: The study presented was conducted in laboratory conditions and required participants to perform a custom-designed visual-motor task that imposed varying levels of demand. The data collected consisted of: physiological measurements (heart inter-beat intervals, breathing rate, pupil diameter, facial thermography); subjective ratings of workload from the participants (ISA and NASA-TLX); and the performance measured within the task.

Results: Facial thermography and pupil diameter were demonstrated to be good candidates for non-invasive mental workload measurements; for 7 out of 10 participants, pupil diameter showed a strong correlation (with R values between 0.61 and 0.79 at a significance value of 0.01) with mean ISA normalized values. Facial thermography measures added on

average 47.7% to the amount of variability in task performance explained by a regression model. As with the ISA ratings, the relationship between the physiological measures and performance showed strong inter-participant differences, with some individuals demonstrating a much stronger relationship between workload and performance measures than others.

5.2 Experiment Design

The study presented in this chapter explores the changes in the physiological parameters that occur as the level of workload varies and examines whether a combination of these parameters could be used for estimating the level of workload. The study uses a task that has varying level of demand with the aim of eliciting different levels of experienced workload which are then captured by subjective and physiological measures. The hypotheses of this study are that:

1. There will be a measurable difference in subjective workload between the two levels of task difficulty
2. The subjective ratings of workload will be associated with changes in physiological measures
3. Multiple physiological measures can be used in combination to analyse workload.

5.2.1 Participants

Fourteen students and staff from the University of Nottingham took part in the study (11 men and 3 women; M age = 28.3 years; SD = 4.9; range = 21-38). The participants were recruited via e-mail and were compensated with a £20 Amazon voucher for their time. The study was approved by the Faculty of Engineering Ethics Committee. Each participant was presented with an information sheet and consent form, stating that they are over 18 years old, had no pre-existing heart-related condition and had no skin conditions or allergies that could prevent them from wearing the heart rate chest strap.

The data from four participants were discarded due to data recording problems and difficulties in tracking the facial features. Data from the remaining ten participants are presented here.

5.2.2 Materials

In order to explore the relationship between workload, variation of performance and objective physiological parameters, a specific computer-based task was designed to impose different

levels of mental demand on the participant. The task consisted of a computer game with 3 stages of two levels of difficulty, in total lasting 29 minutes; each stage consisted of 13 sub-stages (45s each) of varying difficulty, a task paradigm previously used in our research group (Sharples, Edwards, & Balfe, 2012) [77]. Table 5.1 describes the task stages in terms of targets, difficulty level and number of sub-stages.

	Stage 1	Stage 2	Stage 3
Targets	Red balls	Odd numbered balls	Red balls
Difficulty	Level 1 – low difficulty	Level 2 – high difficulty	Level 1 – low difficulty
No. of sub-stages (45s each)	13	13	13

Table 5.1 Task stages description

During each of the stages, the participant is presented with moving coloured balls on a black background. The movement of the balls gives the impression that they are falling from the top of the screen. At the beginning of each of the three stages, the participant is told which the target balls are; the task is to aim at the target balls using a joystick and shoot using a button on the joystick before the balls reach the yellow line and drag it down. During stages 1 and 3 of level 1 difficulty, the target balls are red (Fig. 5.2 Left) while during stage 2 of level 2 difficulty (Fig. 5.2 Right) the colour of the balls no longer represents an identifier of the balls to be targeted. Instead the ones having odd numbers written on them now represent the target, introducing an additional cognitive element with the intent of increasing mental demand. Each of the stages is made up of 13 sub-stages, each presenting the participant with a set number of target balls on the screen at any time; when a target ball is shot, the game generates another one. The number of balls per sub-stage was varied as presented in Fig.5.1 in order to control the level of demand.

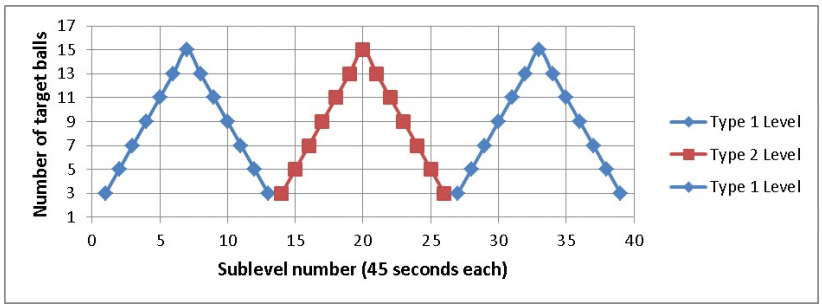

Fig. 5.1 Description of stages

The position of the joystick is indicated by a red circular cursor that turns green once it is within range of the target balls and the participant can make a successful shot (Fig.5.2) using the front button on the joystick. At the beginning of the stage, the horizontal yellow

line finds itself at the top of the screen; when a target ball reaches the yellow line it will drag it down. The participants are told that they have to fight the balls from dragging the yellow line down by shooting at them. Whenever a target ball has been shot, the yellow line goes up by a small increment and whenever the participant misses a shot, the yellow line goes down by the same increment. The main reasons for using the horizontal yellow line were:

- To prevent participants from focusing on the balls that are high on the screen and abandoning the ones that are lower and will soon disappear off the screen, in this way subjecting all participants to the same number of targets at one time;
- To give participants a simple goal to fight towards – keeping the yellow line high up on the screen;
- To obtain a continuous measure of performance in terms of how high on the screen they were able to maintain the yellow line at any moment.

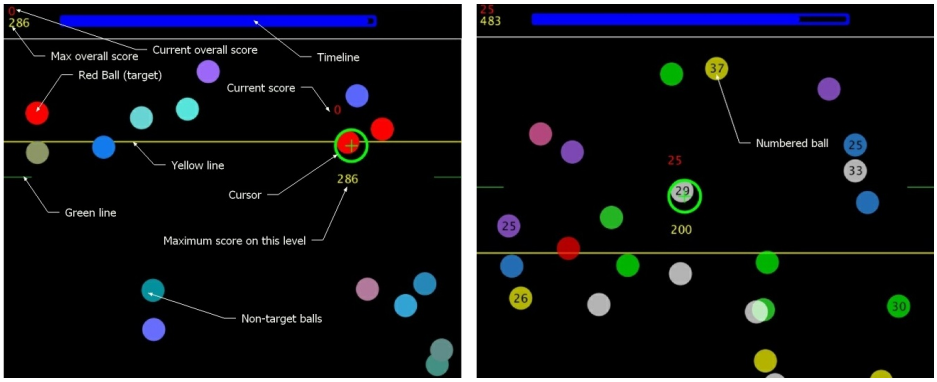

Fig. 5.2 Left: Level 1 difficulty stage (Stage 1) - Right: Level 2 difficulty stage (Stage 2)

After each sub-stage, lasting 45s, the participant was prompted by a voice in the task for their subjective assessment of mental workload, saying: 'Level please'. The task was not frozen while asking for the ISA level, the participant just has to say a number from 1 to 5. At the end of each stage, the task was paused and the participants were shown the task score they achieved in comparison with the other participants as a means of increasing motivation. Sample screen recordings of the task can be found at the following links:

- Stage 1 sample: https://www.youtube.com/watch?v=7a4MaTZ5PzE
- Stage 2 sample: https://www.youtube.com/watch?v=FNwAnWgM024

5.2.3 Design

The independent variable that was manipulated during the study was the task difficulty (i.e. imposed demand). The dependent variables were the physiological measures, the subjective assessment of the perceived level of mental workload and the task performance. The Instantaneous self-assessment workload scale (ISA) (Brennen, 1997) [22] was used once every 45s to collect subjective data about the level of perceived mental workload. The ISA scale was developed primarily as a subjective measure of mental workload for air traffic controllers and it involves the participants self-rating their workload on a scale from 1 (low) to 5 (high). The main reason for using the ISA scale throughout the task was the low level of intrusion, as the participant would verbally rate the perceived level of mental workload when prompted by an auditory message ('Level please'). At the end of each of the three task stages, the participant filled in a NASA-TLX (Hart, California, & Staveland, 1988) [23] questionnaire for a subjective assessment of workload. The reason for using NASA-TLX was to get a more detailed retrospective multidimensional subjective assessment of each of the three stages to determine whether the manipulation of imposed demand through task difficulty had resulted in a perceived experience of increased workload.

5.2.4 Procedure

Each participant was invited to read the information sheet, describing the details of the study, and then fill in a consent form. They were then asked to play a training version of the stimulus task until they became familiar with the rules and the controls. After the training was finished, the participants were invited to attach the Zephyr sensor around their chest in a private space; the thermal and visual cameras were then aligned to match the height of the participant. Before starting the actual task, the eye tracker was calibrated. When the participant was ready, they played stage 1 of the stimulus task, which lasted for almost 10 minutes, at the end of which the participant's score was shown in comparison to the participants before. During the game-play, the participant rated the level of mental workload on the ISA scale once every 45s. After the first stage was over, they filled in the NASA-TLX questionnaire. Stage 2 (higher demand level 2) and 3 (original demand level 1) of the task then followed. Before starting each of the stages, the eye-tracker was recalibrated and after finishing each of the stages, the participant was shown their score and filled in a NASA-TLX questionnaire. After stage 3 ended and the questionnaire had been completed, the participant was invited to remove the Zephyr sensor in a private space.

As ambient temperature could affect facial temperature, room temperature was kept as constant as possible without the use of air conditioning which could result in blowing cold

air towards the participants. Room temperature was recorded every 45s using a room thermometer to insure that there were no external factors that could influence the measurement.

Each participant was offered a £20 voucher as a reward for their time.

5.2.5 Measurements and Equipment

The data collected during the study was of four types:

1. Performance data, generated by the task
2. Subjective data: NASA-TLX and ISA
3. Physiological data collected using the following pieces of equipment[1]:
   (a) Zephyr Bioharness 3
   (b) RED250 Eye Tracker
   (c) FLIR SC7000 Thermal Camera
   (d) fNIRS
4. Room temperature using a digital room thermometer

5.2.6 Data Analysis

The data generated by each of the sensors were post processed in Matlab. The data analysis techniques are described in Chapter 4.

5.2.7 Results

The results are presented in several stages. Firstly the variation of ambient temperature is presented as it could be a factor potentially influencing facial temperature. Secondly the results of the inferential tests to examine the impact of the manipulation of the task demand on the measures of workload and performance are presented. The aim of these tests is to confirm that the demand manipulation affected workload and performance in the manner anticipated. The third analysis examines the relationship between the different measures of workload, using bivariate correlations and reporting both correlation significance and the coefficient of determination to indicate effect size. The final analysis uses multiple linear regression to determine the percentage of variability in task performance explained by the physiological measures and the relative contribution of each of the measures.

[1] All pieces of equipment are described in Chapter 4

Ambient Temperature Data

Ambient temperature was recorded every 45s using a digital room thermometer. The recordings show a gradual slight increase in temperature for each of the participants. Figure 5.3 shows the distribution of temperature variation throught the study. Each blue marker overlayed on top of the boxplot represents the difference between the room temperature at the end and the room temperature at the beginning of the study; for all of the participants, this also represents the maximum variation of ambient temperature. We can observe that for only two of the participants, the variation in ambient temperature was greater than 0.5°C, respectively 0.7°C and 0.9°C.

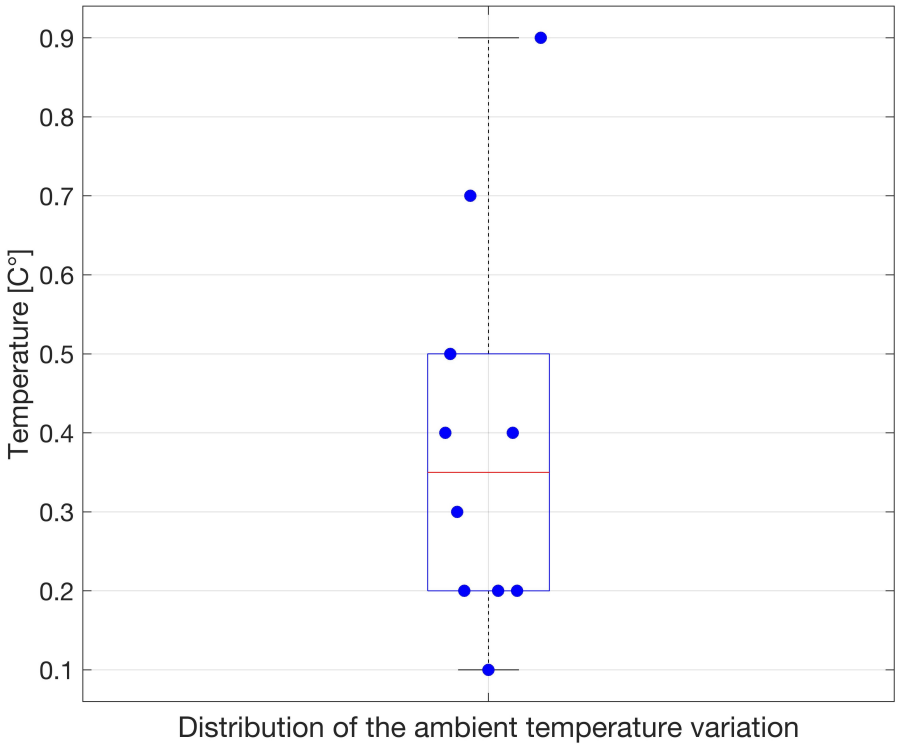

Fig. 5.3 Distribution of the ambient temperature variation

In the case of participant 6, the increase in ambient temperature was the largest, which might explain the general trend of facial temperature increase. Figure 5.15 shows the variation of temperature in the upper nose area for participant 6; while we can observe a slight decrease in temperature around stages 20 and 33, as for the other participants, we can also observe a gradual increase in temperature towards the end of the task.

Subjective and Performance Data

A one way ANOVA (F(1,28) = 4.56, p = 0.041, $\eta^2 = 0.14$) confirmed that there was a difference between the two levels of difficulty in terms of the NASA-TLX mental demand scale, confirming that stage 2 (odd numbered balls as targets) was perceived to be more mentally demanding than stages 1 and 3 (red balls as targets), however the effect size is small, group differences explaining about 14% of the variance. One of the disadvantages of using the ISA technique is subjectivity in interpretation of the absolute meaning of numbers on the rating scale, and thus the limited absolute validity that can be inferred from the ratings. However, it can be assumed that the relative validity of the ratings is robust, and therefore in order to compare the results across the participants, the data were normalized to a common scale ranging between 0 and 1. Fig.5.4 shows the mean performance score for all participants (better performance in the task results in a higher score) at sub-stage scale, plotted against the mean normalized ISA rating for all participants. There is a negative correlation between the two mean scores, the Pearson correlation coefficient is r(37) = -0.74 with p<0.01[2], showing that as the mean subjective level of mental workload increased, the mean task performance decreased. While Fig.5.4 looks at the mean performance and level of mental workload, Table 5.2 shows the individual correlations with performance of both the mean ISA normalized and to each participant's rating. It can be observed that for the individual (non-normalized) ISA ratings, three of the participants did not have significant correlations to the 0.05[3] level and that the r^2 value is smaller in general compared with the mean ISA normalized correlation. Overall these data demonstrates a clear association between performance and subjective workload.

Physiological Data

The physiological data collected consisted of heart R-R inter-beat intervals, breathing rate, pupil diameter and facial skin temperatures measured by thermography. All physiological data reported are the mean of the readings taken during the 45s duration of each of the sub-stages. Due to the fact that physiological data depends so much on the physiology of each of the participants and also on the reaction each of them has to the stimulus task, the correlations of each of the physiological signals with the ISA subjective ratings (both mean normalized and individual values) will be presented in tabular form for each of the

[2]The Pearson correlation coefficient will be presented together with the degrees of freedom, in this case 37; there are N=39 (sub-stages), df = N-2

[3]Note that no familywise corrections such as Bonferonni were applied, as tests were conducted on independent (participant-based) data sets, but it should be acknowledged that as normal when multiple tests are conducted one in twenty will be significant by chance if a p<0.05 level of significance is adopted

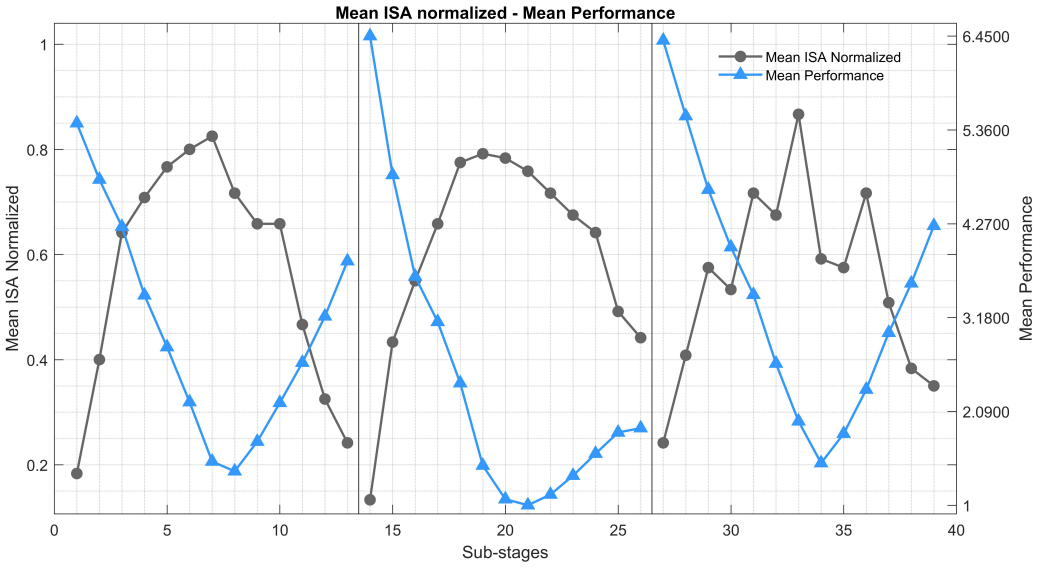

Fig. 5.4 Mean ISA ratings - mean score

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.632**	**0.401**	**<0.01**	-0.3	0.09	0.0632
2	**-0.652**	**0.462**	**<0.01**	**-0.574**	**0.33**	**<0.01**
3	**-0.648**	**0.420**	**<0.01**	**-0.620**	**0.385**	**<0.01**
4	**-0.706**	**0.499**	**<0.01**	**-0.421**	**0.178**	**<0.01**
5	**-0.729**	**0.532**	**<0.01**	**-0.551**	**0.304**	**<0.01**
6	**-0.659**	**0.434**	**<0.01**	**-0.434**	**0.188**	**<0.01**
7	**-0.759**	**0.576**	**<0.01**	-0.229	0.053	0.15
8	**-0.783**	**0.613**	**<0.01**	**-0.754**	**0.569**	**<0.01**
9	**-0.681**	**0.465**	**<0.01**	-0.038	0.001	0.81
10	**-0.742**	**0.551**	**<0.01**	**-0.769**	**0.592**	**<0.01**

Table 5.2 Mean and normalized ISA ratings correlated with Performance

participants individually, together with strong and weak [4] correlation example plots. This helps us understand whether any association between physiology and subjective ratings applies across a population or whether there are different levels of strength of relationships between different predictive variables in different populations. As a subjective measure, the ISA rating is valuable because it is easy to administer, not requiring the interruption of the task, and has high face validity. Even though on some occasions participants might choose to rate their workload lower or higher, for the moment, until other means of measuring

[4]The guide for the absolute value of r was suggested by Evans(1996)[78]

workload are developed, subjective measures will play a big part in understanding how humans perceive demand.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.696**	**0.484**	**<0.01**	**-0.535**	**0.286**	**<0.01**
2	-0.197	0.039	0.22	0.016	0	0.92
3	-0.173	0.030	0.29	-0.183	0.033	0.26
4	**0.47**	**0.221**	**<0.01**	0.079	0.006	0.62
5	-0.276	0.076	0.08	**-0.323**	**0.104**	**<0.04**
6	**-0.573**	**0.328**	**<0.01**	**-0.454**	**0.206**	**<0.01**
7	0.185	0.034	0.25	-0.202	0.041	0.21
8	-0.222	0.049	0.17	-0.198	0.039	0.22
9	**-0.349**	**0.122**	**0.02**	**-0.327**	**0.107**	**0.04**
10	-0.05	0.003	0.75	-0.112	0.013	0.49

Table 5.3 R-R intervals correlated with subjective ISA reports

Table 5.3, shows the correlation of the R-R inter beat intervals with both the mean normalized ISA values and the individual ISA ratings; correlations with p values smaller than 0.05 are bolded. For three of the participants (1, 6 and 9), the R-R values were significantly correlated to both the mean normalized ISA and to their individual ISA ratings. A negative moderate correlation was found for participants 1 and 6 while participant 9 showed a weak correlation with the subjective ISA ratings. The R-R values for participants 1 and 6 showed a moderate negative correlation with their individual ISA ratings but not a significant correlation with the mean normalized values. Participant 4 was the only participant to show a positive significant correlation between R-R and mean ISA normalized. Figure 5.5 shows the R-R measure for participant 1 plotted against mean ISA normalized and individual ISA, representing an example of strong correlation while Figure 5.6 shows the same measures for participant 10, representing the weakest correlation.

Table 5.4 shows the correlations of pupil diameter with both the mean normalized ISA values and the individual ISA ratings; pupil diameter data from all participants except for 7 and 10 have moderate to strong positive correlations with the mean ISA normalized. Participants 7 and 10 show a weak positive correlation with the individual ISA ratings. Only participants 1 and 9 do not show a significant correlation to the individual ISA ratings. For most participants, a clear increase in pupil diameter was observed with the increase of workload. Figure 5.7 shows the pupil diameter measure for Participant 9 plotted against mean ISA normalized and individual ISA, representing an example of strong correlation, whereas Figure 5.8 shows the same measures for Participant 7, representing the weakest correlation.

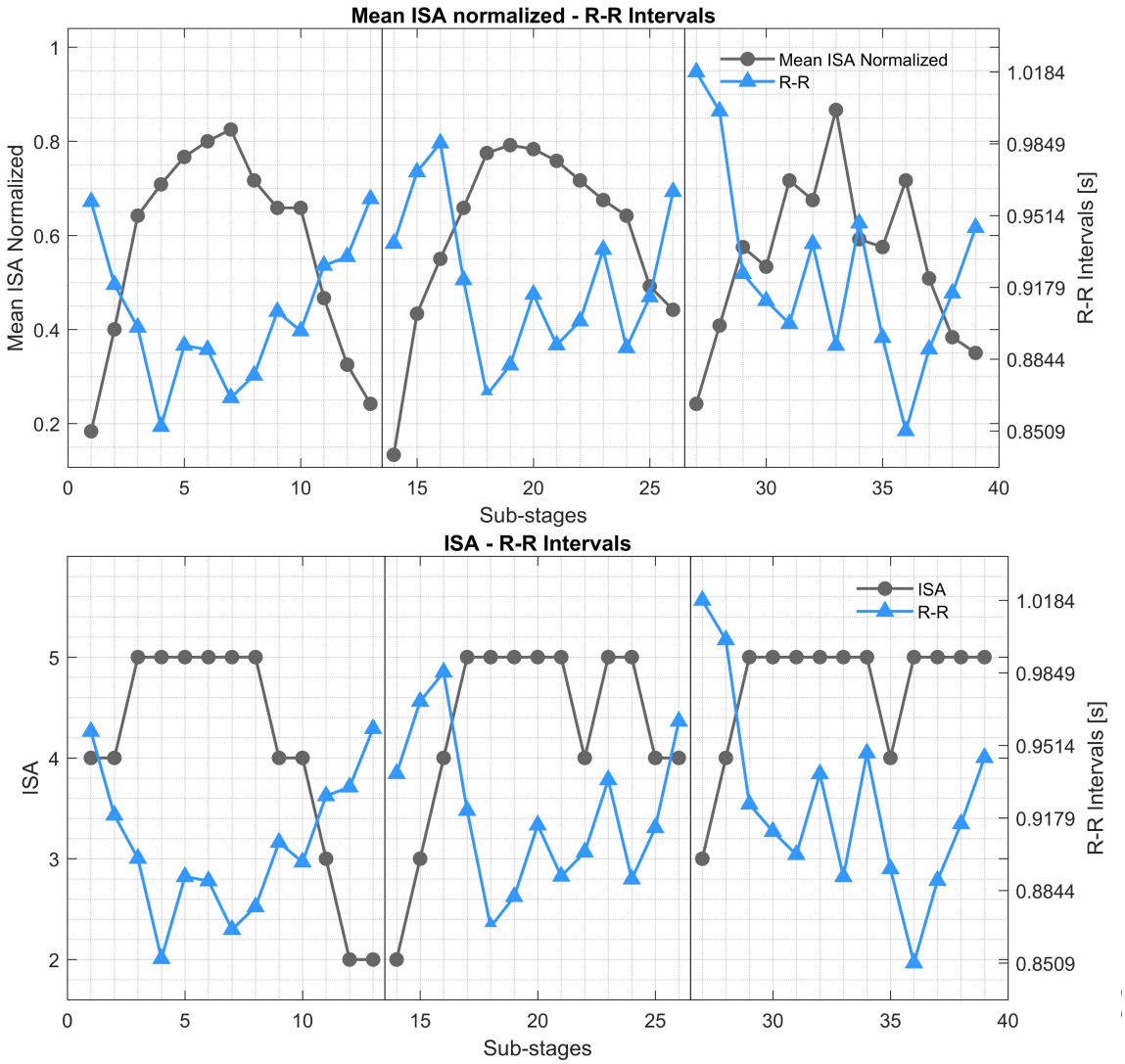

Fig. 5.5 R-R – mean ISA normalized (top) and individual ISA (bottom) for participant 1 (strongest correlation)

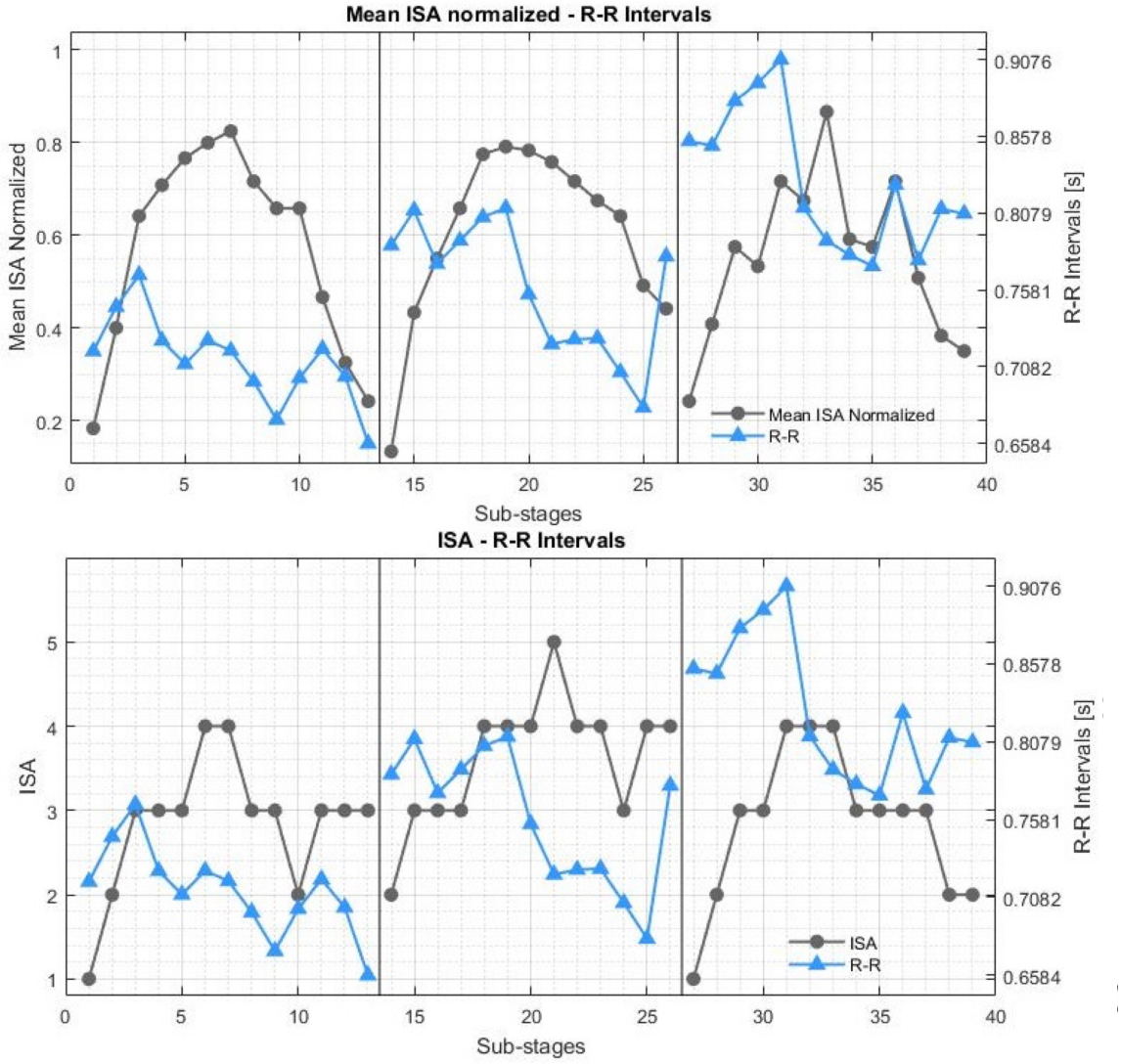

Fig. 5.6 R-R – mean ISA normalized and individual ISA for participant 10 (weakest correlation)

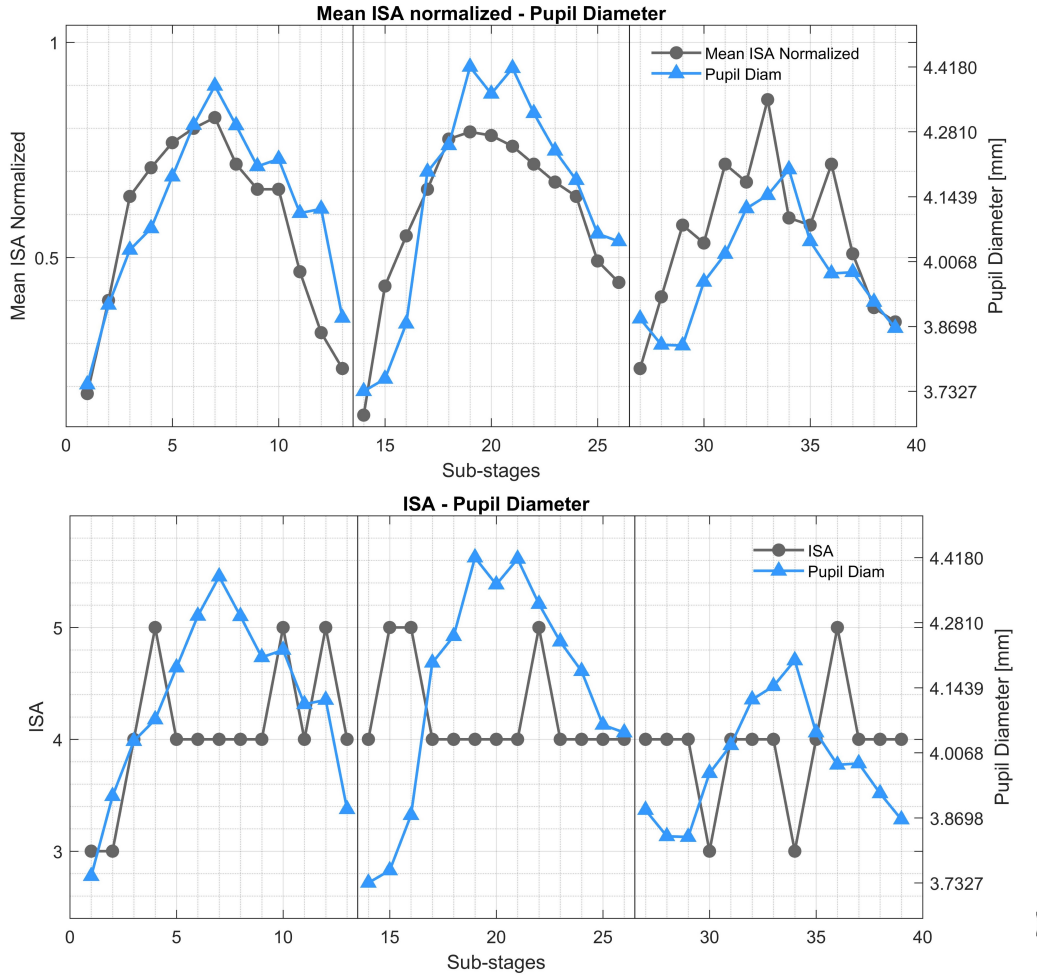

Fig. 5.7 Pupil diameter – mean ISA normalized and individual ISA for participant 9 (strong correlation for mean ISA normalized but weak correlation for individual ISA)

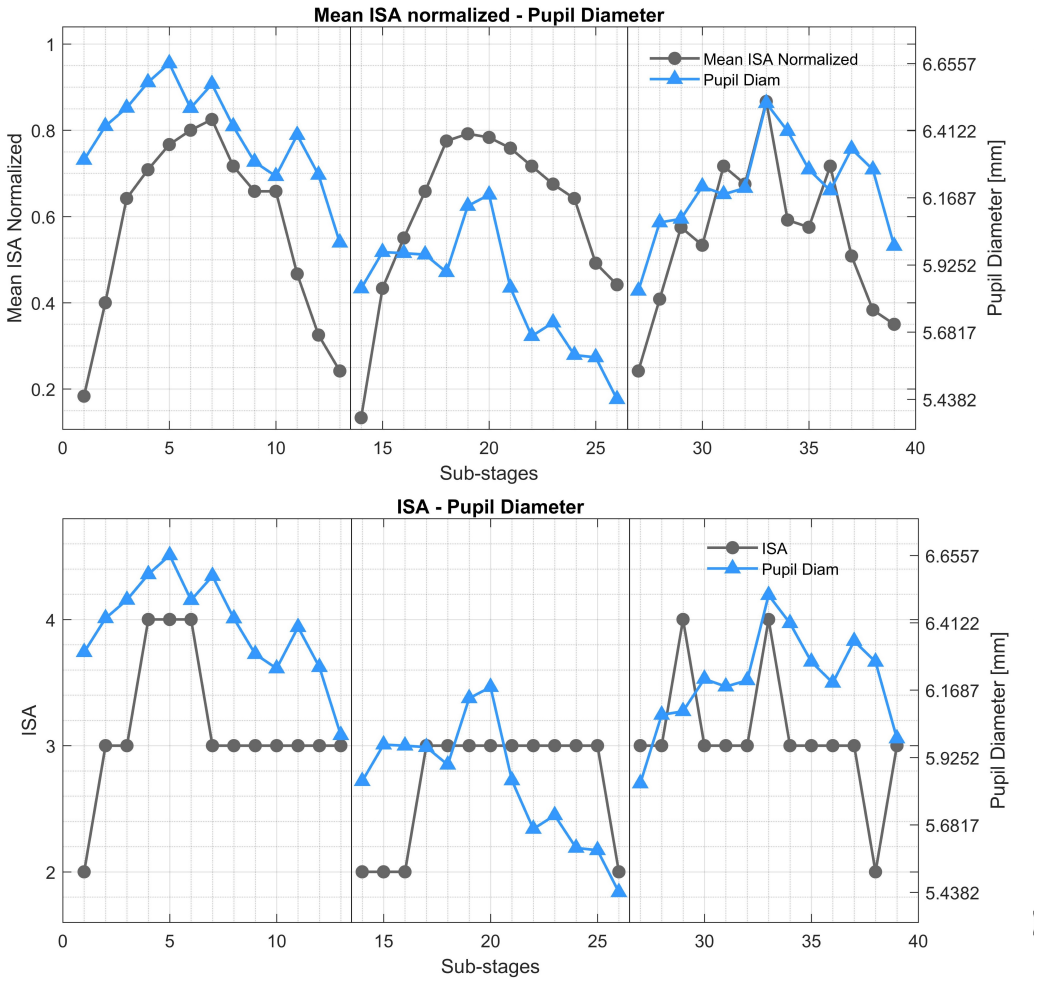

Fig. 5.8 Pupil diameter – mean ISA normalized and individual ISA for participant 7 (non-significant correlation for mean ISA normalized but weak positive correlation with individual ISA ratings)

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**0.617**	**0.381**	**<0.01**	0.309	0.095	0.05
2	**0.675**	**0.456**	**<0.01**	**0.544**	**0.296**	**<0.01**
3	**0.635**	**0.403**	**<0.01**	**0.497**	**0.247**	**<0.01**
4	**0.611**	**0.373**	**<0.01**	**0.677**	**0.458**	**<0.01**
5	**0.449**	**0.202**	**<0.01**	**0.35**	**0.123**	**0.02**
6	**0.705**	**0.497**	**<0.01**	**0.668**	**0.446**	**<0.01**
7	0.268	0.072	0.09	**0.435**	**0.189**	**<0.01**
8	**0.658**	**0.433**	**<0.01**	**0.601**	**0.361**	**<0.01**
9	**0.79**	**0.624**	**<0.01**	0.073	0.005	0.65
10	0.308	0.095	0.05	**0.489**	**0.239**	**<0.01**

Table 5.4 Pupil Diameter correlated with subjective ISA reports

Table 5.5 shows the correlations of breathing rate with both the mean normalized ISA values and the individual ISA ratings; only the breathing rate data for participant 7 showed a moderate positive correlation with the mean normalized ISA values and a weak correlation with the individual ISA values. Participant 1 showed a moderate positive correlation between breathing rate and individual ISA ratings. Figure 5.9 shows the breathing rate measure for Participant 7 plotted against mean ISA normalized and individual ISA, representing an example of strong correlation, whereas Figure 5.9 shows the same measures for Participant 2, representing the weakest correlation.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	0.115	0.013	0.48	**0.419**	**0.176**	**<0.01**
2	0.0005	0	0.99	-0.144	0.021	0.38
3	-0.095	0.009	0.56	-0.169	0.029	0.3
4	0.061	0.004	0.71	0.242	0.059	0.13
5	0.097	0.009	0.55	0.038	0.001	0.81
6	-0.258	0.067	0.11	-0.122	0.015	0.45
7	**-0.661**	**0.437**	**<0.01**	**0.393**	**0.154**	**0.01**
8	0.14	0.02	0.39	0.137	0.019	0.4
9	0.304	0.092	0.05	0.088	0.008	0.59
10	0.297	0.088	0.08	0.232	0.054	0.15

Table 5.5 Breathing Rate correlated with subjective ISA reports

In order to extract the thermal data from the images, a feature tracking algorithm was deployed, splitting the face into regions of interest. For each frame, the temperature was extracted from inside the circular points, from along the lines and from inside some of the

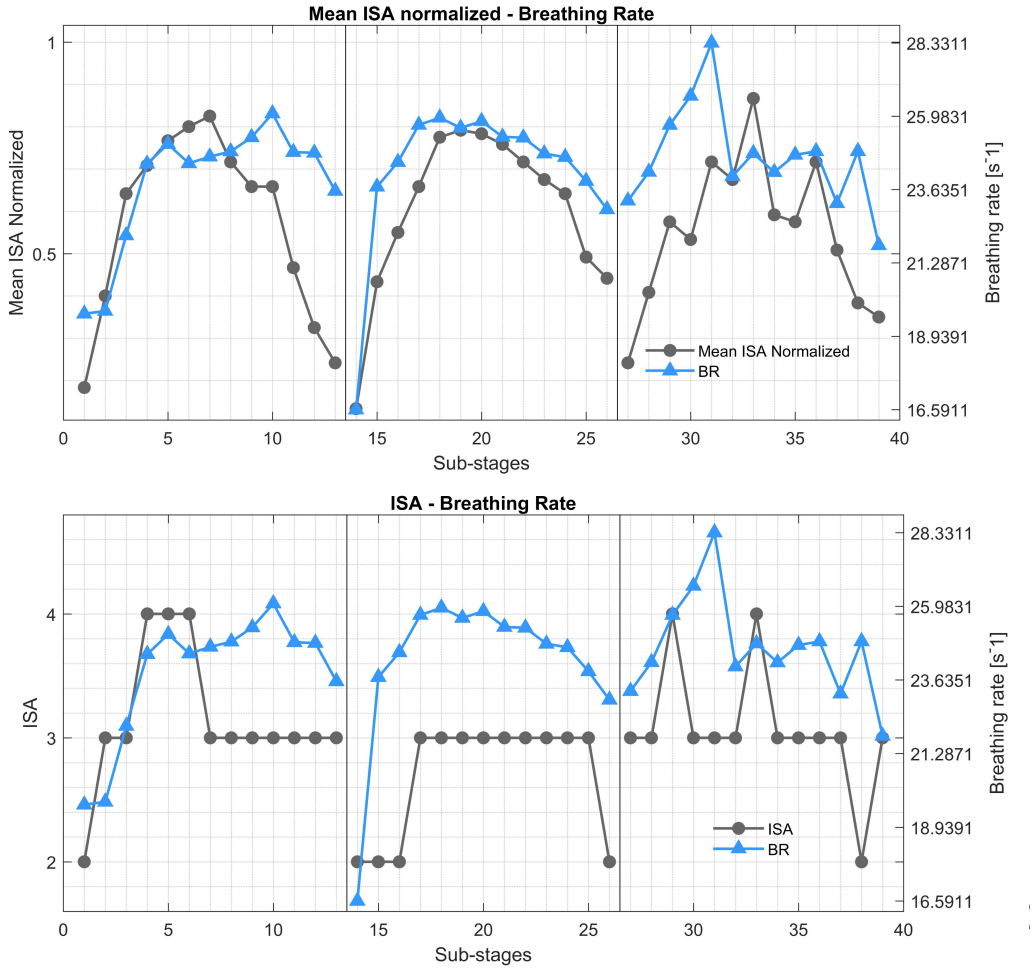

Fig. 5.9 Breathing rate – mean ISA normalized and individual ISA for participant 7 (strongest correlation with mean ISA normalized)

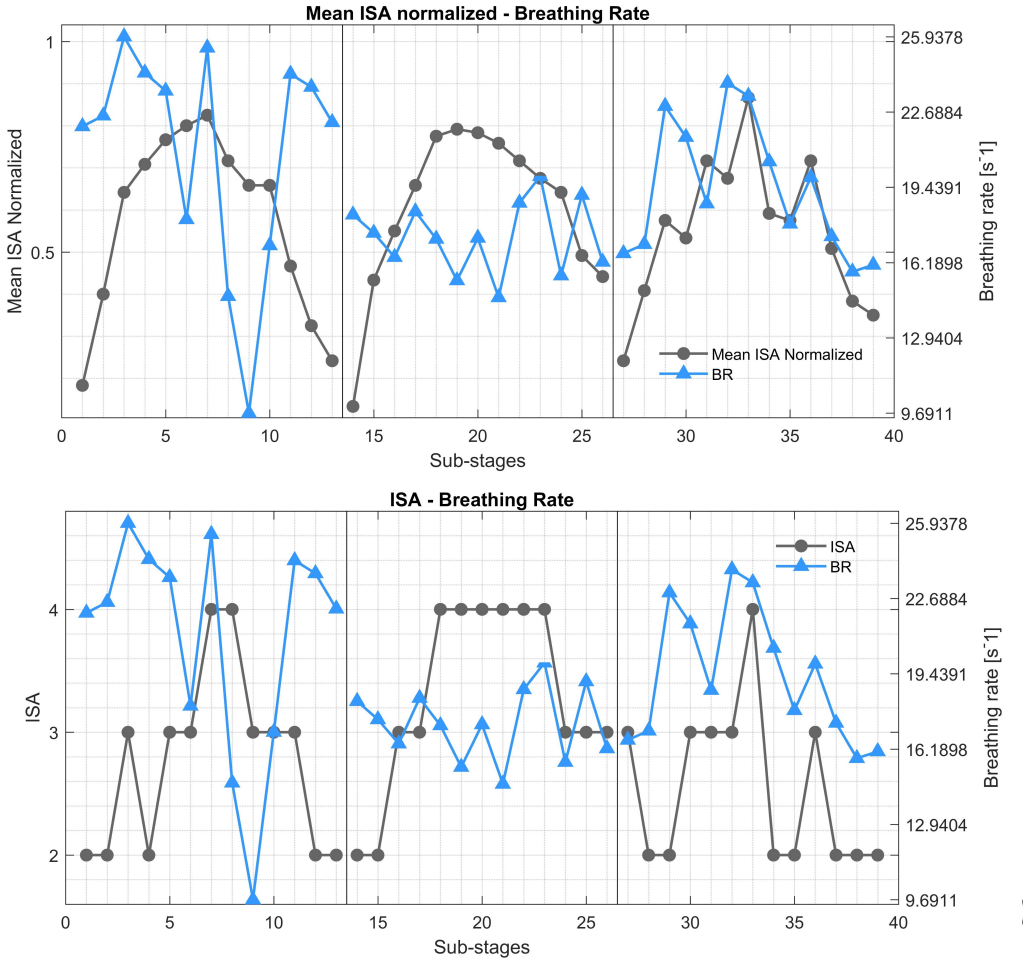

Fig. 5.10 Breathing rate – mean ISA normalized and individual ISA for participant 2 (non-significant correlation example)

triangular areas without using markers, making the technique less intrusive (Fig. 5.11). Features from below the nose were not tracked due to the difficulty imposed by facial hair in some of the participants. The nose and forehead can therefore be considered as ideal sites for skin temperature measurement, as they would normally be un-occluded, which might present a challenge in a real life application as well.

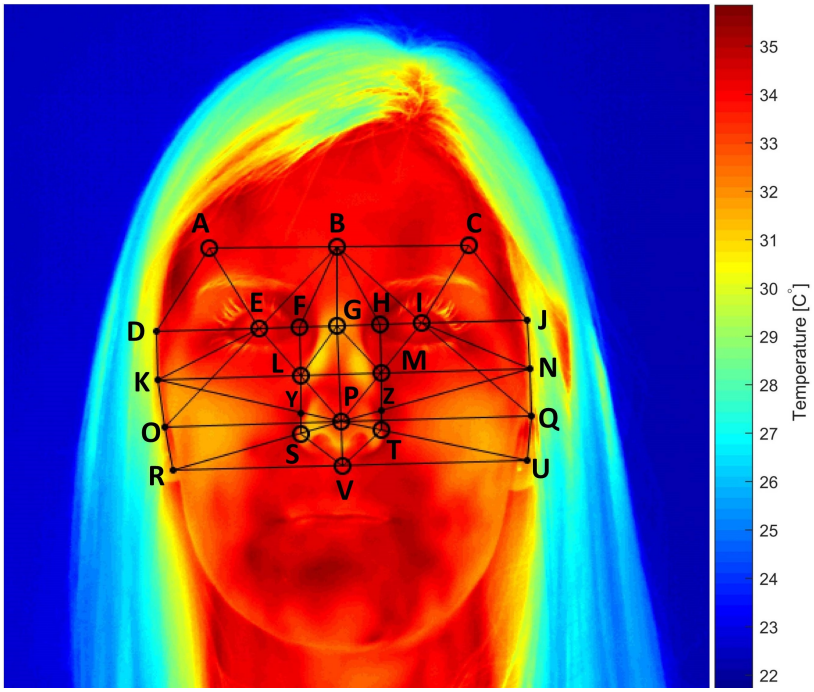

Fig. 5.11 Facial Landmarks Example

Table 5.6 shows the correlations of the average temperature inside point P (nose tip) with both the mean normalized ISA values and the individual ISA ratings; only participants 1 and 9 showed strong and moderate negative correlations to the 0.01 level for the mean ISA normalized. Participants 2, 7 and 10 showed weak negative correlations, significant to the 0.05 level with the mean ISA normalized values. Participant 6 was the only one to show a weak positive correlation with the mean ISA normalized values. Participants 4 and 7 showed stronger correlations with the individual (non-normalized) ISA ratings.

Table 5.7 shows the correlations of the average temperature inside point V with both the mean normalized ISA values and the individual ISA ratings; only participant 1 showed a strong negative correlation to 0.01 level for the mean ISA normalized while participants 2 and 4 showed a weak negative correlation, significant to the 0.05 level with the mean ISA normalized. Participants 1, 4 and 7 showed moderate to weak negative correlations with the individual ISA values. Figure 5.12 shows the temperature in points P and V for Participant 1 plotted against mean ISA normalized and individual ISA, representing an example of strong

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.746**	**0.557**	**<0.01**	**-0.507**	**0.257**	**<0.01**
2	**-0.373**	**0.139**	**0.01**	-0.07	0.005	0.66
3	-0.075	0.006	0.64	-0.137	0.019	0.4
4	-0.152	0.023	0.35	**-0.429**	**0.184**	**<0.01**
5	-0.167	0.028	0.3	-0.008	0	0.95
6	**0.345**	**0.119**	**0.03**	0.188	0.035	0.25
7	**-0.401**	**0.161**	**0.01**	**-0.459**	**0.211**	**<0.01**
8	-0.086	0.007	0.6	-0.042	0.002	0.79
9	**-0.514**	**0.264**	**<0.01**	-0.208	0.043	0.2
10	**-0.329**	**0.108**	**0.04**	-0.028	0.001	0.86

Table 5.6 Point P temperature correlated with subjective ISA reports

correlation, whereas Figure 5.13 shows the same measures for Participant 8, representing the weakest correlation.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.724**	**0.524**	**<0.01**	**-0.468**	**0.219**	**<0.01**
2	**-0.354**	**0.125**	**0.02**	-0.05	0.003	0.76
3	-0.267	0.071	0.09	-0.1	0.01	0.54
4	**-0.382**	**0.146**	**0.01**	**-0.381**	**0.145**	**0.01**
5	-0.284	0.081	0.07	-0.132	0.017	0.41
6	0.146	0.021	0.37	-0.118	0.014	0.47
7	-0.296	0.088	0.06	**-0.382**	**0.146**	**0.01**
8	-0.107	0.011	0.51	-0.075	0.006	0.64
9	-0.035	0.001	0.83	-0.118	0.014	0.47
10	-0.307	0.094	0.05	-0.156	0.024	0.34

Table 5.7 Point V temperature correlated with subjective ISA reports

Table 5.8 shows the correlations of the average temperature inside point L with both the mean normalized ISA values and the individual ISA ratings; participants 1, 2, 9, 10 showed moderate negative correlations while participants 7 showed weak negative correlations with the mean ISA normalized levels. Participants 4 and 7 showed a moderate to weak negative correlation with the individual ISA ratings.

Table 5.9 shows the correlations of the average temperature inside point M with both the mean normalized ISA values and the individual ISA ratings; participants 1, 2, 4, 7, 9 and 10 showed a moderate negative correlation with the mean ISA normalized while participant 5 showed a weak negative correlation with the mean ISA normalized. Participants 2, 4 and

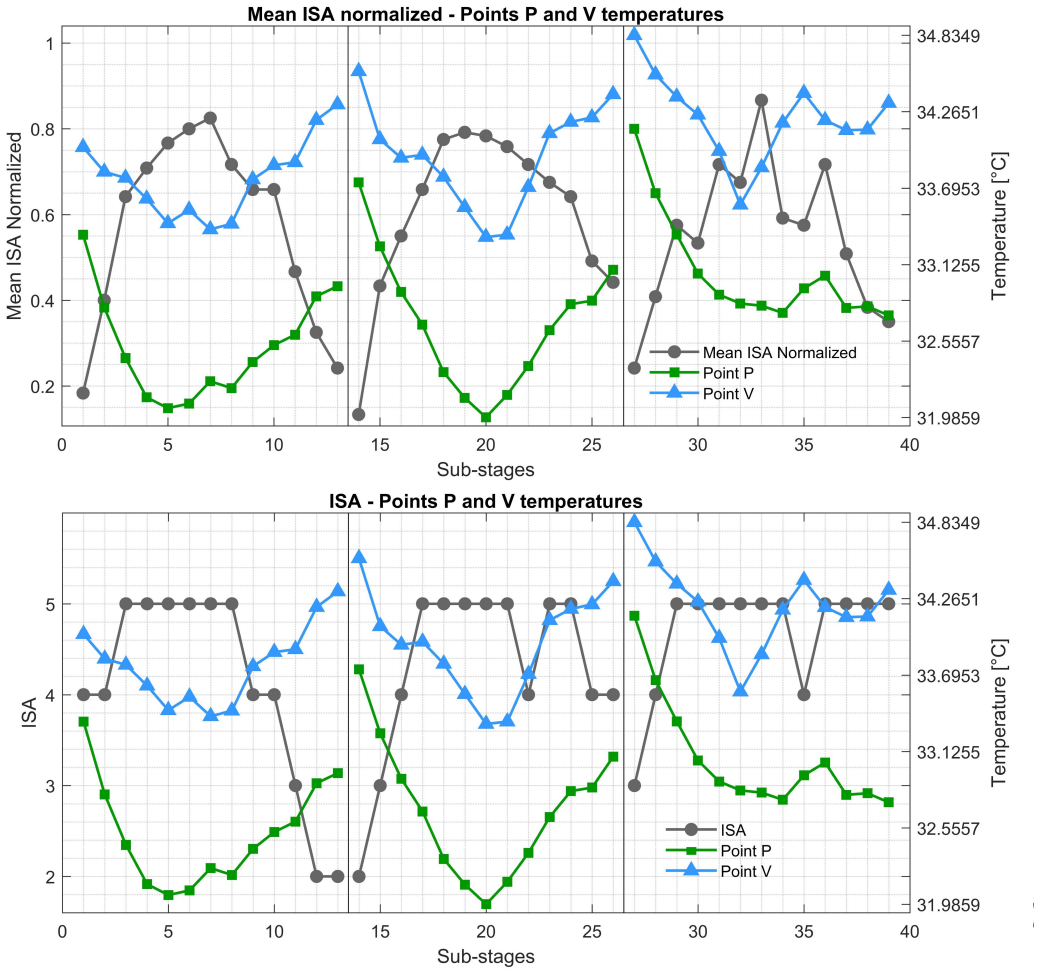

Fig. 5.12 P, V points temperature – mean ISA normalized and individual ISA for participant 1 (strong correlation example)

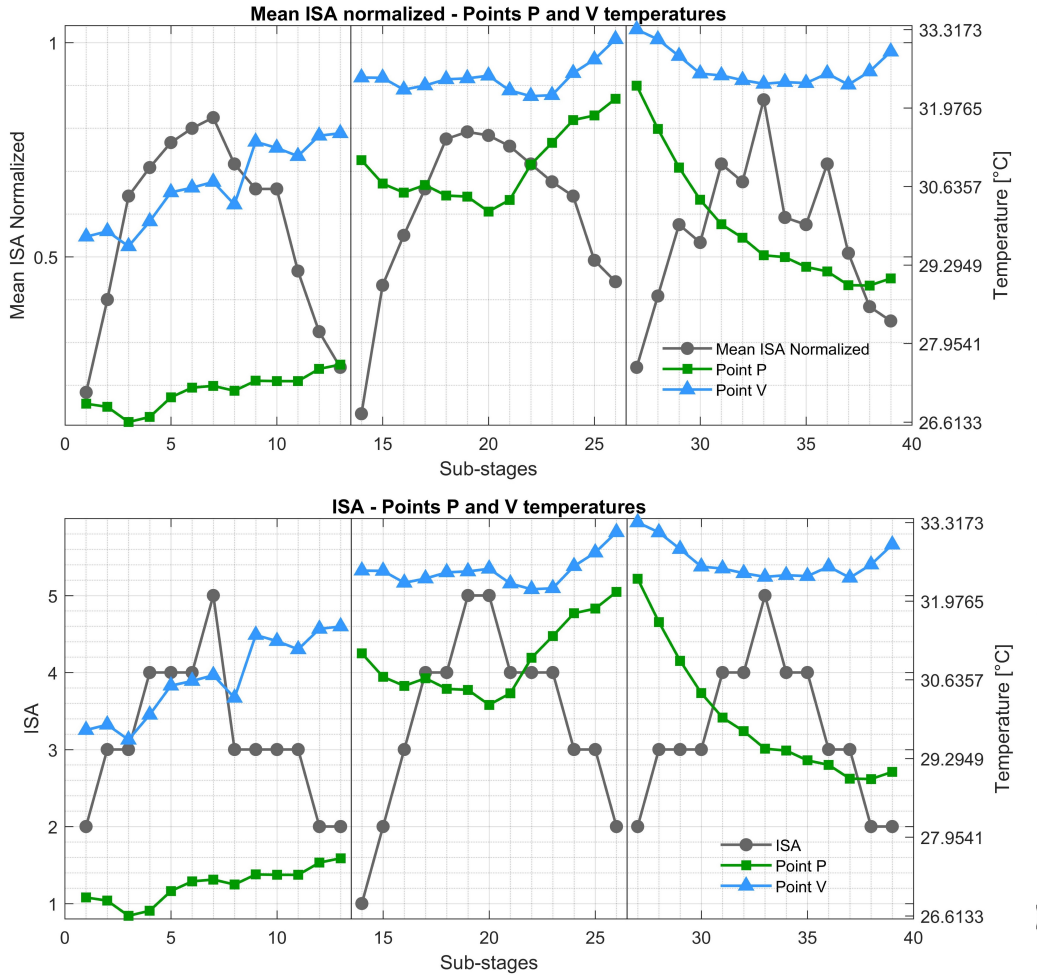

Fig. 5.13 P, V points temperature – mean ISA normalized and individual ISA for participant 8 (non significant correlation example)

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.533**	**0.284**	**<0.01**	-0.249	0.062	0.12
2	**-0.501**	**0.251**	**<0.01**	-0.292	0.085	0.07
3	-0.196	0.038	0.23	-0.006	0	0.96
4	-0.08	0.006	0.59	**-0.471**	**0.222**	**<0.01**
5	-0.289	0.084	0.07	-0.155	0.024	0.34
6	-0.025	0.001	0.87	-0.023	0.001	0.88
7	**-0.373**	**0.139**	**0.01**	**-0.377**	**0.142**	**0.01**
8	-0.16	0.026	0.33	-0.082	0.007	0.61
9	**-0.594**	**0.353**	**<0.01**	-0.157	0.025	0.33
10	**-0.457**	**0.209**	**<0.01**	-0.169	0.029	0.3

Table 5.8 Point L temperature correlated with subjective ISA reports

10 showed a moderate negative correlation with the individual ISA ratings while participant 7 showed a weak negative correlation to the individual ISA ratings. Figure 5.14 shows the temperature in points L and M for Participant 2 plotted against mean ISA normalized and individual ISA, representing an example of strong correlation, whereas Figure 5.15 shows the same measures for Participant 6, representing the weakest correlation.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.472**	**0.223**	**<0.01**	-0.251	0.063	0.12
2	**-0.674**	**0.454**	**<0.01**	**-0.511**	**0.261**	**<0.01**
3	-0.081	0.007	0.62	0.076	0.006	0.64
4	**-0.419**	**0.176**	**<0.01**	**-0.509**	**0.259**	**<0.01**
5	**-0.386**	**0.149**	**0.01**	-0.248	0.062	0.12
6	0.116	0.013	0.48	0.069	0.005	0.67
7	**-0.542**	**0.294**	**0.01**	**-0.358**	**0.128**	**0.02**
8	-0.16	0.026	0.32	-0.071	0.005	0.66
9	**-0.643**	**0.413**	**<0.01**	-0.186	0.035	0.25
10	**-0.543**	**0.295**	**<0.01**	**-0.584**	**0.341**	**<0.01**

Table 5.9 Point M temperature correlated with subjective ISA reports

Multiple physiological measures as an indication of workload level

The previous sections have looked at the correlation of each of the physiological measures with the subjective measures of workload. Some physiological measures have shown higher and some have shown lower or no correlation at all to the subjective workload measures.

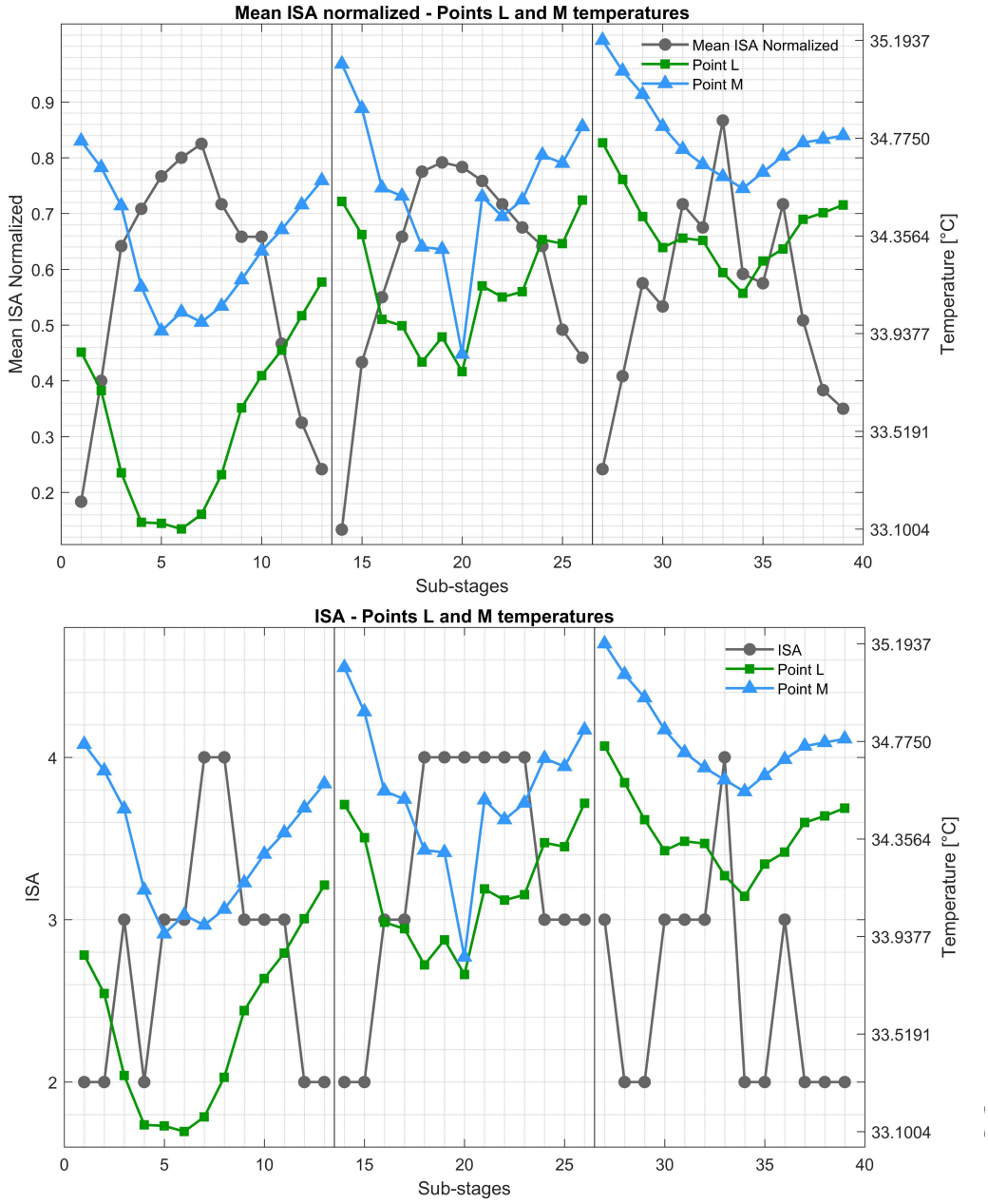

Fig. 5.14 L, M points temperature – mean ISA normalized and individual ISA for participant 2 (strongest correlations)

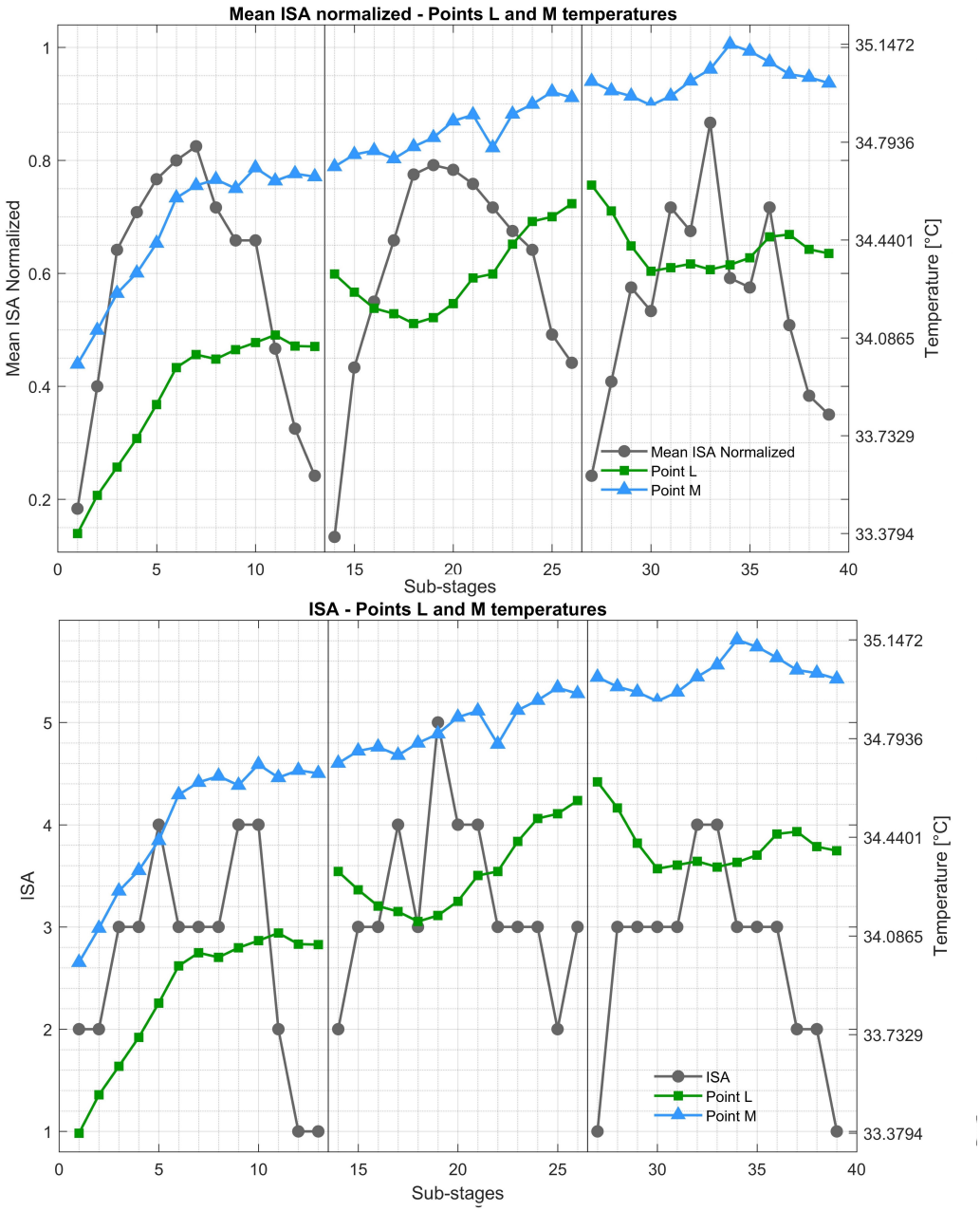

Fig. 5.15 L, M points temperature – mean ISA normalized and individual ISA for participant 6 (non-significant correlations)

This was to be expected, as different measures could show senstivity to different aspects of workload or may have a different bandwidth in which they are sensitive.

This section explores how the different physiological measures can be combined to produce a more accurate measure of workload. Some of the measures presented above show promising correlations to the subjective ISA measure of mental workload. A multiple linear regression was performed for each participant individually, on the entire set of data, on more combinations of the predictor variables to test which one explains more of the variability in the response variable and how different physiological parameters can be combined for more reliable and valid capture of workload. Four combinations of the predictor variables were chosen:

1. Heart (R-R interval) and Breathing Rate data (Mean RR, Mean BR)
2. The Heart and Breathing Rate data and pupil diameter
3. The heart and breathing rate data, pupil diameter and the facial temperatures inside points : 'B', 'F", 'G', 'H', 'L', 'M', 'P', 'V'
4. Facial temperatures inside points : 'B', 'F", 'G', 'H', 'L', 'M', 'P', 'V'

The reason behind the choice of the predictor variables combinations was to start with features from only one of the sensor and gradually add the others; category 1 contains just the features produced by the Zephyr sensor, category 2 adds pupil diameter to category 1, category 3 contains the combined features from the first two categories in addition to the facial thermography measures and category 4 contains just the facial thermography features.

It has to be stated that performance measures are not expected to correlate to either subjective of physiological measures, as performance can be protected and therefore maintained at the same level through investing more resources in accomplishing the task. However, in the particular case of the task used for this study, performance showed a high correlation to subjective measures, being in a way almost a measure of workload in itslef; this should not be generalized on any task. For this reason game performance, rather than ISA ratings, was selected as the response variable for this analysis. Game performance was used as it strongly correlates with the subjective ISA ratings and it is also a continuous variable that was needed for the analysis; game performance is represented by the height the participants managed to maintain the yellow line on the screen.

Some of the predictor variables for some of the participants were highly correlated to each other. Inter-variable correlation influences the ability of multiple linear regression to distinguish between the predictive ability of each individual variable. Our approach to this

limitation was to systematically add and remove predictors based on the F-statistic; the tool used for this was stepwise regression in Matlab. The algorithm starts with a constant model and iteratively adds and removes predictors until the model can no longer be improved substantially. Each of the sections in Tables 10,11,12 and 13 shows the multiple linear regression results for each of the described groups of predictors for each participant. The adjusted r^2 column contains the proportion of variability of the dependent variable accounted for by the regression model. Because the r^2 value increases by adding more predictor variables in the model, the adjusted r^2 value was reported in order to make the comparison between models more meaningful. The table also displays the F statistic of the linear fit versus the constant model, testing the statistical significance of the model; the predictors column contains the names of the predictors selected by the algorithm for each of the regressions. The Beta column contains the estimate standardized coefficients of the terms in the regression, indicating how many standard deviations the dependent variable will change with the change of one standard deviation in the predictor variable, allowing for a comparison of the relative contribution of each of the predictors. The t-statistic test for the significance of each term given the other terms in the model is used to test the null hypothesis that the term is equal to zero (versus the alternate hypothesis that the coefficient is different from zero). The associated p values are also reported in the table.

Predictors	Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
Combination 1: Mean RR Mean BR	1	0.404	0.77	13.91	<0.01	Mean RR	0.56	4.34	<0.01
						Mean BR	0.53	4.07	<0.01
	2	0	-	-		-	-	-	-
	3	0.187	0.9	9.75	<0.01	Mean BR	0.45	3.12	<0.01
	4	0	-	-		-	-	-	-
	5	0	-	-		-	-	-	-
	6	0.434	0.75	30.16	<0.01	Mean RR	0.67	5.49	<0.01
	7	0.265	0.85	14.76	<0.01	Mean BR	-0.53	-3.84	<0.01
	8	0	-	-		-	-	-	-
	9	0.224	0.88	11.97	<0.01	Mean RR	0.49	3.46	<0.01
	10	0.116	0.93	6.01	0.019	Mean BR	-0.37	-2.54	0.019

Table 5.10 Proportion of the variability accounted for by the regression model in the response variable using combination 1 of predictors

The results presented in Tables 5.10,5.11,5.12 and 5.13 show that for combination 3, when using all the predictor variables, for 7 out of 10 participants, the pupil diameter measure was demonstrated to be a good predictor of performance, followed by temperature in point P for 6 out of 10 participants. On average, facial thermography measures added 47.7% to the amount of variability explained by the regression model. Figure 5.16 below shows a boxplot summary of the regression tables presented above, in terms of adjusted r^2 and RMSE. It can be seen that for the predictors in combination 3 the amount of variability explained is

Predictors	Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
Combination 2: Mean RR Mean BR Pupil Diameter	1	0.404	0.77	13.91	<0.01	Mean RR	0.56	4.34	<0.01
						Mean BR	0.53	4.07	<0.01
	2	0	-	-		-	-	-	-
	3	0.187	0.9	9.75	<0.01	Mean BR	0.45	3.12	<0.01
	4	0.338	0.81	20.43	<0.01	Pupil Diameter	-0.59	-4.52	<0.01
	5	0.092	0.95	4.85	<0.05	Pupil Diameter	-0.34	-2.2	<0.05
	6	0.434	0.75	30.16	<0.01	Mean RR	0.67	5.49	<0.01
	7	0.265	0.85	14.76	<0.01	Mean BR	-0.53	-3.84	<0.01
	8	0.280	0.84	15.84	<0.01	Pupil Diameter	-0.54	-3.98	<0.01
	9	0.696	0.55	88.39	<0.01	Pupil Diameter	-0.83	-9.4	<0.01
	10	0.474	0.72	18.15	<0.01	Mean RR	0.45	3.82	<0.01
						Pupil Diameter	-0.62	-5.24	<0.01

Table 5.11 Proportion of the variability accounted for by the regression model in the response variable using combination 2 of predictors

higher than all other combinations but close to combination 4. At the same time, the RMSE is smallest for combination 3, indicating a better fit compared to the other models. Based on the data collected in this study, for most of the participants, pupil diameter together with thermal data measured around the nose area provided the best combination of predictors for inferring the level of performance.

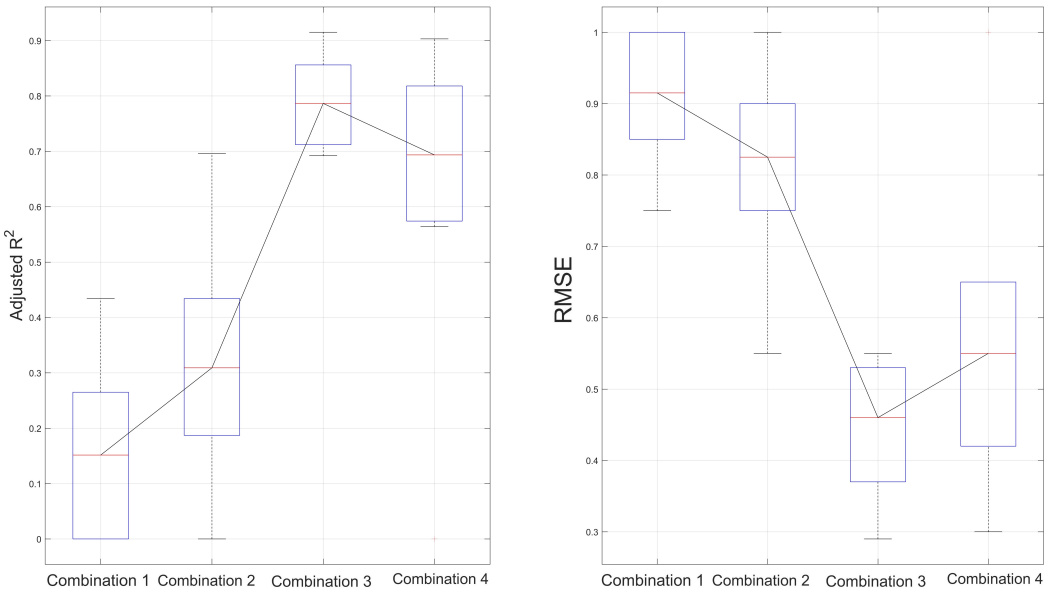

Fig. 5.16 Adjusted R^2 and RMSE for each of the four combinations of predictors

Predictors	Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
Combination 3: Mean RR Mean BR Pupil Diameter Temperatures inside points: 'B' 'F' 'G' 'H' 'L' 'M' 'P' 'V'	1	0.786	0.46	36.03	<0.01	Pupil Diameter	-0.33	-3.72	<0.01
						B	-1.19	-8.62	<0.01
						F	0.76	5.59	<0.01
						P	0.58	6.57	<0.01
	2	0.888	0.33	51.42	<0.01	Mean RR	-0.24	-2.14	<0.05
						Mean BR	0.32	4.4	<0.01
						G	-0.33	-2.83	<0.01
						M	0.82	9.17	<0.01
						P	-0.49	-3.84	<0.01
						V	0.81	7.58	<0.01
	3	0.915	0.29	103.78	<0.01	Pupil Diameter	-0.13	-2.39	<0.05
						B	-1.59	-16.33	<0.01
						G	-0.66	-8.84	<0.01
						H	1.34	18.46	<0.01
	4	0.856	0.37	57.57	<0.01	B	-1.31	-12.4	<0.01
						F	0.26	2.61	<0.05
						M	0.58	7.03	<0.01
						V	0.21	2.87	<0.01
	5	0.692	0.55	29.51	<0.01	Mean BR	0.28	2.54	<0.05
						F	-0.26	-2.82	<0.01
						M	0.85	7.51	<0.01
	6	0.763	0.48	25.6	<0.01	Mean RR	0.62	5.47	<0.01
						Pupil Diameter	-0.27	-2.7	<0.05
						B	1.47	5.89	<0.01
						F	-1.38	-7.35	<0.01
						P	-0.32	-2.35	<0.05
	7	0.787	0.46	24.49	<0.01	Mean RR	-0.26	-2.46	<0.05
						Mean BR	-0.3	-3.03	<0.01
						Pupil Diameter	-0.24	-2.34	<0.05
						L	0.37	2.43	<0.05
						M	1.51	6.85	<0.01
						P	-1.46	-5.01	<0.01
	8	0.712	0.53	24.5	<0.01	Pupil Diameter	-0.67	-6.56	<0.01
						G	-0.63	-2.8	<0.01
						M	1.65	4.83	<0.01
						P	-1.47	-7.38	<0.01
	9	0.841	0.39	51.43	<0.01	Mean BR	-0.2	-3.01	<0.01
						Pupil Diameter	-0.78	-7.49	<0.01
						G	0.34	3.17	<0.01
						P	-0.45	-5.58	<0.01
	10	0.7	0.54	30.55	<0.01	Pupil Diameter	-0.62	-6.54	<0.01
						G	0.49	5.01	<0.01
						V	0.26	2.53	<0.05

Table 5.12 Proportion of the variability accounted for by the regression model in the response variable using combination 3 of predictors

5.3 Discussion

The results presented in this chapter demonstrate that physiological monitoring can be used for non-invasive real-time measurement of workload, assuming models have been appropriately trained on previously recorded data from the user population. Facial thermography combined with measurement of pupil diameter are strong candidates for real-time monitoring of workload due to the availability and non-intrusive nature of current technology. The study also

Predictors	Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
Combination 4: Temperatures inside points: 'B' 'F' 'G' 'H' 'L' 'M' 'P' 'V'	1	0.708	0.54	31.71	<0.01	B	-0.94	-6.65	<0.01
						F	0.58	3.92	<0.01
						P	0.68	6.94	<0.01
	2	0.818	0.42	43.77	<0.01	G	-0.33	-2.63	<0.05
						M	1.01	10.1	<0.01
						P	-0.76	-5.22	<0.01
						V	0.55	4.8	<0.01
	3	0.903	0.3	120.15	<0.01	B	-1.55	-15.18	<0.01
						G	-0.64	-8.09	<0.01
						H	1.38	18.33	<0.01
	4	0.856	0.37	57.57	<0.01	B	-1.31	-12.4	<0.01
						F	0.26	2.61	<0.05
						M	0.58	7.03	<0.01
						V	0.21	2.87	<0.01
	5	0.679	0.56	27.81	<0.01	F	-0.55	-3.68	<0.01
						H	0.33	2.17	<0.05
						M	0.52	4.42	<0.01
	6	0.641	0.59	18.02	<0.01	B	0.79	2.61	<0.05
						F	-1.5	-6.66	<0.01
						G	-0.93	-4.25	<0.01
						L	1.14	5.73	<0.01
	7	0.724	0.52	20.97	<0.01	B	-0.49	-3.27	<0.01
						G	-0.97	-3.88	<0.01
						L	1.08	4.32	<0.01
						M	1.27	5.04	<0.01
						P	-0.83	-3.07	<0.01
	8	0	-	-	-	-	-	-	-
	9	0.564	0.65	25.63	<0.01	G	0.89	6.91	<0.01
						P	-0.3	-2.33	<0.05
	10	0.574	0.65	18.1	<0.01	G	1.29	6.71	<0.01
						L	-1.21	-3.9	<0.01
						P	0.58	2.46	<0.05

Table 5.13 Proportion of the variability accounted for by the regression model in the response variable using combination 4 of predictors

demonstrates the importance of identifying whether an individual is one who demonstrates a strong relationship between physiological measures and experienced workload measures before physiological measures are applied uniformly. This is a feasible proposition in settings such as aircraft cockpits, where pilots are drawn from a relatively small, targeted and managed population.

This research presents novel insights into the relative value of physiological and subjective techniques for assessment of workload and human performance. The main novelty lies in the fact that multiple continuous physiological measures were recorded and synchronized with task performance and subjective ratings. The hypotheses explored in this study were:

1. There will be a measurable difference in subjective workload between the two levels of task difficulty This hypothesis was found to be true: The mental demand measured using NASA-TLX confirmed that there was a measurable difference between the two

levels of difficulty and stage 2 was perceived to be more mentally demanding than stages 1 and 3.

2. The subjective ratings of workload will be associated with changes in physiological measures Hypothesis 2 was partially supported: The study explored which physiological measures showed a change in accordance to the change in mental workload as measured subjectively on the ISA scale. It was found that for some of the participants, the mean normalized ISA ratings showed a stronger correlation with some of their physiological measures than it did with the individual ISA rating. Table 5.14 summarizes the results by displaying the number of participants that showed moderate to strong correlations with mean ISA normalized or individual ISA ratings for each of the physiological measures presented above. Overall, the correlations of the thermal data with the individual ISA ratings were weaker for all participants.

Measure	No. of participants showing moderate to strong correlations	
	Mean ISA Normalized	Individual ISA
R-R Intervals	3/10	2/10
Breathing Rate	1/10	1/10
Pupil Diameter	8/10	7/10
Point P Temperature	3/10	3/10
Point V Temperature	1/10	1/10
Point L Temperature	4/10	1/10
Point M Temperature	6/10	3/10

Table 5.14 No. of participants showing moderate to strong correlations with the ISA rating

3. Multiple physiological measures can be used in combination to analyze workload Hypothesis 3 was tested by using a multiple linear regression on the data from each of the participants, showing that when using facial thermography data are combined with other physiological data, the predictive model explains on average 47.7% more of the variability in performance compared to solely using a combination of R-R inter-beat intervals, breathing rate and pupil diameter. As mean performance across the participants was strongly correlated with the mean ISA normalized, it is an indication that these physiological measures could also provide good prediction results for the level of subjectively experienced mental workload.

In their discussion section, Ora & Duffy (2007) [5] recommended that further examination under more controlled conditions and the test of additional psycho physiological measures such as pupil dilation should be performed in the hope of developing a more robust approach to the estimation of mental workload in a non-invasive way. In this study, the variation

of demand was done in more controlled conditions and additional physiological measures (such as heart rate, breathing rate and pupil diameter) were collected and their relative group contribution was tested. In terms of facial thermography, the landmark tracking was done automatically and included more areas of the face. One of the limitations of the study was the small number of participants; for the limited number of participants (10), there was no physiological measure that proved to work best at predicting mental workload or performance levels across all participants. Although from a physiological point of view people responded differently when being subjected to the type of demand induced by the task, some of the physiological measures, especially pupil diameter and temperatures in points G, M and P, proved to be good and consistent indicators of the level of performance (and implicitly the level of demand) for more than half of the participants. Further studies will concentrate on the collection of more data in environments closer to the real workplace setting and the use of machine learning algorithms to improve prediction accuracy, confirm feasibility of applying the physiological and analytical methods in situ, and ensure generalisability of results. Future work should also consider how facial thermography measurements would vary over longer time periods than have been examined in this study.

5.4 Chapter Summary

The results presented in this chapter demonstrate that physiological measures, especially face temperature and pupil diameter, can be used for non-invasive real-time measurement of workload when combined with a facial landmark tracking algorithm, assuming models have been appropriately trained on previously recorded data from the user population. This is a feasible proposition in a setting such as cockpits. This chapter has addressed research questions 1, 2 and 3:

1. How do human physiological responses change in response to variations in task demand and task performance? This was explored by recording multiple channels of physiological data while the participant was exposed to a task eliciting a predefined pattern of demand.

2. How are physiological responses associated with variations in subjective reports of mental workload? This was tested by comparing the response of multiple physiological data pieces with the subjective ratings of mental workload.

3. Can multiple combined physiological parameters explain more of the variability in mental workload or performance than individual parameters? This was explored by

performing a multiple linear regression on the data from each of the participants, showing that using facial thermography data improves the predictive model.

The next chapter will focus on testing the fourth research question: How do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator? In order to test this, a flight simulator study was planned in a high fidelity helicopter simulator, having as participants active helicopter pilots.

Chapter 6

Physiological Measures of Workload in a Flight Simulator

6.1 Introduction

The aim of this chapter is to examine the physiological response of highly trained helicopter pilots to variations in task demand. The study presented in the previous chapter examined the relationship between experienced workload and physiological response by no-invasive monitoring of physiological parameters and was performed in laboratory conditions. The study presented in this chapter aims at using non-invasive physiological measures in a highly realistic environment.

Methods: The study presented in this chapter was performed in a highly realistic helicopter flight simulator and had as participants active helicopter pilots. The task consisted of flight scenarios of three levels of difficulty and the collected data consisted of: physiological measurements (heart inter-beat intervals, breathing rate and facial thermography) and ISA subjective ratings of workload.

Results: The study demonstrated that it is feasible to use physiological measures such as facial thermography in aircraft cockpits and that the changes in physiology shown by trained helicopter pilots in response to changes in demand were similar to the ones experienced by the general population and presented in the previous chapter.

6.2 Study Design

The study presented in this chapter explores the changes of physiological parameters that occur as the level of workload varies. As opposed to the other studies presented in this thesis,

this study was performed in an ecologically valid aircraft simulator and the participants were highly trained active pilots.

The tasks were designed in collaboration with a flight training instructor who planned specific scenarios that exposed the participants to three different levels of demand.

The independent variable that was manipulated during the study was task difficulty. The dependent variables were the physiological measures and the subjective assessment of the perceived level of workload. In order to avoid interrupting the simulation, the Instantaneous self-assessment workload scale (ISA) was used [22]. For the other studies presented in this thesis, ISA ratings were sampled at fixed 45s time intervals; in the case of the simulator study, the pilots could not be interrupted mid-scenario when they were engaged in important communication, for example, to diagnose a failure. For this reason, the ISA ratings were sampled at convenient times so as not to interfere with the scenario.

6.2.1 Study Task

The study tasks were designed together with a flight training instructor and aimed at varying the level of task difficulty in three stages: Easy, Medium and Difficult. The helicopter was flown by two pilots, one having the role of flying pilot and the second one of copilot; data were collected only from the flying pilot. In order to make the most of the time and resources, after performing one task, composed of a series of scenarios of different difficulty levels, the participants swapped places. This way, the copilot became the flying pilot; the problem with this approach was that the newly appointed flying pilot was already used to the scenarios. For this reason, two types of very similar tasks were designed, following the same pattern in variation of difficulty; from here on they will be referred to as Type I and Type II tasks. Type I and Type II tasks were meant to be composed of scenarios of the same difficulty and following the same pattern. As they were not perfectly identical, it was decided to treat each of them separately. Each task consisted of 9 scenarios alternating in difficulty scenarios.

The level of difficulty of each scenario varied in time according to the pattern described in Fig.6.1, the black horizontal lines symbolically represent the scenario duration in time. Notations t_1 to t_9 were used to denote their duration while the time intervals between the scenarios used notations of the format $t_{scenario-before-scenario-after}$. The same notations were used in Tables 6.1 and 6.2 to describe the task durations for both Type I and Type II scenarios.

The duration of the scenarios varied for each of the participants depending on how they approached solving the task. Figures 6.2 and 6.3 show box plots of the time taken for the participants to complete each of the scenarios in both the Type I and Type II tasks. The circular markers in each of the box plots represent the actual time for each of the participants.

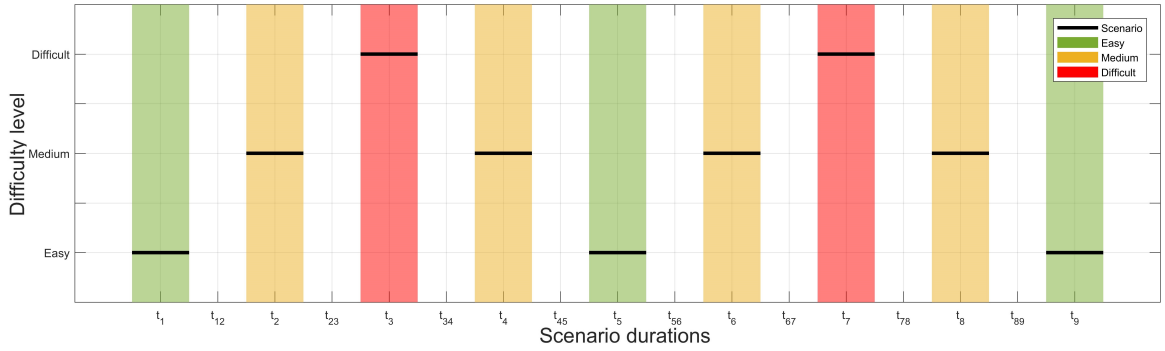

Fig. 6.1 Description of conditions

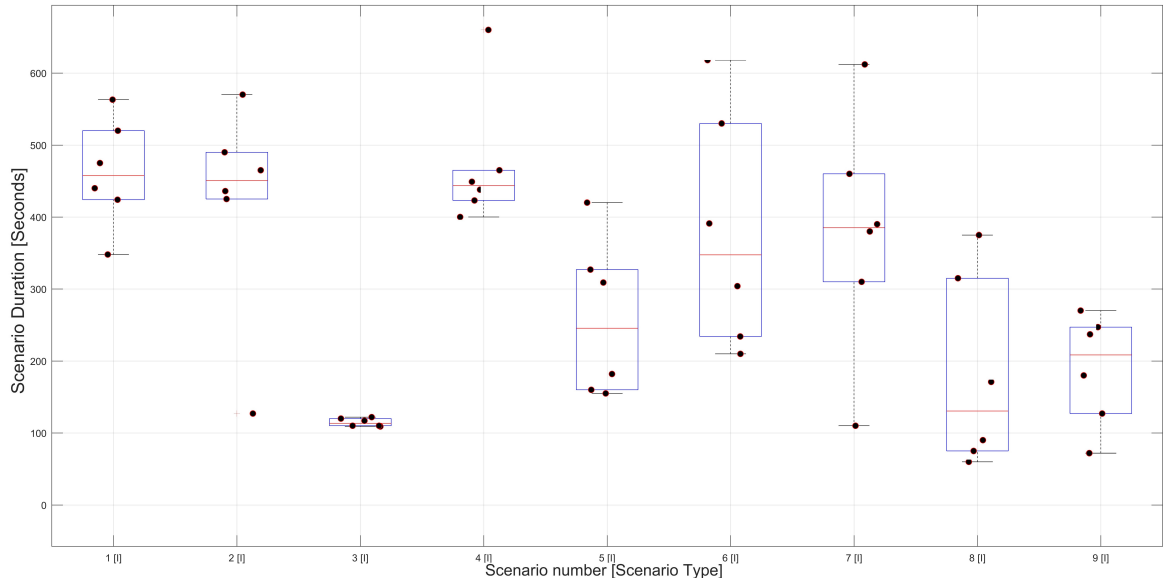

Fig. 6.2 Duration boxplot of Type I Scenarios.

Time Duration	Average Duration	Min Duration	Max Duration	SD Duration
t_1	7m 42s	5m 48s	9m 23s	1m 16s
t_{12}	3m 11s	2m 22s	4m 21s	46s
t_2	6m 59s	2m 7s	9m 30s	2m 32s
t_{23}	1m 27s	59s	2m 43s	40s
t_3	1m 55s	1m 49s	2m 2s	6s
t_{34}	2m 48s	2m	4m 44s	59s
t_4	7m 52s	6m 40s	10m 60s	1m 35s
t_{45}	1m 32s	1s	4m 41s	1m 37s
t_5	4m 19s	2m 35s	7m	1m 49s
t_{56}	1m 6s	1s	2m 40s	52s
t_6	6m 21s	3m 30s	10m 18s	2m 44s
t_{67}	1m 52s	1m 33s	2m 14s	18s
t_7	6m 17s	1m 50s	10m 12s	2m 46s
t_{78}	3m 10s	1m 5s	11m 5s	3m 55s
t_8	3m 1s	1m	6m 15s	2m 14s
t_{89}	1m 12s	45s	3m	53s
t_9	3m 9s	1m 12s	4m 30s	1m 17s

Table 6.1 Type 1 Scenario Durations. Symbols t_n represent the duration of the nth scenario while t_{nm} represent the time duration between the nth and the mth scenarios

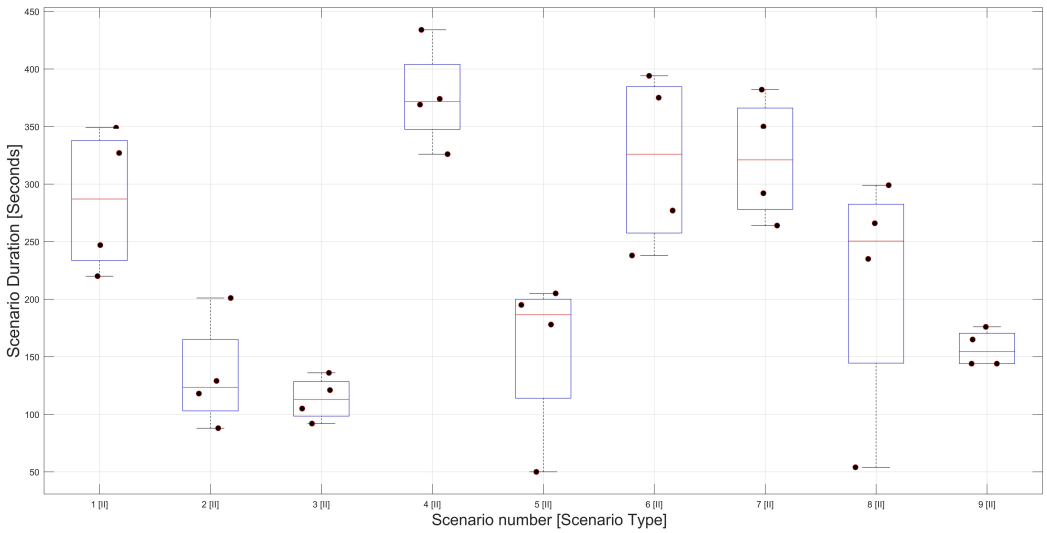

Fig. 6.3 Duration boxplot of Type II Scenarios

Time Duration	Average Duration	Min Duration	Max Duration	SD Duration
t_1	4m 46s	3m 40s	5m 49s	1m 2s
t_{12}	1m 52s	1m 12s	2m 60s	51s
t_2	2m 14s	1m 28s	3m 21s	48s
t_{23}	1m 21s	56s	2m 4s	30s
t_3	1m 53s	1m 32s	2m 16s	19s
t_{34}	2m 14s	2m 5s	2m 23s	9s
t_4	6m 16s	5m 26s	7m 14s	44s
t_{45}	1m 17s	41s	1m 43s	27s
t_5	2m 37s	50s	3m 25s	1m 12s
t_{56}	2m 57s	2m 27s	3m 25s	28s
t_6	5m 21s	3m 58s	6m 34s	1m 15s
t_{67}	1m 46s	1m 21s	2m 18s	24s
t_7	5m 22s	4m 24s	6m 22s	54s
t_{78}	2m 17s	1m 47s	2m 50s	26s
t_8	3m 33s	54s	4m 59s	1m 49s
t_{89}	1m 21s	45s	2m 32s	48s
t_9	2m 37s	2m 24s	2m 56s	16s

Table 6.2 Type 2 Scenario Durations. Symbols t_n represent the duration of the n^{th} scenario while t_{nm} represent the time duration between the n^{th} and the m^{th} scenarios

The scenarios were pre-programmed into the simulator but manually started by the instructor when the pilots were ready. In case an emergency occured during the scenario, the pilots were instructed to first deal with the emergency; if on approach, continue approach, do not go around and do not use upper modes unless told they are available. The scenarios are briefly described below:

Type I Task:

- Scenario 1 (Easy): Flight to oil rig with wind change
- Scenario 2 (Medium): IFR (Instrument Flight Rules) approach with coupled engine failure. Autopilot allowed
- Scenario 3 (Difficult): Double engine failure
- Scenario 4 (Medium): Collective Trim Failure with autopilot allowed. The collective changes the pitch angle of all main rotor blades collectively. In levelled flight, changing the collective would cause a climb or a descent whereas if the helicopter is pitched forward a change in collective would induce forward movement and ascent.
- Scenario 5 (Easy): Landing gear fails to extend normally

- Scenario 6 (Medium): Take-off with governor failure. The governor is a system that controls engine power in order that the rotor maintain constant RPM.
- Scenario 7 (Difficult): Tail rotor failure. In this scenario, due to the failure of the tail rotor , the pilots would have to disengage the engine and land using the autorotation manoeuvre. This is an extremely demanding procedure in which the pilots have to work together to land the helicopter without the engine and requires very good synchronization for flaring (raising the nose) the aircraft at just the right moment so that the air rushing through the blades cushions the landing.
- Scenario 8 (Medium): Engine failure
- Scenario 9 (Easy): Engine failure hover

Type II Task:

- Scenario 1 (Easy): VFR (Visual flight rules) take-off landing gear will not go up
- Scenario 2 (Medium): Engine failure before TDP (take-off decision point)
- Scenario 3 (Difficult): Double engine failure
- Scenario 4 (Medium): IFR (Instrument Flight Rules) approach manual, bad weather
- Scenario 5 (Easy): Engine failure hover
- Scenario 6 (Medium): ARA (Airborne Radar Approach) with weather radar failure. Autopilot allowed
- Scenario 7 (Difficult): MGB (Main Gear Box) shaft failure at night over water. Autopilot allowed
- Scenario 8 (Medium): Governor failure after TDP (take-off decision point). Autopilot allowed
- Scenario 9 (Easy): EID failure

6.2.2 Study Protocol

The participants were welcomed into a waiting room where they were explained the purpose of the study and invited to read the information sheet, describing the details of the study, and

then fill in a consent form. The participants were then invited to attach the Zephyr sensor on, in a private space.

Once the Zephyr sensor was attached and recording data, the participants were offered headphones (required for in-flight communications) and walked into the simulator.

The scenarios were controlled by the instructor. I was seated in the back seat of the helicopter monitoring the recording. At various time intervals I asked the flying pilot for the ISA rating over the communication system by saying the phrase we had agreed on before the flight: 'Level please'. The start and end times of the scenarios were recorded as well as the times of the ISA ratings.

6.2.3 Measurements and Equipment

The study took place in a high fidelity Eurocopter EC225 helicopter simulator in Aberdeen (Fig.6.4). The simulator has a 210 degree horizontal field of view, a +30 degree above horizon and -50 degree below horizon. The simulator also has a six degree of freedom motion capability with vibration platform and sound [79]. An inside view of the cockpit and the positioning of the thermal camera can be seen in Fig.6.5).

Fig. 6.4 Eurocopter EC225 Helicopter Flight Simulator

The data collected during the study were of two types:

1. Subjective data:

Fig. 6.5 Eurocopter EC225 Helicopter cockpit view and thermal camera positioning

(a) Instantaneous Self Assessment (ISA) workload scale

2. Physiological data collected using the following pieces of equipment[1]:

(a) Zephyr Bioharness 3

(b) FLIR A65sc Thermal Camera

6.2.4 Study Hypothesis

The study hypotheses were that:

1. There will be a measurable difference in subjective workload ratings between the three levels of task difficulty

2. The physiological parameters will show sufficient changes as to differentiate between the three different levels of demand imposed

3. The physiological response will be similar to non-trained individuals performing a laboratory based task of varying levels of difficulty. This third hypothesis addresses the fourth research question of the thesis: "How do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator?" Mental workload, as a multidimensional construct, is influenced by a multitude of factors such as physical and cognitive task demands, performance and external and internal influences. Skill, training level and experience fall in the category of internal influences. This hypothesis is exploring whether someone who has undergone extensive training in dealing with

[1]The equipment and data analysis are described in Chapter 4

extremely demanding situations will show changes in the physiological measures that were used.

6.2.5 Participants

Eight participants were recruited by Airbus Helicopters to take part in the study, all of them active helicopter pilots (two of them performed both scenarios). All participants had previously undergone training in this flight simulator and were familiar with the equipment. The participants were all men with a mean age of 42.7 years, SD = 8.1; range 30-51. In terms of flying experience the mean number of flight hours was 7500, SD = 4898; range 2000-16500.

The study was approved by the Faculty of Engineering Ethics Committee. Each participant was presented with an information sheet and consent form, stating that they do not suffer from motion or travel sickness, migraine, epilepsy, dizziness and blurred vision. They were informed that they could withdraw at any moment without giving any reason.

The data were recorded only from the flying pilot. Each participant played the role of flying pilot on one of the two scenarios and copilot on the other as described in Table 6.3.

Crew	Type 1 Scenario		Type 2 Scenario	
	Flying pilot	Copilot	Flying pilot	Copilot
1	Participant 1	Participant 2	Participant 2	Participant 1
2	Participant 3	Participant 4	-[2]	-
3	Participant 2	Participant 6	Participant 6	Participant 2
4	Participant 7	Participant 1	Participant 1	Participant 7
5	Participant 11	Participant 10	-	-
6	Participant 9	Participant 10[3]	Participant 10	Participant 9

Table 6.3 Participant roles for each of the scenarios

6.3 Results

In this section the results of both subjective and physiological data will be presented in several stages. Firstly the results of the inferential tests to examine the impact of the manipulation of the task demand on the subjective measures of workload, aiming to confirm that the demand manipulation affected workload in the anticipated manner. The second part of the analysis focuses on the physiological data and how they were affected by the manipulation of task demand. In the final part of the analysis, individual observations are made on some of the

pieces of data as well as comparisons to the results obtained in the study presented in Chapter 5.

6.3.1 Subjective data

ISA subjective ratings were recorded during the tasks. The number of ISA sampling points differs from scenario to scenario and participant to participant as it was not possible to interrupt the pilots at any time. Figures 6.6 and 6.7 below show a box plot description of the mean ISA ratings for each of the nine scenarios of both Type I and Type II tasks. This is reported in order to show that the general variation of demand perceived by the participants is very similar to the intended variation in demand Fig.6.1.

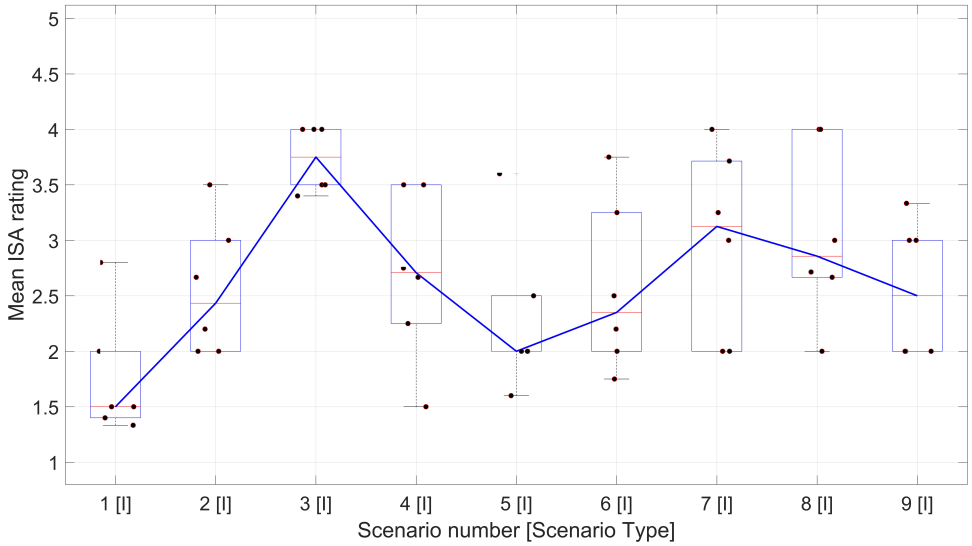

Fig. 6.6 Mean ISA level for each scenario of Type I

As the demand also varied during each of the scenarios, Figures 6.8 and 6.9 show the maximum ISA rating during each of the scenarios; in this case as well it can be observed that the perceived variation in demand is similar to Fig.6.1.

A one way ANOVA ($F(2,185) = 23.68$, $p < 0.01$, $\eta^2 = 0.2$) was performed on the ISA ratings grouped in three groups by the level of difficulty of the scenario (easy, medium, difficult) for the Type I task as well as for the Type II task ($F(2,109) = 31$, $p < 0.01$, $\eta^2 = 0.2$). A multiple test comparison using the Bonferroni method was used to determine which type of scenario makes a difference in the perceived demand. The test revealed that each of the scenarios showed a significant difference from the others Fig. 6.10. Tables 6.4, 6.5 contain the results of the test.

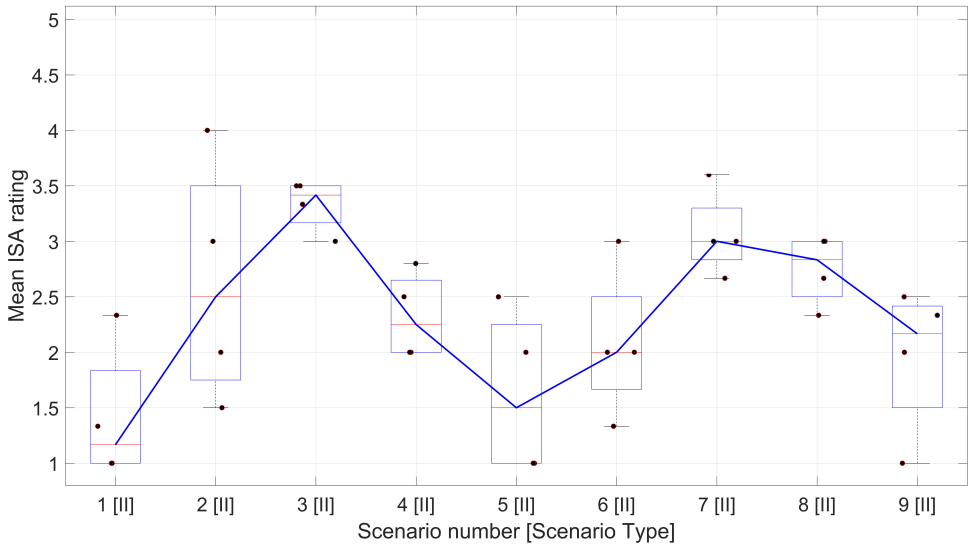

Fig. 6.7 Mean ISA level for each scenario of Type II

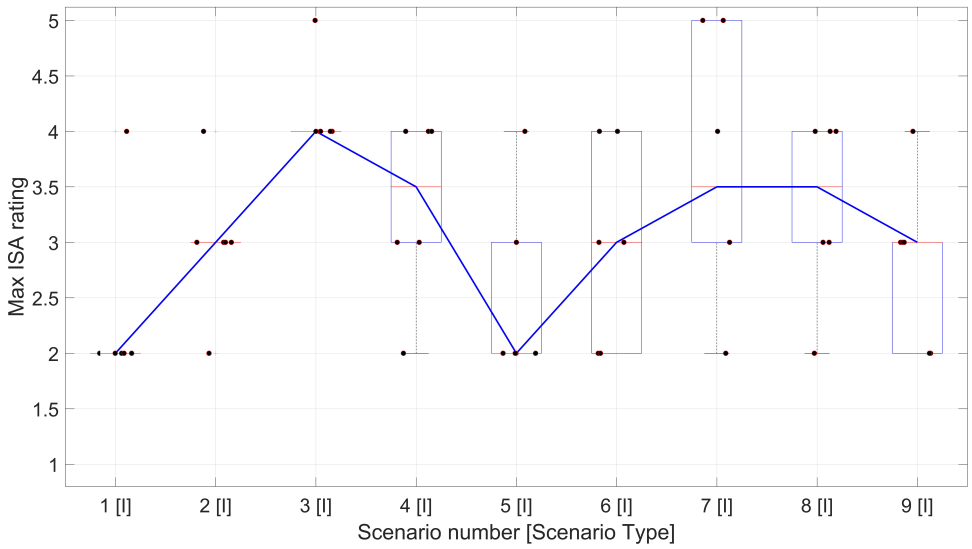

Fig. 6.8 Max ISA level for each scenario of Type I

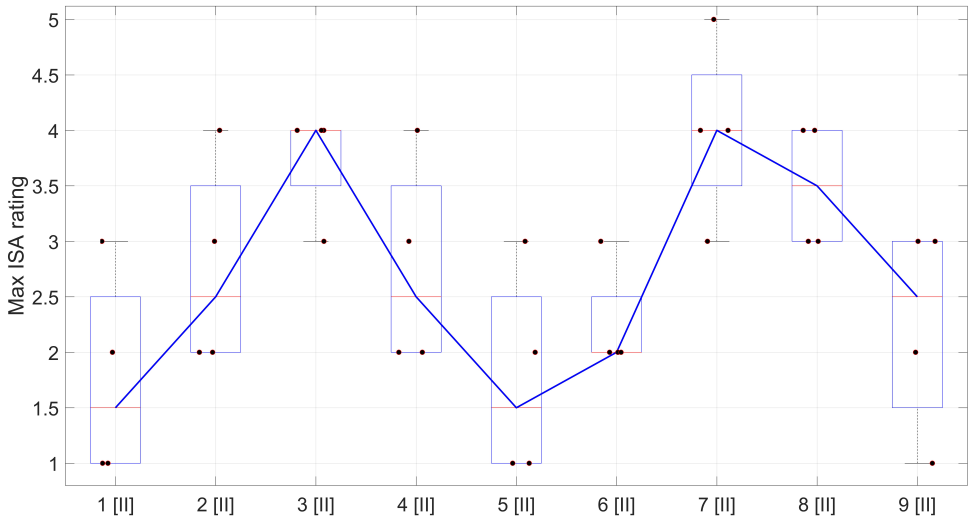

Fig. 6.9 Max ISA level for each scenario of Type II

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	-0.4643	<0.05	-0.8153	-0.1134
Easy - Difficult	-1.208	<0.05	-1.6325	-0.7834
Difficult - Medium	-0.7436	<0.05	-1.1343	-0.353

Table 6.4 Multiple Comparison Test Results for the Type I task

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	-0.6599	<0.05	-1.0930	-0.2268
Easy - Difficult	-1.5152	<0.05	-1.9864	-1.0439
Difficult - Medium	-0.8553	<0.05	-1.2760	-0.4346

Table 6.5 Multiple Comparison Test Results for the Type II task

This confirms ***Hypothesis 1***, that there will be a measurable difference in subjective workload ratings between the three levels of task difficulty.

6.3.2 Physiological data

The physiological data collected consisted of heart rate, breathing rate and facial skin temperatures measured by thermography. One of the challenges in reporting the data comes from the fact that neither the subjective rating sampling times nor the time duration of the levels were fixed. Another challenge is that there were no data with regards to the exact

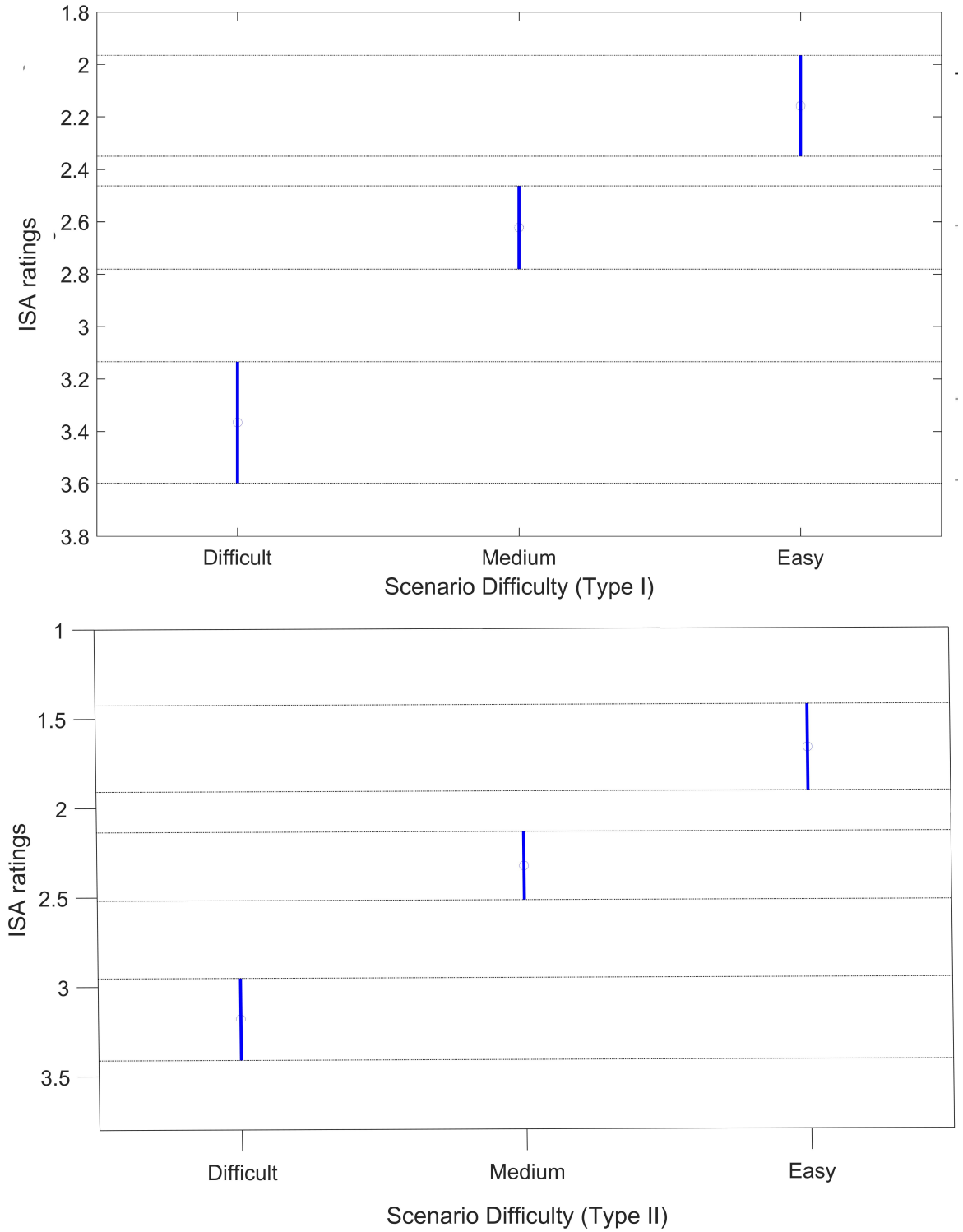

Fig. 6.10 Multiple comparison between difficulty levels for both types of tasks in terms of ISA ratings

timing of when an emergency occurred during a scenario, for this reason, during some parts of some scenarios, the physiological data show no changes.

As physiological data depend so much on the physiology of the participants, the range of values for the same physiological measure can vary between participants. In order to overcome this and be able to compare physiological changes between participants, the z-score of the data was computed. The z-score was computed ignoring the physiological data recorded in-between scenarios.

The results will be presented for each physiological measure: heart rate, breathing rate and facial thermography. Although a large number of facial areas were considered, not all of them were tracked with the same accuracy, causing some of the resulting thermal data to be noisy. Considering this, a subset of the data will be presented. The data will also be presented separately for Type I and Type II tasks.

Heart Rate

Figure 6.11 shows the heart rate time series for participant 1 during the Type I task as an example of what the data look like. The colours overlaid on the heart rate signal symbolise the duration of the scenarios and also the difficulty. The black dots represent the ISA ratings (right axis) and the time they were sampled.

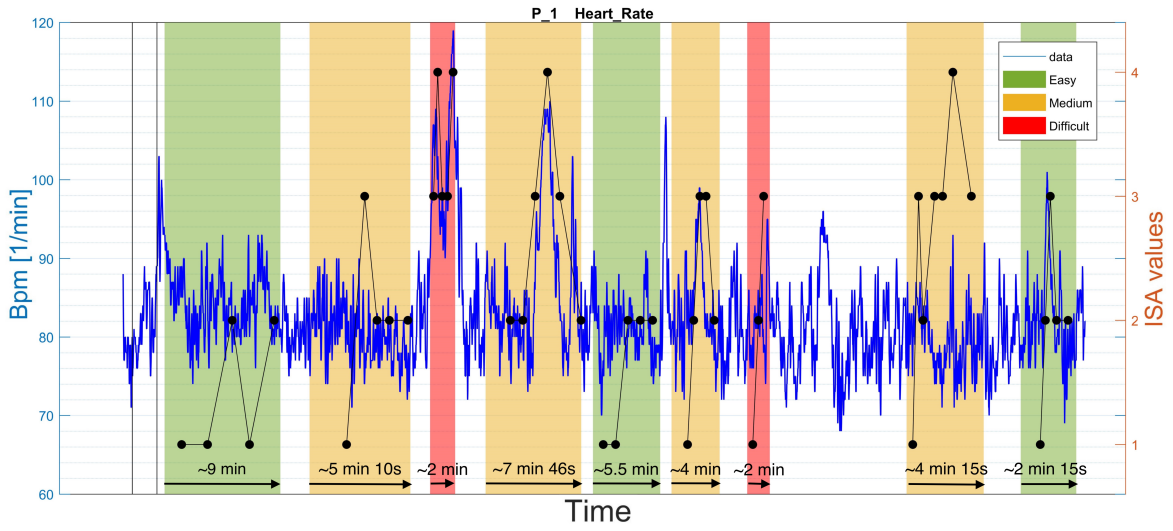

Fig. 6.11 Heart rate example data: Participant 1 - Type I task

Figures 6.12 shows box plots of the mean heart rate z-score for both participants performing the Type I task (left) and Type II task (right). It can be observed that the relative pattern of variation in demand is maintained. One thing to observe is that there was a stronger heart

rate response to Scenario 3 of Type I, which can also be observed in the mean ISA ratings Fig.6.6.

A one way ANOVA (F(2,51)=21.1, p<0.01, $\eta^2 = 0.46$) was performed on the average z-scores of the heart rate data for the participants performing Type I task as well as for the participants performing the Type II task (F(2,33)=4.76, p=<0.01, $\eta^2 = 0.44$), grouped by the three levels of difficulty (easy, medium, difficult). A multiple comparison test using the Bonferroni method was used to test if the means of the groups are significantly different. For both the Type I (Fig. 6.13 left) task and Type II (Fig. 6.13 right) task, the difficult scenarios had heart rate z-score means that were significantly different from the ones during the easy and medium scenarios. Tables 6.6 and 6.7 contain the results of the test.

In some cases, averaging the z-score over the entire scenario might not tell the whole story as demand varied during each of the scenarios. For example, in Fig. 6.11, Scenario 4, there is an increase in heart rate that seems to be reflected also in the increase of the ISA ratings. In order to capture some of these changes, Table 6.8 reports the variation in the z-score (ΔS expressed in number of standard deviations from the mean) of such large changes as well as mean heart rate during the scenario.

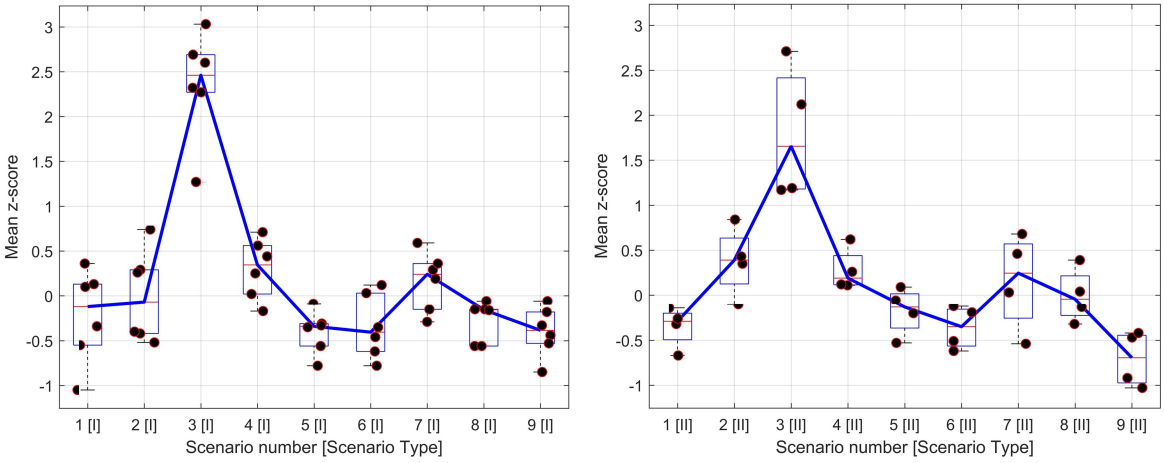

Fig. 6.12 Heart rate variation across the Type I (left) and Type II (right) scenarios

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	-0.2614	0.66	-0.7833	0.2605
Easy - Difficult	-1.6064	<0.05	-2.2302	-0.9826
Difficult - Medium	-1.345	<0.05	-1.9368	-0.7532

Table 6.6 Multiple Comparison Test results between difficulty levels for the Type I task and mean Heart Rate z-score

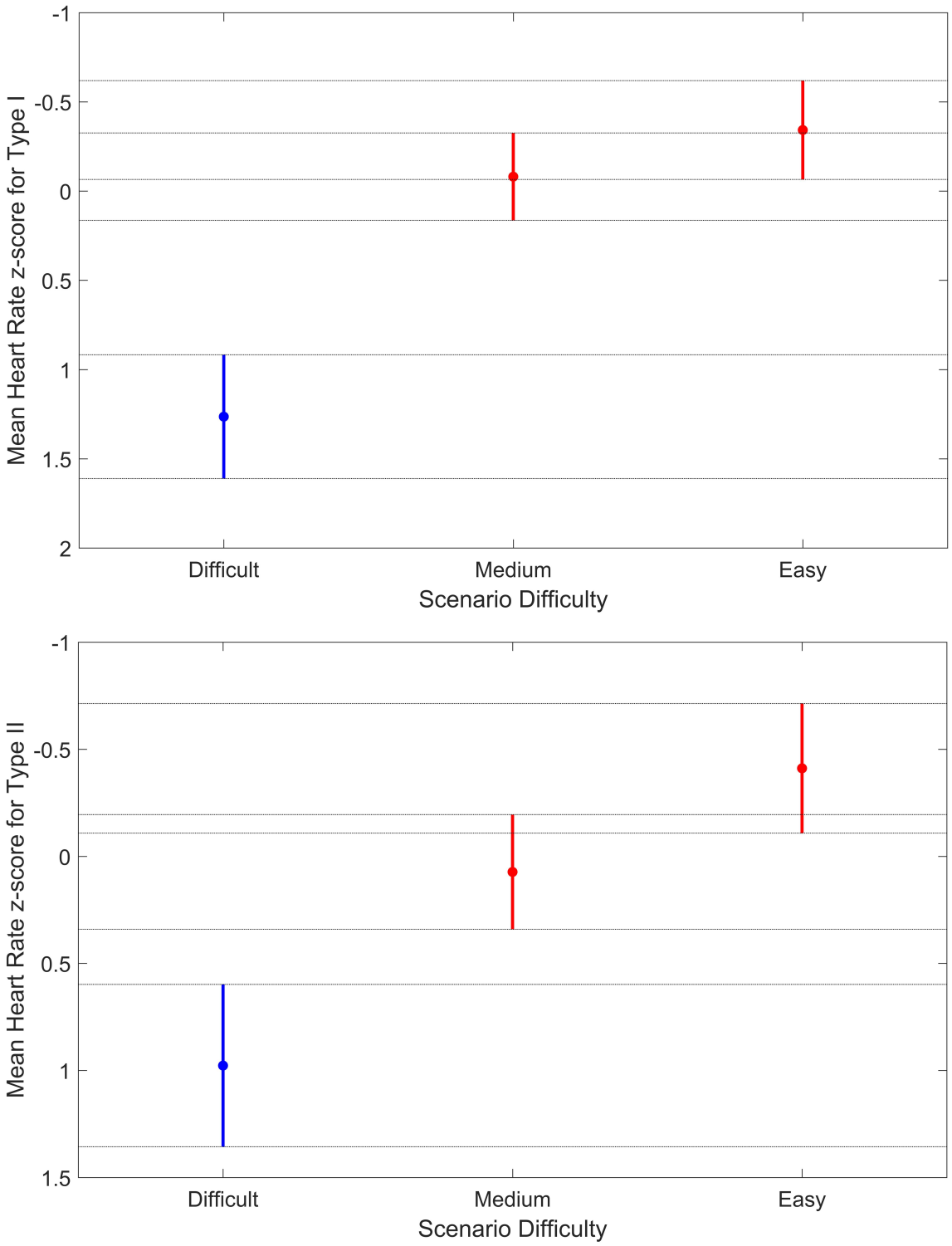

Fig. 6.13 Multiple comparison between difficulty levels and mean Heart Rate z-score

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	-0.484	0.119	-1.0545	0.0866
Easy - Difficult	-1.3883	<0.05	-2.0703	-0.7064
Difficult - Medium	-0.9044	<0.05	-1.5513	-0.2574

Table 6.7 Multiple Comparison Test results between difficulty levels for the Type II task and mean Heart Rate z-score

Hear Rate (bpm)	Measure	P1	P2	P3	P5	P6	P7	P8	P9	P10	P11
1. Easy	Mean	84.12	68.45	104.08	76.81	57.4	72.86	77.54	84.47	81.26	69.09
	ΔS	-	-	-	4.13	2.89	3.01	-	1.73	-	-
2. Medium	Mean	80.37	70.69	110.17	76.36	58.44	78.49	80.83	86	96.22	68.06
	ΔS	-	-	-	-	4.55	2.69	3.35	-	2.11	-
3. Difficult	Mean	102.57	78.85	114.59	86.56	64.56	93.16	84.45	113.79	114.78	88.86
	ΔS	3.22	3.03	3.96	4.82	4.55	4.84	3.56	5.57	1.51	4.26
4. Medium	Mean	86.53	69.58	112.06	74.49	60.16	73.93	79.33	90.31	94.04	74.4
	ΔS	4.76	3.9	-	-	2.07	2.79	3.14	2.97	-	-
5. Easy	Mean	80.86	69.48	107.43	71.86	57.94	68.57	76.24	84.26	87.32	70.6
	ΔS	-	-	-	3.44	-	-	-	-	-	-
6. Medium	Mean	83.64	68.22	106.03	75.74	56.45	64.33	78.19	86.5	81.77	68.42
	ΔS	2.38	-	-	3.44	5.79	-	-	-	-	-
7. Difficult	Mean	82.35	69.22	110.18	73.97	61.15	73.37	82.03	96.28	82.6	73.24
	ΔS	2.94	-	2.86	-	3.72	-	4.82	3.84	-	3.58
8. Medium	Mean	79.4	68.47	108.17	72.8	59.09	70.24	77.25	88.47	91.74	70.77
	ΔS	-	-	-	-	3.93	2.79	-	-	1.91	2.04
9. Easy	Mean	82.08	64.31	107.33	74.96	56.63	63.71	74.38	85.55	83.71	68.02
	ΔS	3.5	-	-	-	2.07	2.36	-	-	-	-

Table 6.8 Heart rate variations showing the mean heart rate for each of the scenarios for each participant as well as ΔS that in this case denotes the number of standard deviation from the mean that were observed during high demand periods; if no such changes were observed the field would be left empty

Breathing Rate

Figure 6.14 shows the breathing rate time series for participant 1 during the Type I task as an example of what the data look like. The colours overlaid on the breathing rate signal symbolise the duration of the scenarios and also the difficulty. The black dots represent the ISA ratings (right axis) and the time they were sampled.

Figures 6.15 shows box plots of the mean breathing rate z-score for both participants performing the Type I task (left) and Type II task (right). It can be observed that the relative pattern of variation in demand is not maintained as well as for the heart rate data. For the Type I task, only scenarios 2,3 and 4 show a pattern similar to the one shown by the ISA ratings and the level of difficulty whereas for the Type II task, scenarios 2, 3, 4 and 6,7, 8

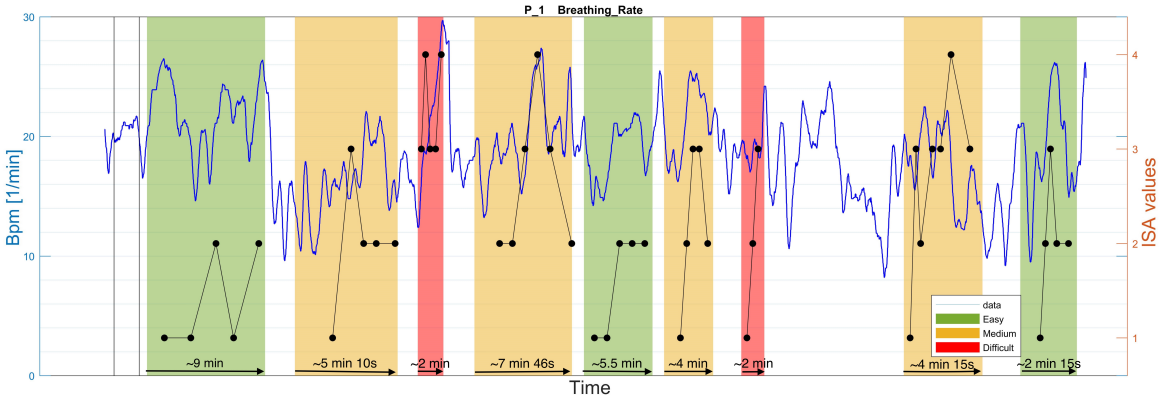

Fig. 6.14 Breathing rate example data: Participant 1 - Type I task

show these patterns. As in the case of the heart rate, scenario 3 showed the strongest response for both task types.

A one way ANOVA (F(2,51)=2.82, p=0.07, $\eta^2 = 0.09$) was performed on the average z-scores of the breathing rate data for the participants performing Type I task as well as for the participants performing the Type II task (F(2,33)=4.52, p=0.018, $\eta^2 = 0.21$), grouped by the three levels of difficulty (easy, medium, difficult). For both Type I and II levels, the p value for the F-statistic was not small enough to indicate the group means were significant. A multiple comparison test using the Bonferroni method to test if the means of the groups are significantly different was used. For the Type I task (Fig. 6.13 left) none of the groups were significantly different from each other. For the Type II task (Fig. 6.16 right), the difficult scenarios had breathing rate z-score means that were significantly different from the medium scenarios, as it can also be observed in Fig. 6.15. Even though not statistically significant, there could be other reasons for these results, including that breathing rate could be sensitive to other aspects of mental workload compared to the subjective measures or it could have been affected by the participant using radio communication to accomplish the task, as speach will influence breathing rate.

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	0.3018	0.3375	-0.1608	0.7645
Easy - Difficult	-0.1686	1	-0.7216	0.3844
Difficult - Medium	-0.4704	0.0927	-0.9950	0.0542

Table 6.9 Multiple Comparison Test results between difficulty levels for the Type I task and mean Breathing Rate z-score

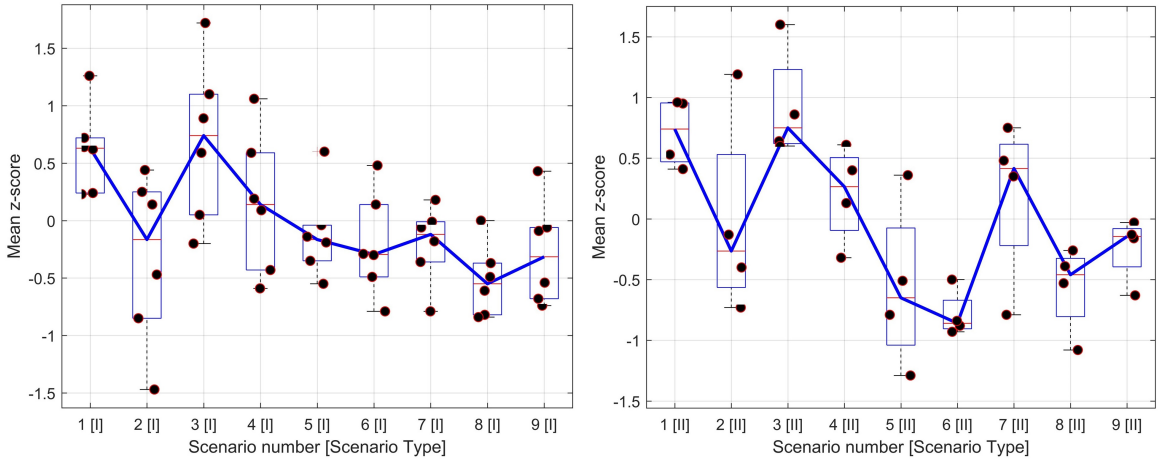

Fig. 6.15 Breathing rate variation across the Type I (left) and Type II (right) scenarios

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	0.2637	0.8978	-0.3670	0.8945
Easy - Difficult	-0.5887	0.1720	-1.3427	0.1652
Difficult - Medium	-0.8525	0.0151	-1.5677	-0.1373

Table 6.10 Multiple Comparison Test results between difficulty levels for the Type II task and mean Breathing Rate z-score

It needs to be mentioned that during the scenarios, the pilots were ingaged in radio communications which might have affected both heart rate and respiration. This is one of the challenges of performing studies in high fidelity simulators.

Facial Thermography

This section will discuss the facial thermography results in a similar manner as to the heart and breathing rate data. The facial landmarks used to extract the thermal data are described in Chapter 4; Fig. 4.27 shows the landmarks that were tracked for both non-glasses and glasses wearing participants. The left side of Table 4.5 lists the regions of interest that temperature was extracted from in terms of points, lines and areas for the Aberdeen study; the bolded ones were common for both non-glasses and glasses wearing participants.

In general, landmarks found around the central area of the face (e.g. nose area) were more accurately tracked as opposed to landmarks on the sides of the face which were more influenced by head movement. In this analysis, a subset of regions of interest from each of the following facial areas was selected: nose area, cheek area and forehead area. The data that will be presented are the mean temperature inside point 30, represented in red in the

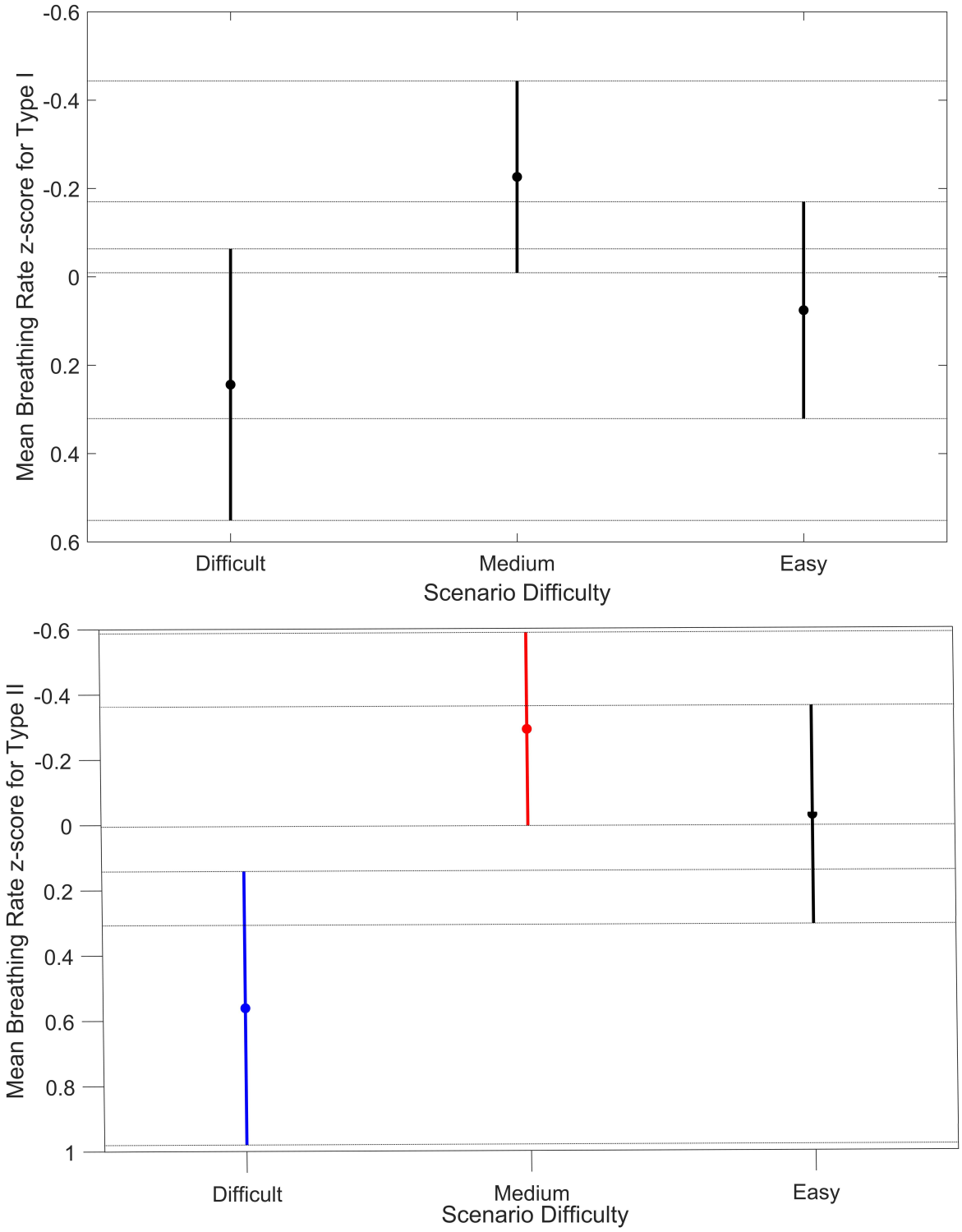

Fig. 6.16 Multiple comparison between difficulty levels and mean Breathing Rate z-score

left hand side of Fig. 6.17, on the nose as it was better tracked and as the nose temperature showed the most clear changes. The other areas, points and lines near the nose show similar results while the other areas of the face do not show such consistent changes to variations of demand because the changes are very small (this will be examined towards the end of this chapter), because there are no significant changes or the tracking in the area was not accurate enough.

To show an example of what the thermal data look like, Fig. 6.17 represents the average temperature for point 30 for participant 1 during a Type I task.

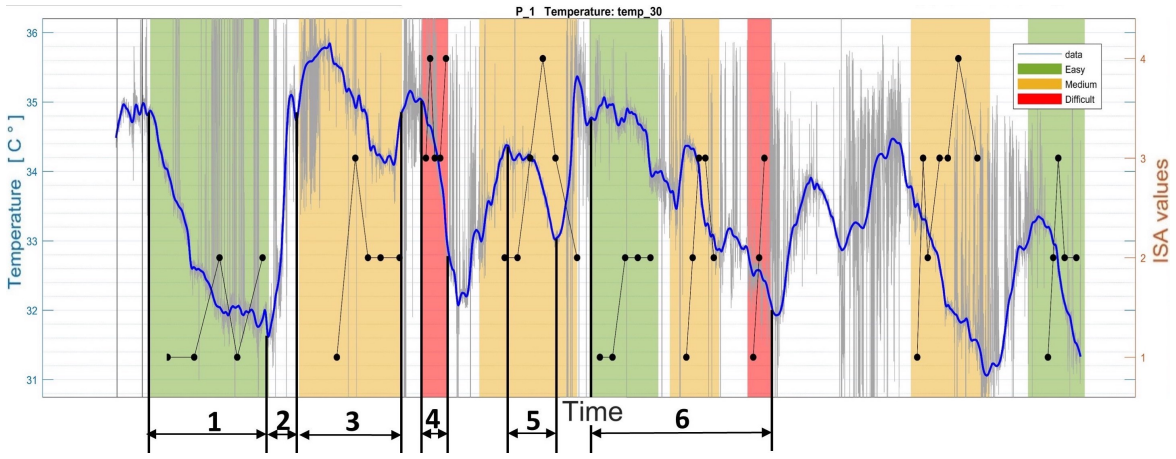

Fig. 6.17 Nose temperature example: Participant 1, Type I task, Point 30; Point 30 is marked in red in the left hand side of the image

Six time intervals are marked in the figure. As established in the study presented in Chapter 5, nose temperature is expected to drop when the participant is under high demand. During the time interval marked 1, representing an easy scenario, it can be observed that temperature dropped by about 3.2°C, almost the lowest temperature in during this trial. One of the possible reasons for this large change is that, as we found out later, at that point the participant had not flown a helicopter in about 4 months; temperature drops experienced during the other easy scenarios were not as high. The time interval marked by number 2 is interesting as the nose temperature inside point 30 goes back roughly to the initial temperature at the beginning of scenario 1, in a time interval of about 2 minutes and 22 seconds. During the time interval marked by number 3, the temperature starts decreasing (by about 1.9°C) as the scenario becomes more difficult and then starts returning to the initial value. The time interval marked by number 4, shows the most difficult scenario that required an autorotation to be performed, that in this particular case ended with crash. During this scenario, the temperature decreased by about 2.3°C in a time interval of about 2 minutes and 2 seconds; the temperature continues to decrease for a short time interval even after

the scenario is over before starting to recover. In this case, we can see the temperatures increasing at a slower rate when the next scenario starts, not having time to reach the initial temperature before the difficult part starts and it decreases again during time interval marked 5. Time interval 6 is interesting to observe because as the scenarios became increasingly more difficult, and the time intervals between them were not long enough for the temperature to come back to the initial value, we seem to be observing a cumulative effect of the increase in demand on the decrease of nose temperature.

While it is clear from most of the plots that the nose temperature decreases with the increase in demand, a similar approach to the one used above for heart rate and breathing rate provided no results. Instead of averaging the z-scores over the scenario time interval, a different approach was used in this case. Clear and consistent time intervals during which decreases in temperature took place (most likely caused by the onset of high demand situations) were selected manually for each participant and the variation of the z-score over the selected time interval was used to discriminate between various levels of difficulty.

A one way ANOVA ($F(2,35)=6.52$, $p<0.01$, $\eta^2 = 0.27$)[4] was performed on the z-score of the temperature data collected from point 30 on the nose for the participants performing Type I task as well as for the participants performing the Type II task ($F(2,26)=2.23$, $p=0.12$, $\eta^2 = 0.14$), grouped by the three levels of difficulty (easy, medium, difficult). For the Type II task, the p value for the F-statistic was not small enough to indicate the group means were significantly different. A multiple comparison test using the Bonferroni method was used to test if the means of the groups are significantly different. For the Type I task (Fig. 6.18 left) the difficult scenarios had means that were significantly different from the easy and medium difficulty scenarios. For the Type II task (Fig. 6.18 right), none of the groups were significantly different from each other.

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	0.1846	1	-0.7821	1.1512
Easy - Difficult	1.3281	<0.05	0.2639	2.3924
Difficult - Medium	1.1436	<0.05	0.2374	2.0498

Table 6.11 Multiple Comparison Test results between difficulty levels for the Type I task and point 30 temperature z-score

Hypothesis 2 stating that the physiological parameters will show sufficient changes as to differentiate between the three different levels of demand imposed was partially supported as the ANOVA tests for each of the measures showed that only some group means were

[4] η^2 indicates effect size and in this particular case it means that 27% of the total variance can be accounted for by group membership [80]

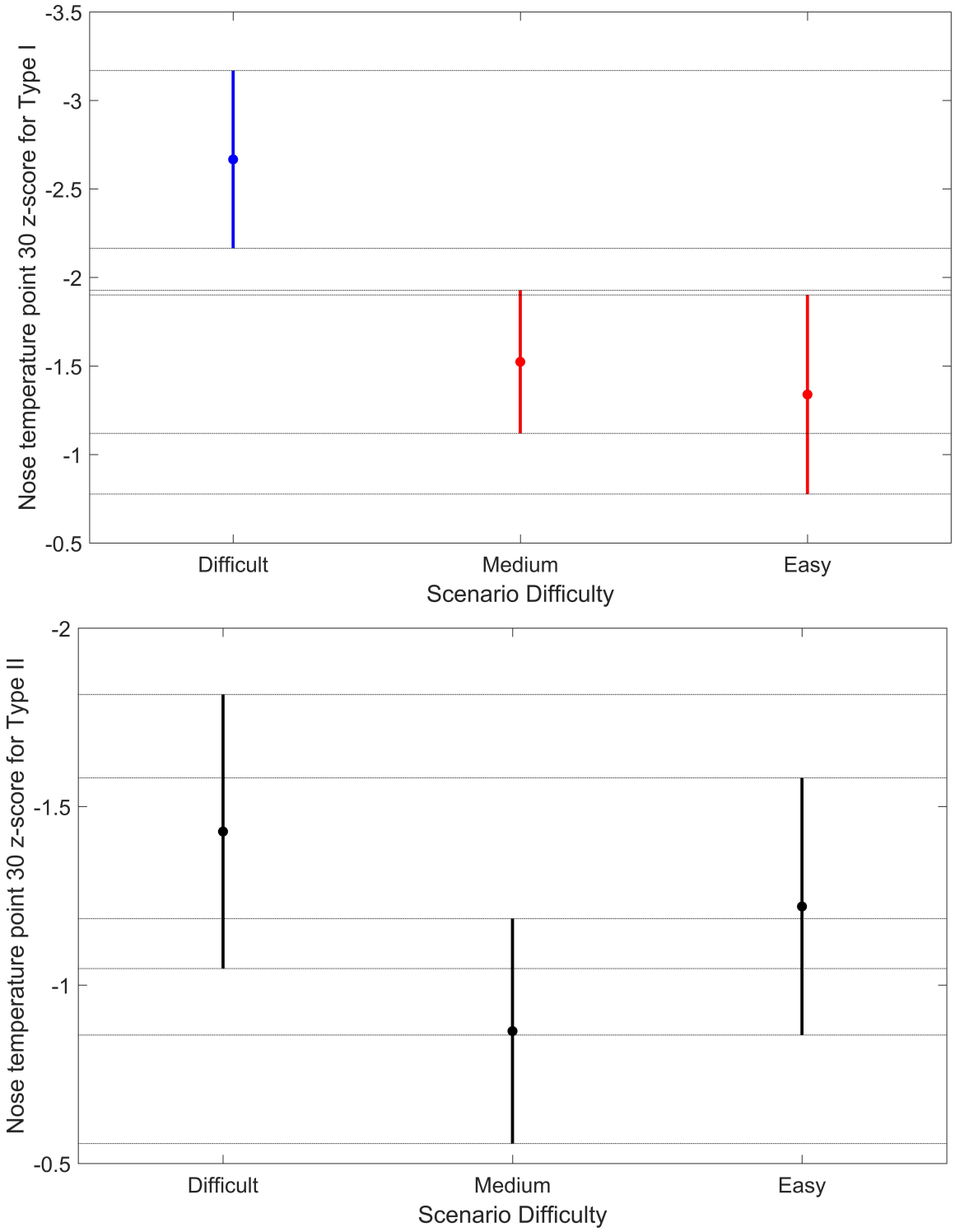

Fig. 6.18 Multiple comparison between difficulty levels and point 30 temperature z-score

Scenario Difficulty Groups	Mean Difference	Significance	95% Confidence Interval	
			Lower Bound	Upper Bound
Easy - Medium	-0.3492	0.5922	-1.0246	0.3263
Easy - Difficult	0.21	1	-0.5343	0.9543
Difficult - Medium	0.5592	0.1528	-0.1400	1.2583

Table 6.12 Multiple Comparison Test results between difficulty levels for the Type II task and point 30 temperature z-score

significantly different for all three levels of demand. Heart rate and nose temperature means were significantly different between the difficult scenarios and the others while breathing rate has showed significant differences only between the difficult and the medium scenarios for the Type II task.

Hypothesis 3 was that the physiological response of highly trained individuals to changes in workload will be similar to the ones of non-trained individuals performing a laboratory based task. There were obviously big differences between the task used for the laboratory study presented in Chapter 5 and the task performed by the pilots in the helicopter flight simulator. Nevertheless I will try to compare the physiological responses.

In most of the cases, heart rate increased with the increase of workload for both the pilots and the participants in the laboratory study. Large heart rate increases for pilots, that could clearly be attributed to high demand were mostly seen during the difficult scenarios. Large variations in the z-score of heart rate were manually selected and compared between the studies. On average, for the flight simulator study the changes were around 3.38 standard deviations with values between 1.5 and 5.8 standard deviations from the mean while for the participants performing the laboratory study the changes were on average 2.77 standard deviations, with values between 1.23 and 4.65. While this might not be a very significant measure, visually analysing the heart rate plots does seem to show more clear changes in the case of the pilots. Figure 6.19 shows a box plot of the variation in heart rate z-scores for both studies.

As the heart rate data, in most cases breathing rate increased with the increase in workload in both studies. Large variations in breathing rate z-scores were manually marked and compared between studies. During high workload the participants in the laboratory study experienced an average increase in the breathing rate z-score of 2.2 standard deviations from the mean with a range of 1.16 to 3.49 standard deviations from the mean, while the pilots had an average increase of 2.75 standard deviations from the mean with a range of 1 to 4.65 standard deviations from the mean. Overall the results are relatively similar although in the case of the pilots verbal communication during the scenarios may have influenced breathing rate. Figure 6.20 shows a box plot of the variation in breathing rate z-scores for both studies.

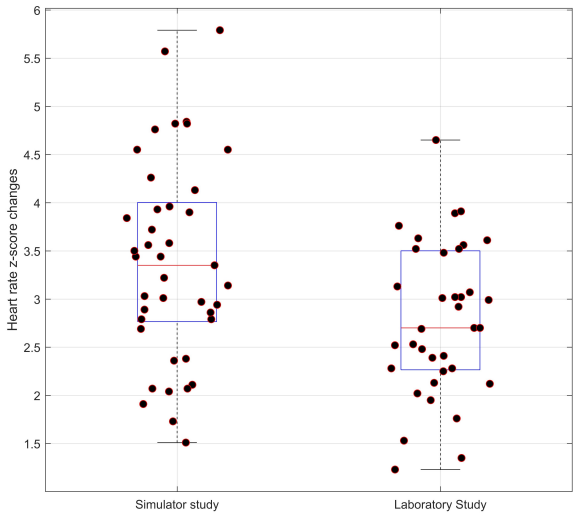

Fig. 6.19 Comparison between simulator and laboratory heart rate z-score results

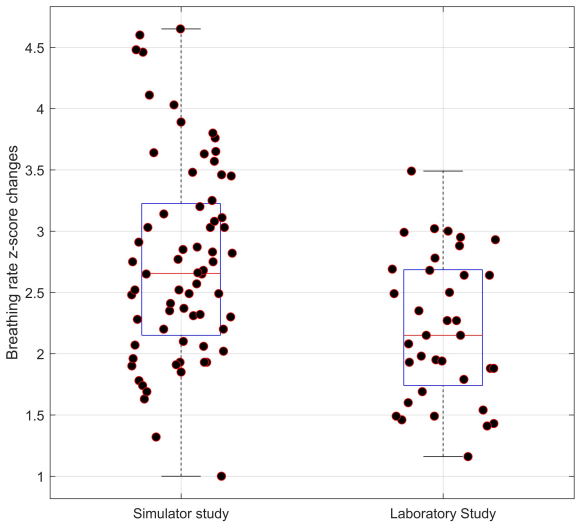

Fig. 6.20 Comparison between simulator and laboratory breathing rate z-score results

The nose temperature from point 30 presented in this chapter will be compared to the data from point P in the laboratory study. Overall the changes in nose temperature for the simulator and laboratory studies were similar. In the case of the simulator study the changes ranged between -4.75 and -0.27 standard deviations from the mean while for the laboratory study the range was -4.03 to -0.18 standard deviations from the mean. Figure 6.21 presents the box plots of the variations in nose temperature z-scores for both studies.

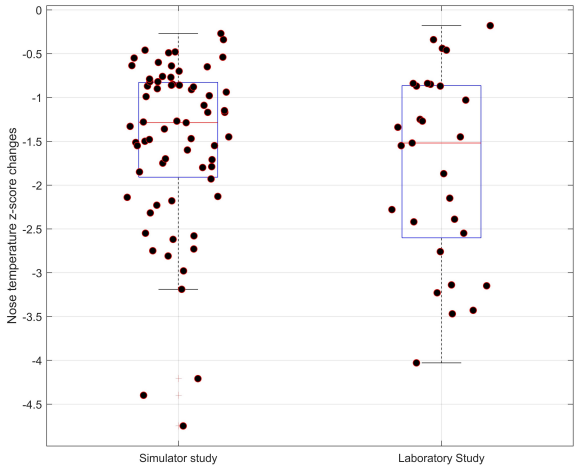

Fig. 6.21 Comparison between simulator and laboratory nose z-score results

Facial Thermography Individual Observations

Further on, some observations will be made on a participant basis. The first observation is related to the temperature decrease around the nose area with the increase in demand. Looking at the data it appears that the temperature decrease effect can be observed at a smaller scale in other areas of the face. Some time intervals where the facial landmark tracking performed better and the data were less noisy were selected to illustrate the temperature decrease in various areas. The participant thermal picture on the left hand side of each of the figures is of participant 1 and was used in all examples just to identify the areas from which temperature was considered for the examples. The colours in the plots are the same as the ones overlaid on the areas of the face. The areas on the right side of the face were labeled with capital letter while the areas on the left side of the face were labeled with the same letter but in lower case; these notations were used to refer to the facial areas that were analysed. Next to each plot, a thermal picture of the face is presented to more easily identify where the data in the plot were sampled from.

Figure 6.22 shows temperature data from Line 29-30 (on the nose) and areas A, B, C, D and E on the right side of the face, during scenarios 3 (difficult), 6 (medium) and 9 (easy).

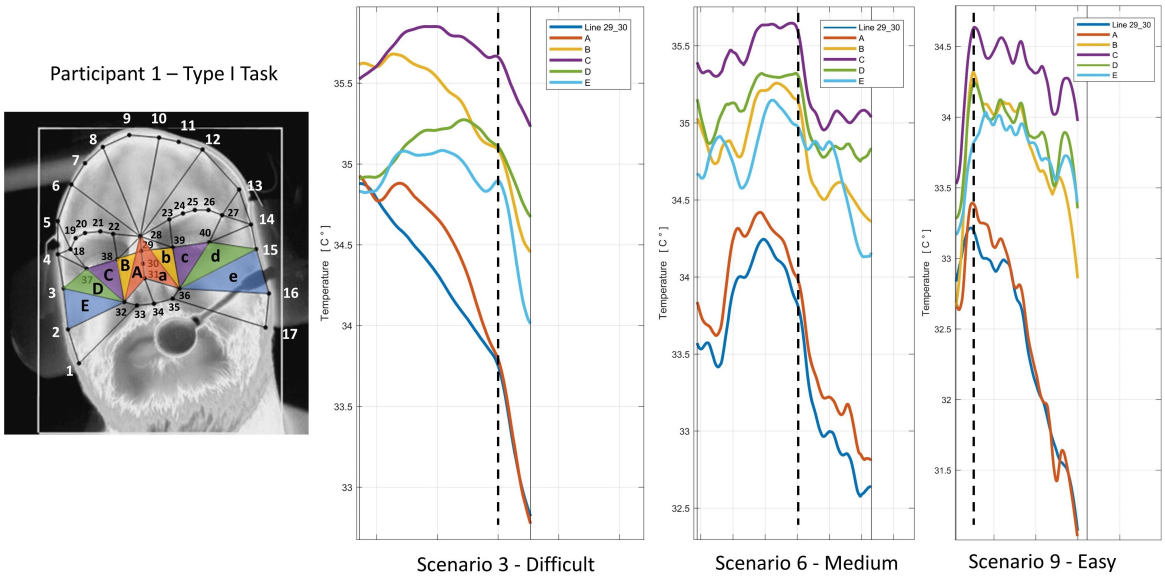

Fig. 6.22 Temperature around the nose area for participant 1, scenarios 3, 6, 9 during the Type I task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots. The vertical dotted line marks the point from which the temperatures in all of the areas starts decreasing.

For scenario 3 (difficult) shown in Fig.6.22, the temperature along line 29-30 and in area A started decreasing from the beginning of the scenario, with temperature in area B following very shortly. The other areas that are presented (C, D and E) are temporarily increasing and then start decreasing at a slower pace than the areas close to the nose. The vertical dotted line marks the point from which the temperatures in all of the areas starts decreasing.

Overall, during scenario 3, the average temperature of line 29-30 and of area A for participant 1 decreased by about 2°C at a rate of about 0.018°C/s. From the dotted line until the end of the scenario, the temperatures decreased by amounts of between 0.4°C to 1°C. Line 29-30 and area A had the largest temperature decrease and at the highest rate.

Scenario 6 of medium difficulty in Fig.6.22 shows a similar trend, the temperatures on Line 29-30 and on area A start decreasing while the others remain relatively constant; from the dotted line onwards, temperatures in all areas start decreasing. While towards the end of the scenario the rate of change for the nose areas slows down, areas like C and E seem to have already reached a relatively constant level. Scenario 9 of easy difficulty in the same figure shows a similar pattern.

Figure 6.23 shows how temperature in the same area vary for participant 3 during a Type I task for scenarios 3 and 7, both of them difficult. It can be observed, especially for scenario

7 that all areas respond to the variation of demand and towards the end of the scenario, temperature decreased by a range of 0.45°C to 1°C.

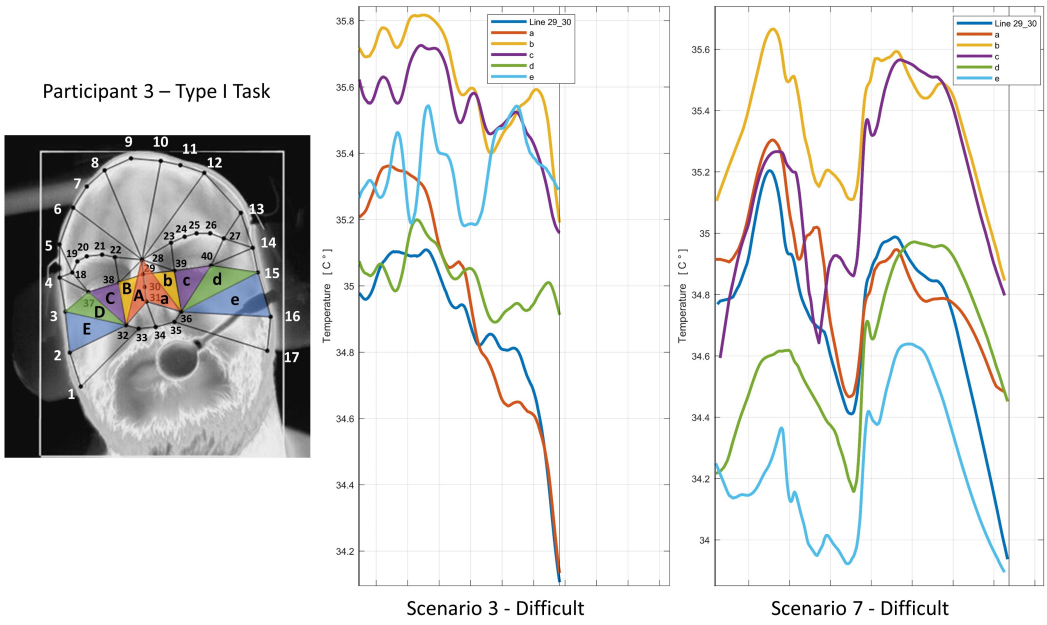

Fig. 6.23 Temperature around the nose area for participant 3, scenarios 3 and 7 during the Type I task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

Figure 6.24 shows the temperature response for participant 6 during a Type II task for scenarios 7, 8 and 9. Unlike the participants presented so far, for participant 6, as temperature decreased in area A and Line 29-30 during scenario 7, temperature in all other areas increased slightly.

Figure 6.25 shows the change in temperature for participant 7 during scenarios 3 (difficult) and 4 (medium) of a Type I task. The data collected during scenario 3 reveal that temperature decreased in all examined areas of the face (A, B, C, D, E). Line 29-30 and area A displayed temperature changes of -1.4°C and -1.3 °C, about twice as high as the next largest change in area B (-0.75°C) Fig. 6.25: Scenario 3-Difficult. As scenario 4 of medium difficulty level began, temperatures increased in all areas until the collective trim failure took place and the increase in temperature stopped 6.25: Scenario 4-Medium.

Figure 6.26 shows the thermal data collected from participant 9 during scenario 7 (difficult) of a Type I task. The largest changes in temperature as a response to increased demand were observed for area a (-2.3°C) and line 29-30 (-2°C). Further away from the nose, the next

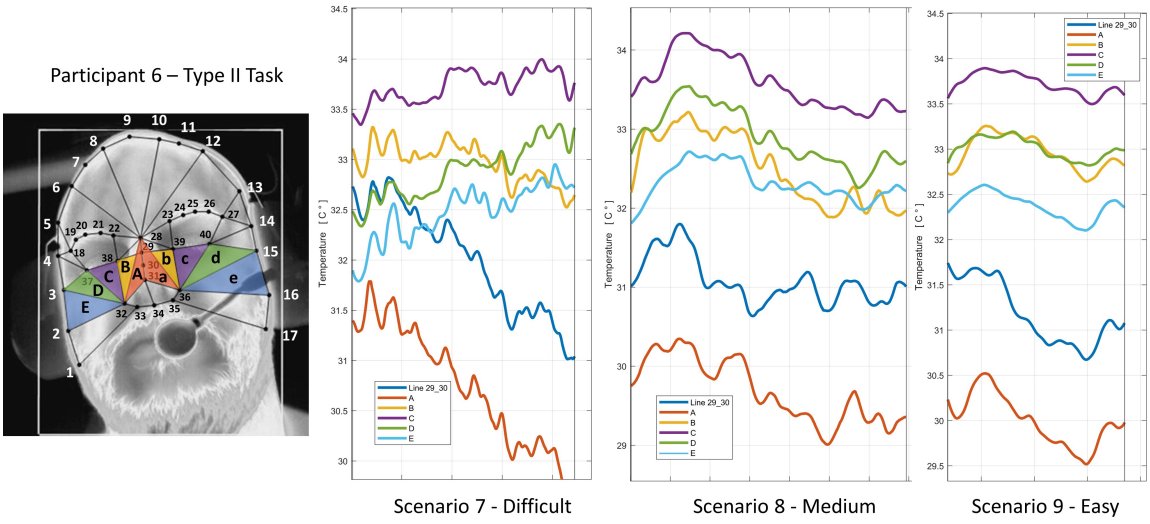

Fig. 6.24 Temperature around the nose area for participant 6, scenarios 7, 8, 9 during the Type II task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

largest change was measured in area b (-1°C) while areas c, d and e showed no significant changes.

Figure 6.27 shows the temperature variation for participant 10 during scenarios 4 and 8 of medium difficulty. During scenario 4, a small decrease in temperature can be observed for all observed areas, followed by an increase towards the initial levels. For scenario 8, all areas show a decrease in temperature, ranging from 0.4°C to 0.65°C, the largest change being recorded for area A.

Figure 6.28 shows the temperature variation for participant 11 during the difficult scenarios 3 and 7 of a Type I task. For scenario 3, the range of temperature variation for the examined areas was between -0.24°C and -1.2°C, the largest changes observed for line 29-30 and area a. Similar changes were observed for scenario 7, during which temperature in most areas decreased for the first part of the scenario and then started increasing. The biggest temperature change was observed for area a (-1.6°C) followed by line 29-30 (-1.2°C), area b (-0.76°C). The further away from the nose, the smallest the change in temperature.

Examples of data from seven participants were presented. The data were selected from scenarios during which the landmark tracking provided the best results with as little noise as possible. The data appear to show that in most cases, the nose area is the one where

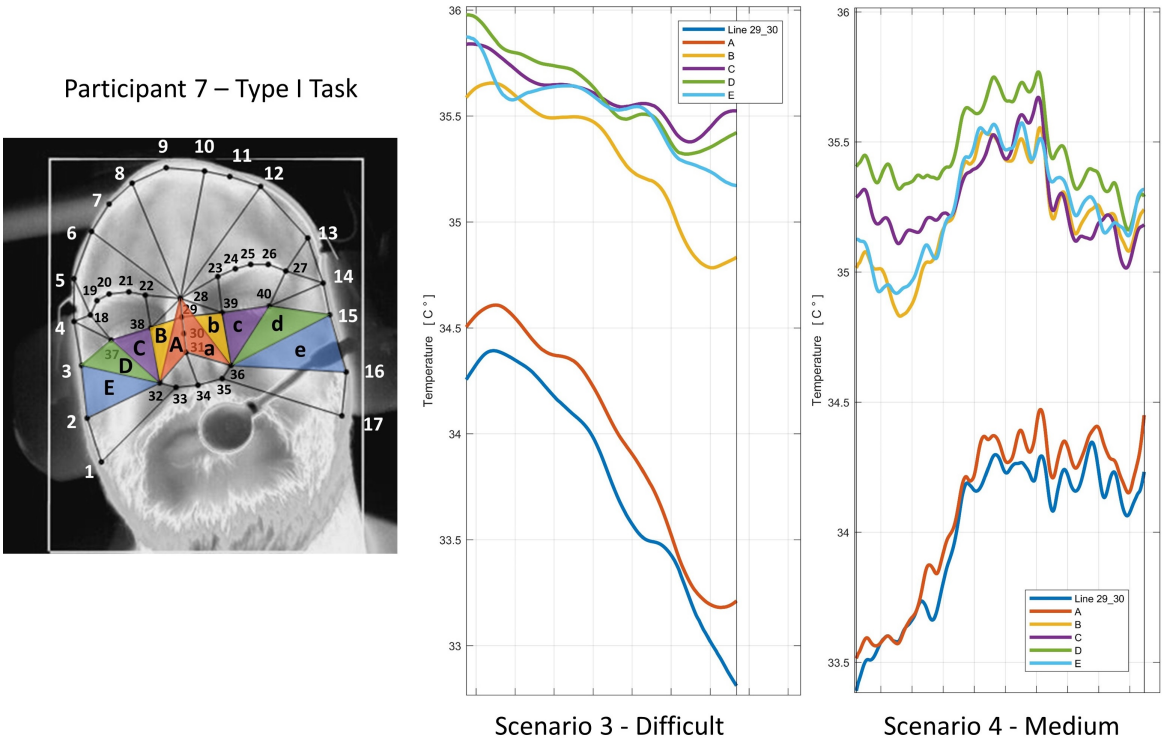

Fig. 6.25 Temperature around the nose area for participant 7, scenarios 3 and 4 during the Type I task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

temperature decrease happens first and where the temperature drop is the highest. Most researchers report that the temperature in the nose area decreases with the increase in workload. The examples presented above confirm this and open the possibility that the temperature decrease is taking place in other areas of the face but having a more reduced effect when measured further away from the nose.

Another observation to be made is related to forehead temperatures. It has been reported that forehead temperature remains constant during variations in workload [45], [5], [47] and it has also been used as a reference temperature [8]. The data from this study show that while in most situations forehead temperature does not change by much, there are changes that appear, especially during higher demand scenarios. Figures 6.29 and 6.30 show a few examples of forehead temperature decreasing at the same time with nose temperature in point 30, marked in the left side of the figures with blue. The sample area for forehead temperature is area 8-9-10-28, marked in the figure with light red.

In the examples presented in these figures, the forehead temperature decreased on average by 0.61°C, with values between -0.31°C to -0.86°C, which represented between 42% to 84%

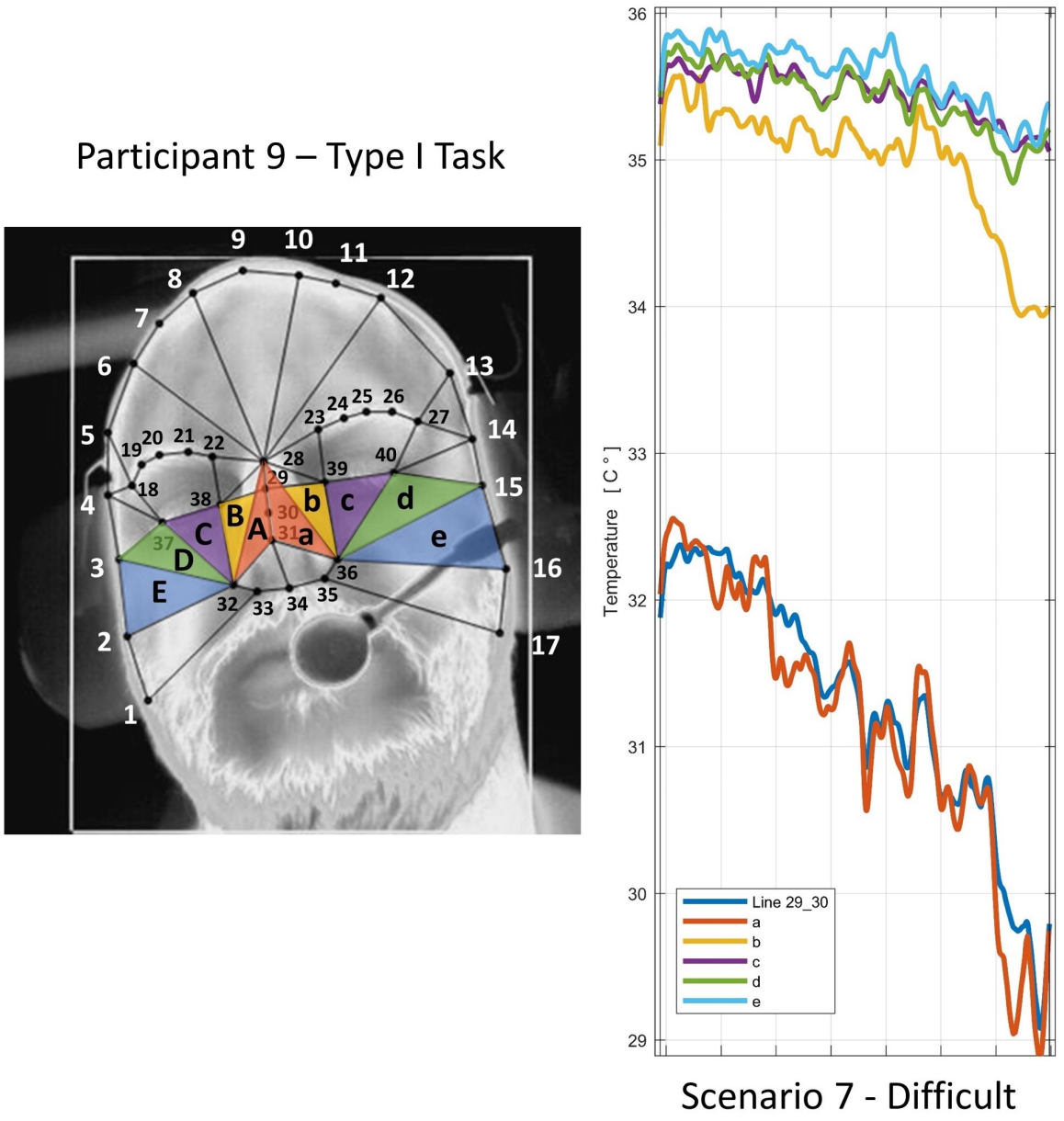

Fig. 6.26 Temperature around the nose area for participant 9, scenarios 7 and 9 during the Type I task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

of the changes occurring in the nose area. Although this type of large change in the forehead area occurred in 5 out of 10 participants, it might be something to consider for future studies.

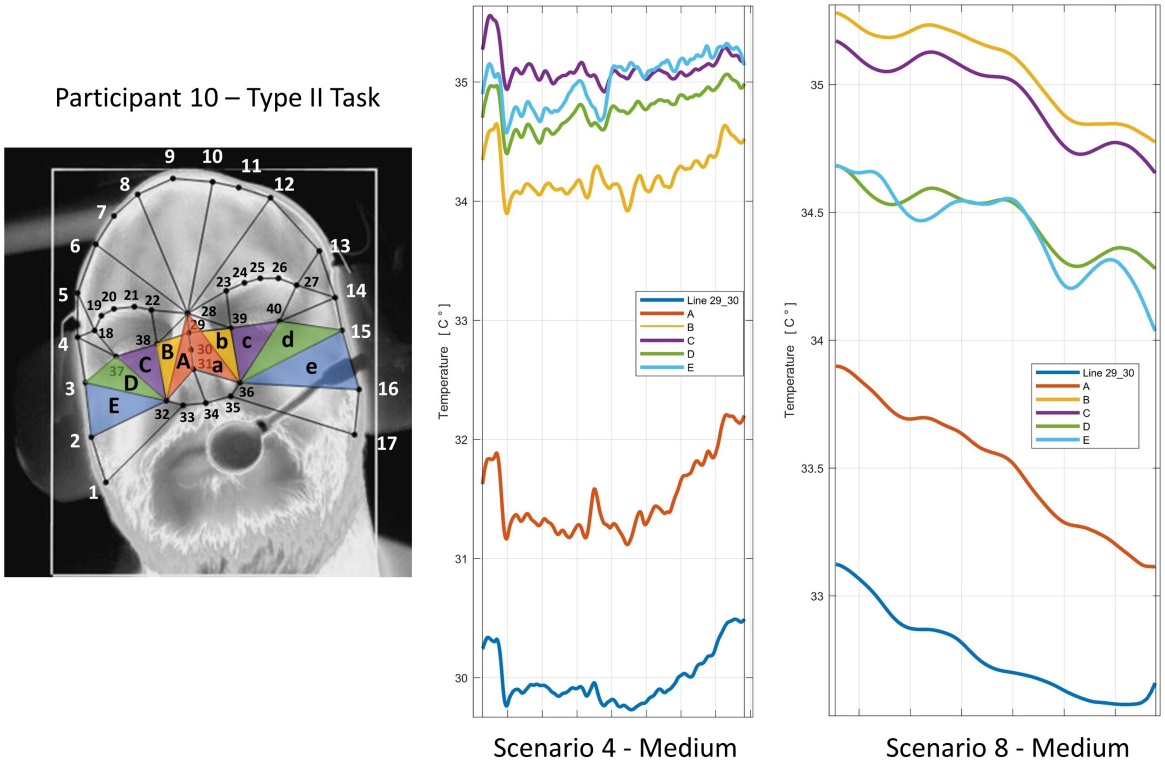

Fig. 6.27 Temperature around the nose area for participant 10, scenarios 4 and 8 during the Type II task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

6.4 Discussion

The results presented in this chapter demonstrate that physiological monitoring can be used for non-invasive real-time measurement of workload, in the case of highly trained individuals such as helicopter pilots. The results showed that heart rate presented significant changes during the high demand scenarios while breathing rate showed significant changes only in differentiating the medium from the difficult scenarios of the Type II task. The nose area temperature presented significant changes only during the difficult scenarios of the Type I task. In a comparison between the physiological reaction of the participants in the laboratory study presented in Chapter 5 and the pilots presented in this chapter, the physiological changes showed similar results with a mention that a visual analysis of the heart rate changes showed more clear variations in the case of the pilots. It can also be argued that the scenarios that the pilots were performing subjected them to higher levels of demand as well as potentially more stressful situations. This study shows that it is feasible to deploy physiological measures

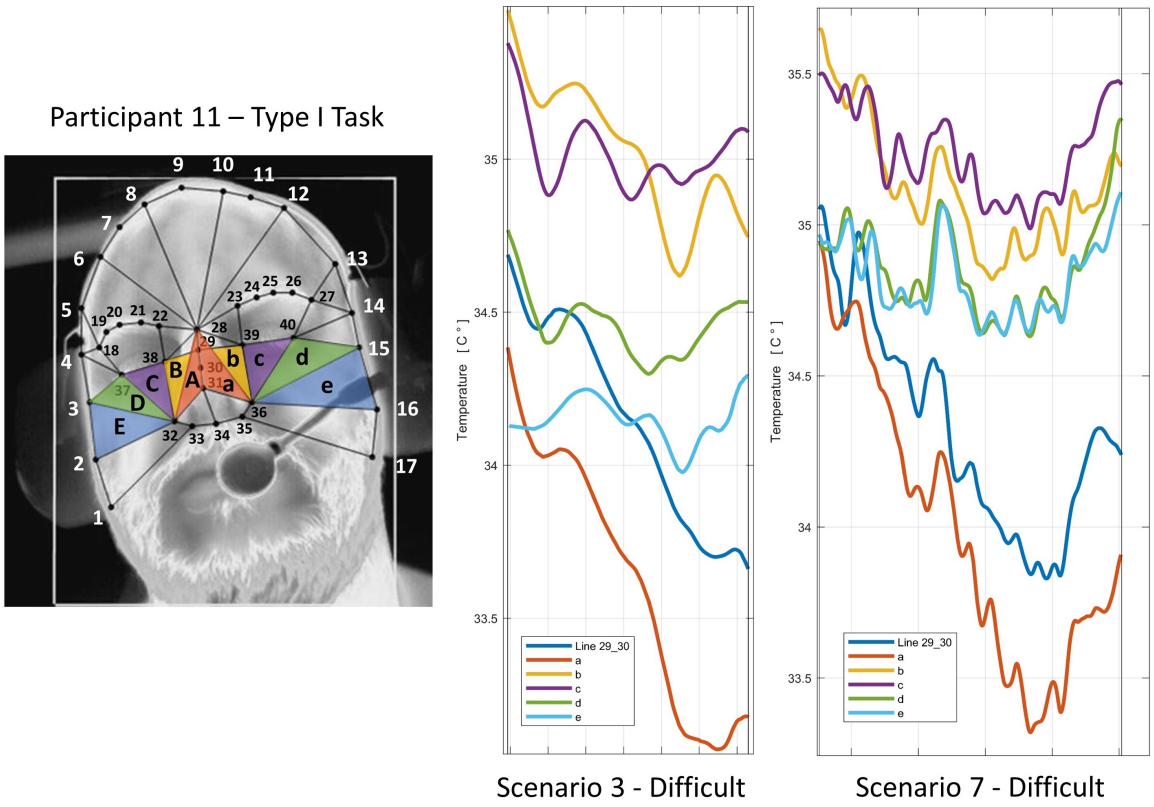

Fig. 6.28 Temperature around the nose area for participant 11, scenarios 3 and 7 during the Type I task; Data showing how facial temperature varies as it is measured further away from the nose. Three scenarios are presented and commented on. The colours marking the areas on the left side of the image are maintained in the plots

such as facial thermography in aircraft cockpits although more data would need to be studied from a larger population.

The main novelty lies in the fact that multiple continuous physiological measures were recorded from active helicopter pilots performing tasks in a high fidelity helicopter flight simulator. The hypotheses explored in this study were:

Hypothesis 1 proposing that there will be a measurable difference in subjective workload ratings between the three levels of task difficulty was supported, the ANOVA analysis on the ISA subjective ratings showed that there are significant differences between the group means defined by the three levels of difficulty scenarios.

Hypothesis 2 proposing that physiological parameters will show sufficient changes as to differentiate between the three different levels of demand imposed was partially supported, none of the physiological measures showed significant differences between the means of the easy and medium levels. Heart rate and nose temperature (only for Task I type) means were significantly different only between the difficult scenario and the others, but not between the

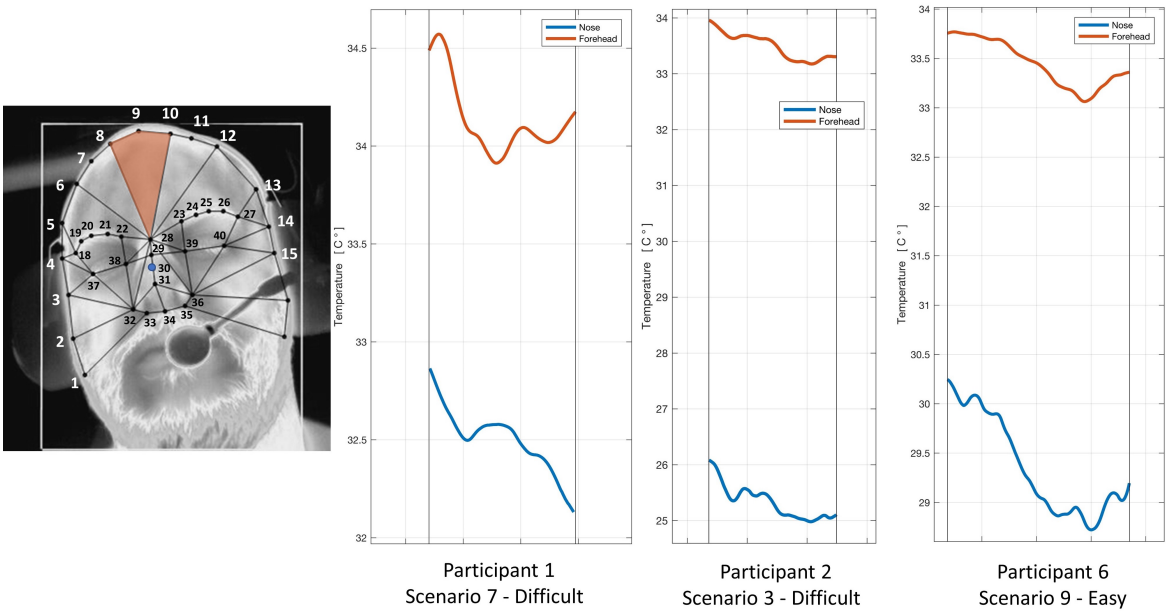

Fig. 6.29 Comparison between nose and forehead temperature - samples from participants 1, 2 and 6; The forehead area and nose point that temperature was sampled from are represented in orange and blue in the left hand side of the image. The plots on the right hand side maintain these colour codes

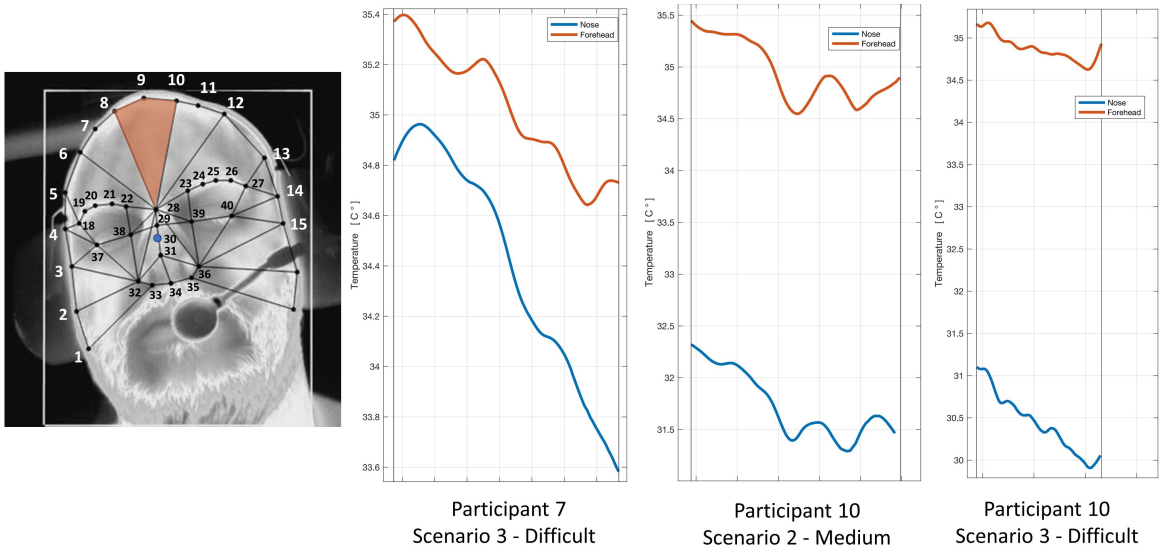

Fig. 6.30 Comparison between nose and forehead temperature - samples from participants 7 and 10; The forehead area and nose point that temperature was sampled from are represented in orange and blue in the left hand side of the image. The plots on the right hand side maintain these colour codes

easy and medium ones. Breathing rate showed significant changes only between the medium and difficult scenarios of the Type 2 task.

Hypothesis 3 aimed to test if the physiological response will be similar to non-trained individuals performing a laboratory based task of varying levels of difficulty was partially proven in the sense that, indeed there are physiological changes that can be seen in heart rate, breathing rate and nose temperature also for highly trained individuals, but the comparison may not be that valuable as the tasks were so different. Another observation to be made about the simulator study is that the large physiological changes mainly occured during very high demand situations which are unlikely to take place on a normal basis.

Potential limitations of this study include the small number of participants and the fact that they swapped between roles as data could be recorded only from one participant at a time. Even though after changing role the scenarios were different, having assisted to the previous round of scenarios, as well as the fact that most of the scenarios would include a different type of failure, this could have had an effect on expectancy.

6.5 Chapter Summary

This chapter presented the results obtained in an ecologically valid helicopter flight simulator, together with the challenges of performing a study outside the laboratory. The results showed that physiological changes during elevated levels of workload occur also in the case of highly trained individuals. The changes can be summarized as increases in heart rate, breathing rate and decreases in nose temperature.

This chapter has addressed research questions 1, 2 and 4:

1. How do human physiological responses change in response to variations in task demand and task performance? This was explored by recording multiple channels of physiological data while the participant was exposed to a task eliciting a predefined pattern of demand.

2. How are physiological responses associated with variations in subjective reports of mental workload? This was tested by comparing the response of multiple physiological data pieces with the levels of demand that were imposed by the task. The levels of demand were shown to be differentiated by the subjective ratings of mental workload (Hypothesis 1 of this study)

3. How do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator? This was tested by recording physiological data from highly trained pilots performing flying tasks in a high fidelity helicopter simulator. The physiological reaction was also compared to the one of the students performing a laboratory task.

The next chapter presents an extension of the study presented in Chapter 5 also with the introduction of the fNIRS sensor which was proven to be sensitive to changes in workload and also has high face validity.

Chapter 7

Physiological Measures of Workload - fNIRS Comparison

7.1 Introduction

The main aim of the study presented in this chapter is the same as for the study presented in Chapter 5, to examine the relationship between experienced workload and physiological response by non-invasive monitoring of physiological parameters. First of all, this study extends the study presented in Chapter 5 and also improves on the first study by also collecting fNIRS data which were proven to be sensitive to changes in workload and also has higher face validity. The fNIRS data act almost as a validation measure as it is able to estimate blood oxygenation in an area of the brain which is known to be associated with working memory, executive and attention processes [81]. The fNIRS data were collected by a colleague and does not represent my own contribution to the thesis.

Methods: The study presented was conducted in laboratory conditions and required participants to perform a custom-designed visual-motor task that imposed varying levels of demand. This is the same task as the one for the study presented in Chapter 5. The data collected consisted of: physiological measurements (heart inter-beat intervals, breathing rate, facial thermography, fNIRS - blood oxygenation); subjective ratings of workload from the participants (ISA and NASA-TLX); and the performance measured within the task.

Results: Facial thermography demonstrated to be a good candidate for non-invasive mental workload measurements; facial thermography measures added on average 40% to the amount of variability in task performance explained by a regression model. Blood oxygenation in the prefrontal cortex showed strong correlations with the subjective ratings for all participants and improved the amount of variability in performance accounted for by

the regression model by an average of 71% compared to the heart and breathing rate model. As with the ISA ratings, the relationship between the physiological measures and performance showed strong inter-participant differences, with some individuals demonstrating a much stronger relationship between workload and performance measures than others.

7.2 Experiment Design

The study presented in this chapter explores the changes in the physiological parameters that occur as the level of mental workload varies and examines whether a combination of these parameters could be used for estimating the level of mental workload. The study uses the same task as the study presented in Chapter 5 aiming at eliciting different levels of experienced workload which are then captured by subjective and physiological measures. The hypotheses of this study are that:

1. There will be a measurable difference in subjective workload between the two levels of task difficulty
2. The subjective ratings of workload will be associated with changes in physiological measures
3. Multiple physiological measures should be used in combination to analyse workload.

7.2.1 Participants

Eleven students and staff from the University of Nottingham took part in the study (6 men and 5 women). The average age of the participants was 29 years old, SD = 6.8, range = 19-42. The participants were recruited via e-mail and were compensated with a £20 Amazon voucher for their time. The study was approved by the Faculty of Engineering Ethics Committee. Each participant was presented with an information sheet and consent form, stating that they are over 18 years old, had no pre-existing heart-related condition and had no skin conditions or allergies that could prevent them from wearing the heart rate chest strap.

7.2.2 Materials

The materials used for the study are described in Chapter 5, Section 5.2.2.

7.2.3 Design

The independent variable that was manipulated during the study was the task difficulty (i.e. imposed demand). The dependent variables were the physiological measures, the subjective assessment of the perceived level of mental workload and the task performance. The Instantaneous self-assessment workload scale (ISA) (Brennen, 1997) [22] was used once every 45s to collect subjective data about the level of perceived mental workload. The ISA scale was developed primarily as a subjective measure of mental workload for air traffic controllers and it involves the participants self-rating their workload on a scale from 1 (low) to 5 (high). The main reason for using the ISA scale throughout the task was the low level of intrusion, as the participant would verbally rate the perceived level of mental workload when prompted by an auditory message ('Level please'). At the end of each of the three task stages, the participant filled in a NASA-TLX (Hart, California, & Staveland, 1988) [23] questionnaire for a subjective assessment of workload. The reason for using NASA-TLX was to get a more detailed retrospective multidimensional subjective assessment of each of the three stages to determine whether the manipulation of imposed demand through task difficulty had resulted in a perceived experience of increased workload.

7.2.4 Procedure

Each participant was invited to read the information sheet, describing the details of the study, and then fill in a consent form. They would then play a training version of the stimulus task until they became familiar with the rules and the controls; after the training was finished, the participants would be invited to attach the Zephyr sensor on, in a private space; the placement of the fNIRS device on their forehead would follow next; the thermal and visual cameras would then be aligned to match the height of the participant. When the participant was ready, they would play Stage 1 of the stimulus task, which would last for almost 10 minutes, at the end of which, the participant's score would be shown in comparison to the participants before. During the game-play, the participant would rate the level of mental workload on the ISA scale once every 45s (when prompted by the voice). After finishing each of the stages, the participant would be shown their score in comparison to the others and would be asked to fill in a NASA-TLX questionnaire. After stage 3 was over and the questionnaire filled, the participant would be invited to remove the Zephyr sensor in a private space.

As ambient temperature could affect facial temperature, room temperature was kept as constant as possible without the use of air conditioning which could result in blowing cold air towards the participants. Room temperature was recorded every 45s using a room thermometer to insure that there were no external factors that could influence the measurement.

Each participant was offered a £20 voucher as a reward for their time.

7.2.5 Measurements and Equipment

The data collected during the study were of three types:

1. Performance data, generated by the task
2. Subjective data: NASA-TLX and ISA
3. Physiological data collected using the following pieces of equipment[1]:
   (a) Zephyr Bioharness 3
   (b) FLIR SC7000 Thermal Camera
   (c) fNIRS
4. Room temperature using a digital room thermometer

7.2.6 Data Analysis

The data generated by each of the sensors were post processed in Matlab. The data analysis techniques are described in Chapter 4. The results will be presented in the same format as in Chapter 5.

7.3 Results

The results will be presented in several stages. Firstly the variation of ambient temperature is presented as it could be a factor potentially influencing facial temperature. Secondly the results of the inferential tests to examine the impact of the manipulation of the task demand on the measures of workload and performance are presented. The aim of these analyses is to confirm that the demand manipulation affected workload and performance in the manner anticipated. The third part of analysis examines the relationship between the different measures of workload, using bivariate correlations and reporting both correlation significance and the coefficient of determination to indicate effect size. The final analysis uses multiple linear regression to determine the percentage of variability in task performance explained by the physiological measures and the relative contribution of each of the measures.

[1] All pieces of equipment are described in Chapter 4

Ambient Temperature Data

Ambient temperature was recorded every 45s using a digital room thermometer. The recordings show a gradual slight increase in temperature for each of the participants. Figure 7.1 shows the distribution of temperature variation throught the study. Each blue marker overlayed on top of the boxplot represents the difference between the room temperature at the end and the room temperature at the beginning of the study; for all of the participants, this also represents the maximum variation of ambient temperature. We can observe that for only two of the participants, the variation in ambient temperature was greater than 0.6°C, respectively 0.8°C and 2.1°C.

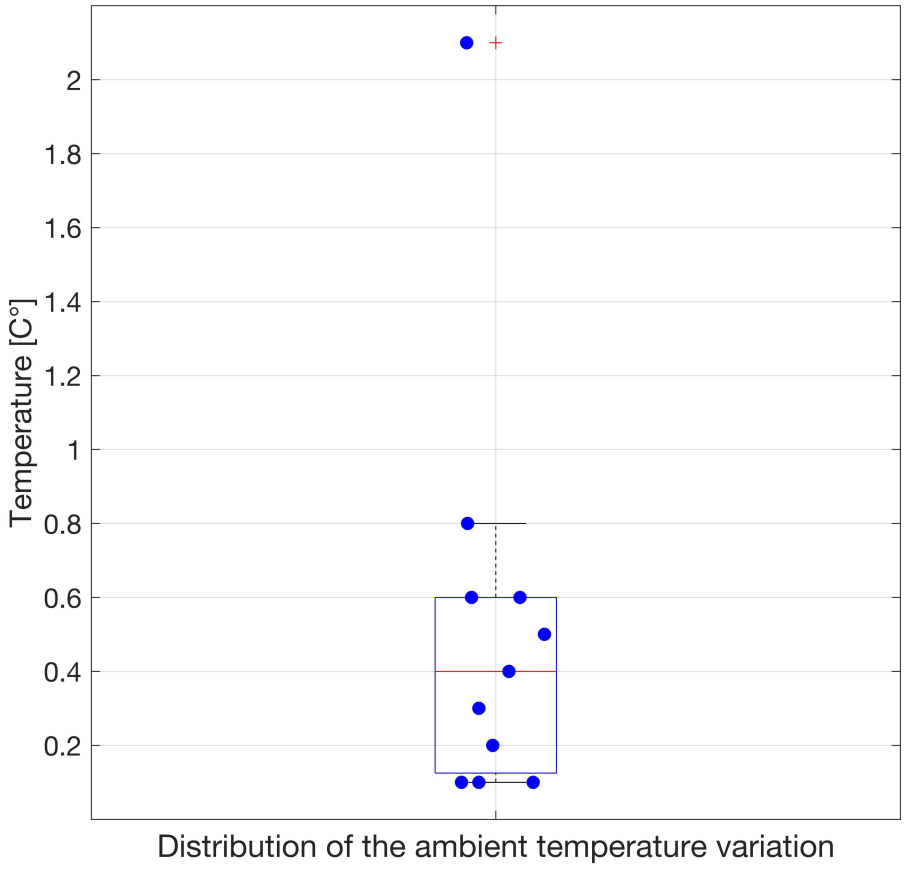

Fig. 7.1 Distribution of the ambient temperature variation

In the case of participant 3, the increase in ambient temperature was the largest, which might explain the general trend of facial temperature increase. Figure 7.2 shows the variation of temperature in the nose area for participant 3; while we can observe a slight decrease in temperature around stages 19 and 24, as for the other participants, we can also observe a

gradual increase in temperature towards the end of the task. This might have been influeneced by the large increase in ambient temperature (2.1°C).

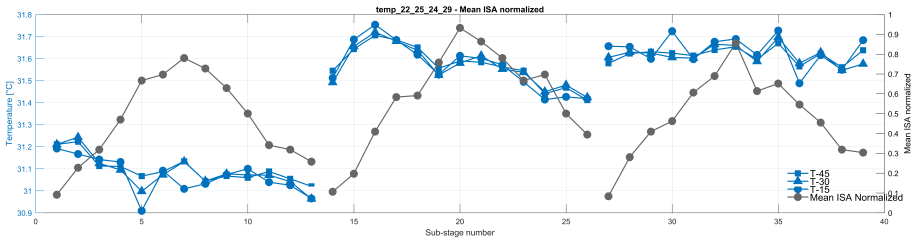

Fig. 7.2 Variation of nose temperature for Participant 3

7.3.1 Subjective and Performance Data

A one way ANOVA (F(1,31)=2.75), p = 0.1, $\eta^2 = 0.08$) indicated that there was not significant difference between the two levels of difficulty in terms of the NASA-TLX mental demand scale. As opposed to the results obtained in the first study, this shows that stage 2 (odd numbered balls as targets) was not perceived to be more mentally demanding than stages 1 and 3 (red balls as targets). As in Chapter 5, in order to overcome the disadvantage of the ISA subjectivity and assuming the relative validity of the ratings is robust, the data were normalized to a common scale ranging from 0 to 1. This should allow for a better comparison between participants. Fig. 7.3 shows the mean performance score for all participants (a better performance in the task results in a higher score) at sub-stage scale, plotted against the mean normalized ISA rating for all participants. There is a very strong negative correlation between the two mean scores, the Pearson correlation coefficient is r(37)=-0.87 with p<0.01, showing that as the mean subjective level of mental workload increased, the mean task performance decreased.

While Fig.7.3 looks at the mean performance and level of mental workload, Table 7.1 shows the individual correlations with performance of both the mean ISA normalized and to each participant's rating. For both the normalised ISA ratings and individual (non-normalised) ISA ratings the correlation coefficients were strong and the correlations were significant to the 0.01[2] level. The r^2 values were identical in both cases. These data demonstrated a clear association between performance and subjective workload.

[2]Note that no familywise corrections such as Bonferonni were applied, as tests were conducted on independent (participant-based) data sets, but it should be acknowledged that as normal when multiple tests are conducted one in twenty will be significant by chance if a p<0.05 level of significance is adopted

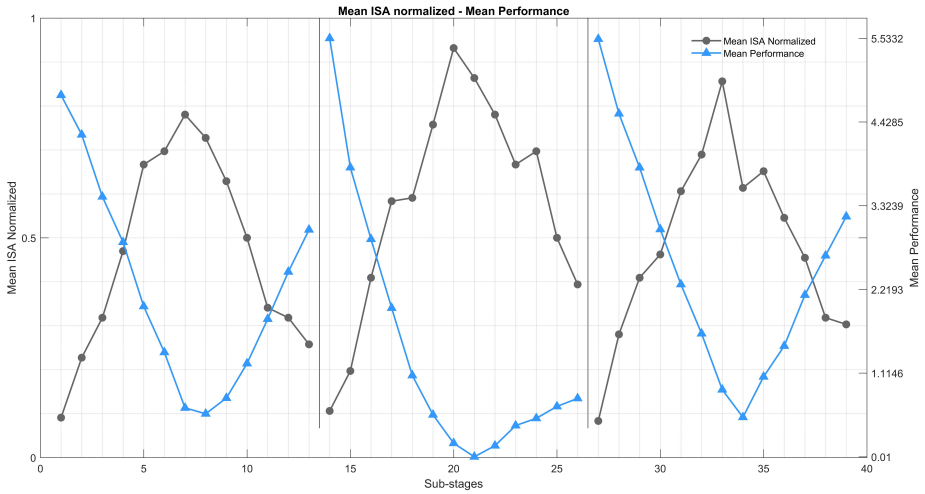

Fig. 7.3 Mean ISA ratings - mean score

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	P	r(37)	r^2	P
1	**-0.607**	**0.368**	**<0.01**	**-0.607**	**0.368**	**<0.01**
2	**-0.579**	**0.335**	**<0.01**	**-0.579**	**0.335**	**<0.01**
3	**-0.569**	**0.324**	**<0.01**	**-0.569**	**0.324**	**<0.01**
4	**-0.694**	**0.482**	**<0.01**	**-0.694**	**0.482**	**<0.01**
5	**-0.691**	**0.477**	**<0.01**	**-0.691**	**0.477**	**<0.01**
6	**-0.684**	**0.468**	**<0.01**	**-0.684**	**0.468**	**<0.01**
7	**-0.532**	**0.283**	**<0.01**	**-0.532**	**0.283**	**<0.01**
8	**-0.656**	**0.43**	**<0.01**	**-0.656**	**0.43**	**<0.01**
9	**-0.818**	**0.669**	**<0.01**	**-0.818**	**0.669**	**<0.01**
10	**-0.578**	**0.334**	**<0.01**	**-0.578**	**0.334**	**<0.01**
11	**-0.634**	**0.402**	**<0.01**	**-0.634**	**0.402**	**<0.01**

Table 7.1 Mean and normalized ISA ratings correlated with Performance

7.3.2 Physiological Data

The physiological data collected consisted of heart R-R inter-beat intervals, breathing rate, facial skin temperatures measured by thermography and fNIRS data.

Due to the fact that physiological data depends so much on the physiology of each of the participants and also on the reaction each of them has to the stimulus task, the correlations of each of the physiological signals with the ISA subjective ratings (both mean normalized and individual values) will be presented in tabular form for each of the participants individually, together with strong and weak correlation example plots. This helps us understand whether any association between physiology and subjective ratings applies across a population or whether there are different levels of strength of relationships between different predictive

variables in different populations. There are no R-R and R-R derived measures and breathing rate results reported for participant 5 as the Zephyr chest strap was too large for them to wear it.

Mean R-R Inter-Beat Intervals

Table 7.2 shows the correlation of the R-R inter beat intervals (averaged over a period of 45 seconds) with both the mean normalized ISA values and the individual ISA ratings. For participants 1, 8 and 10 a moderate negative correlation was found to both mean ISA normalized and individual ISA. Participant 3 showed a weak positive correlation with the individual ISA. This parameter showed the best correlations when averaged 45 seconds before the ISA sampling time.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.524**	**0.275**	**<0.01**	**-0.605**	**0.366**	**<0.01**
2	0.196	0.039	0.231	0.219	0.048	0.181
3	-0.068	0.005	0.682	**0.389**	**0.151**	**<0.05**
4	-0.091	0.008	0.583	0.153	0.024	0.351
5	-	-	-	-	-	-
6	0.029	0.001	0.859	-0.195	0.038	0.234
7	-0.08	0.006	0.63	-0.224	0.05	0.171
8	**-0.554**	**0.307**	**<0.01**	**-0.54**	**0.292**	**<0.01**
9	-0.169	0.029	0.303	0.051	0.003	0.757
10	**-0.525**	**0.275**	**<0.01**	**-0.403**	**0.162**	**<0.05**
11	-0.223	0.05	0.173	-0.281	0.079	0.083

Table 7.2 Feature mean-RR (45s) correlated with subjective ISA reports

Figure 7.4 shows the R-R values averaged over 45s, 30s and 15s before the ISA sampling point (in blue), the mean ISA normalized and individual ISA for participant 1, that showed the strongest correlation with the individual ISA.

Even though except for participants 1, 8 and 10 the correlations presented are not significant over the entire duration of the task, it is worth discussing about some local changes in the mean R-R. For example, participant 11 (Figure 7.5) shows an interesting decrease in mean R-R between sub-stages 4-5 and a slow return by sub-level 9 to roughly the mean R-R values of sub-level 4; in much the same way, the mean R-R values decreased from sub-stage 14 to 25 or sub-stage 32 to 34, in a pattern similar to the participants having high correlations.

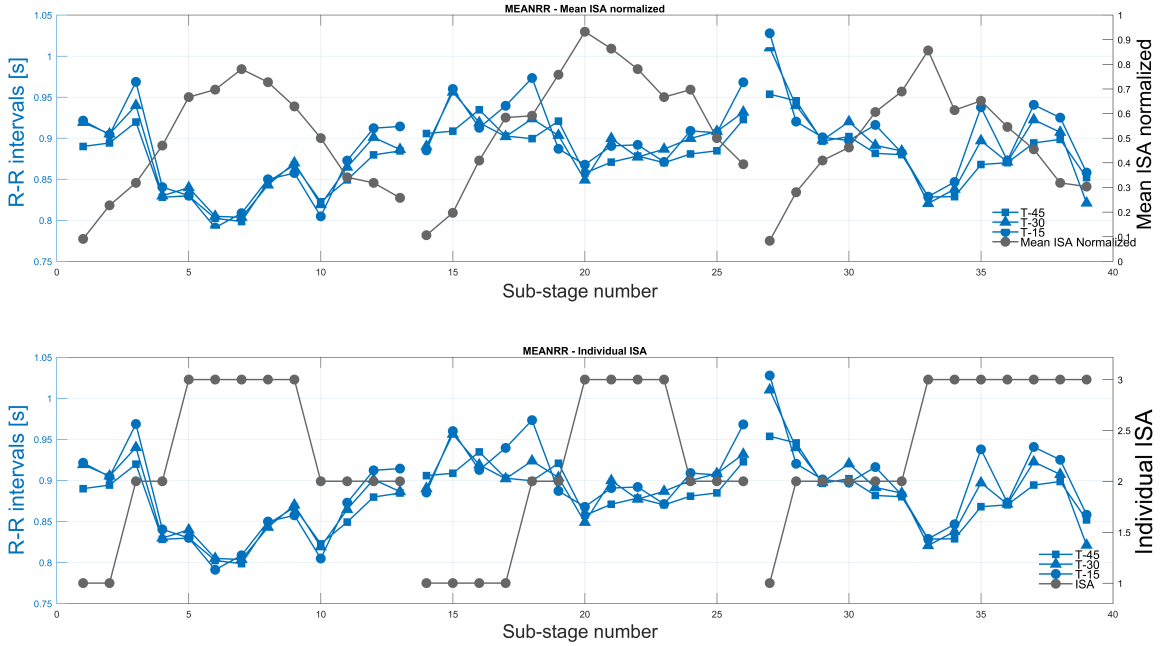

Fig. 7.4 mean R-R - mean ISA normalized and individual ISA for Participant 1

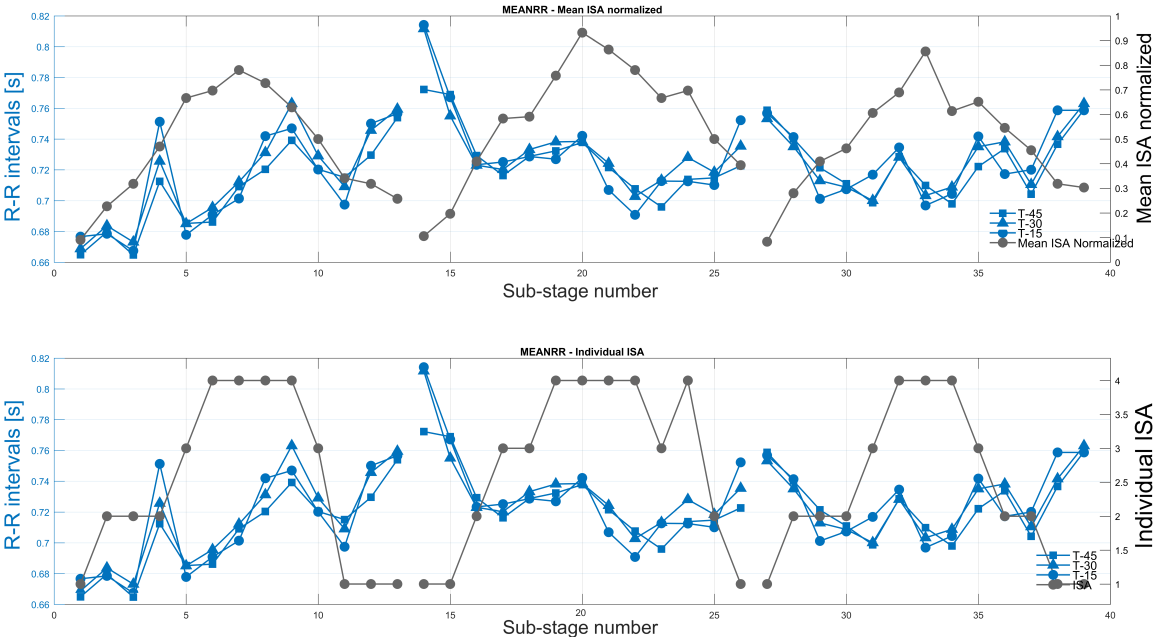

Fig. 7.5 mean R-R - mean ISA normalized and individual ISA for Participant 11

Participant 6 shows one of the weakest correlations between mean R-R and ISA. Even in this case, there are decreases in the mean R-R values when the task was expected to become more difficult (e g. Figure 7.6 sub-stages: 8, 11, 15, 23, 30, 33).

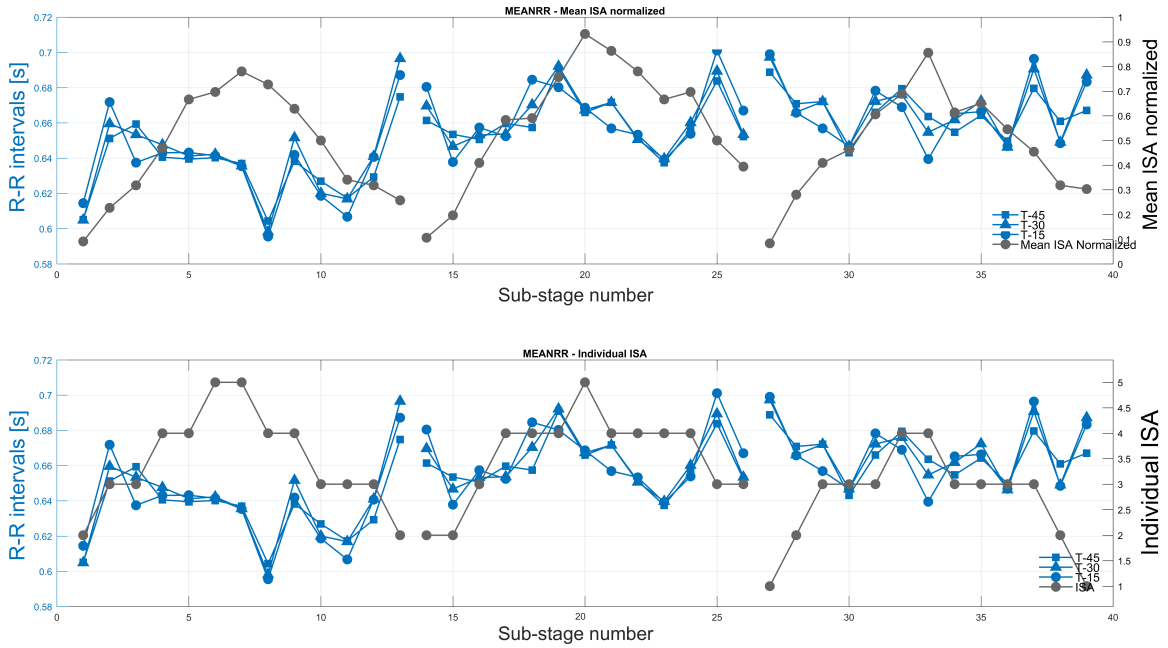

Fig. 7.6 mean R-R - mean ISA normalized and individual ISA for Participant 6

SDNN - Standard Deviation of the R-R Data

The SDNN[3]: measure showed the best correlation with the ISA subjective when computed over 45 s time intervals prior to the ISA sampling time. Three out of ten participants showed significant correlations with the mean ISA normalized while the data from only two of them were significantly correlated to the individual ISA values. Participants 1 and 4 showed moderate correlations of the SDNN feature with the mean ISA normalized while participant 7 showed a weak correlation. Participant 1 had a moderate correlation with the individual ISA values as well while participant 6 showed a weak correlation only with the individual ISA values - Table 7.3

Figure 7.7 shows the SDNN values computed over 45s, 30s and 15s before the ISA sampling point (in blue), the mean ISA normalized and individual ISA for participant 1, that showed the strongest correlation with the subjective measures. Figure 7.8 shows the SDNN values for participant 11, displaying one of the weakest correlations.

[3]Described in 4.2.1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.523**	**0.274**	**<0.01**	**-0.53**	**0.281**	**<0.01**
2	-0.184	0.034	0.262	0.073	0.005	0.658
3	-0.251	0.063	0.124	0.091	0.008	0.582
4	**-0.413**	**0.17**	**<0.01**	-0.258	0.067	0.113
5	-	-	-	-	-	-
6	-0.162	0.026	0.323	**-0.334**	**0.112**	**<0.05**
7	**-0.327**	**0.107**	**<0.05**	-0.147	0.022	0.371
8	0.004	0	0.982	0.021	0	0.9
9	-0.126	0.016	0.446	-0.093	0.009	0.572
10	-0.209	0.044	0.202	-0.099	0.01	0.55
11	-0.018	0	0.916	0.005	0	0.976

Table 7.3 Feature SDNN (45s) correlated with subjective ISA reports

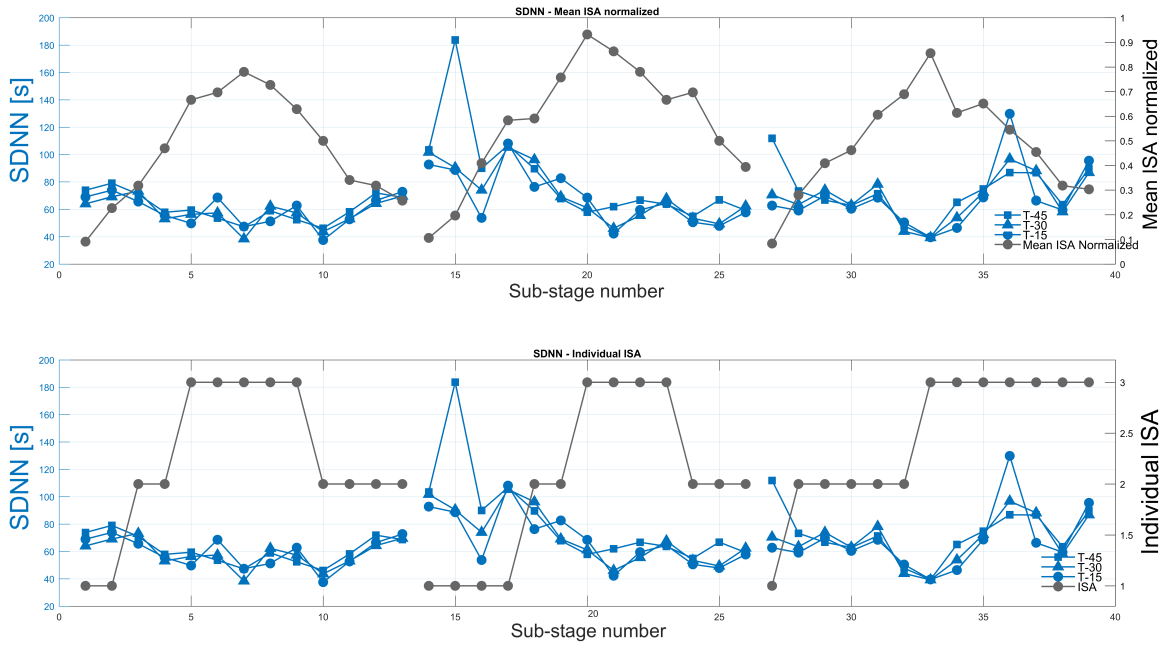

Fig. 7.7 SDNN - mean ISA normalized and individual ISA for Participant 1

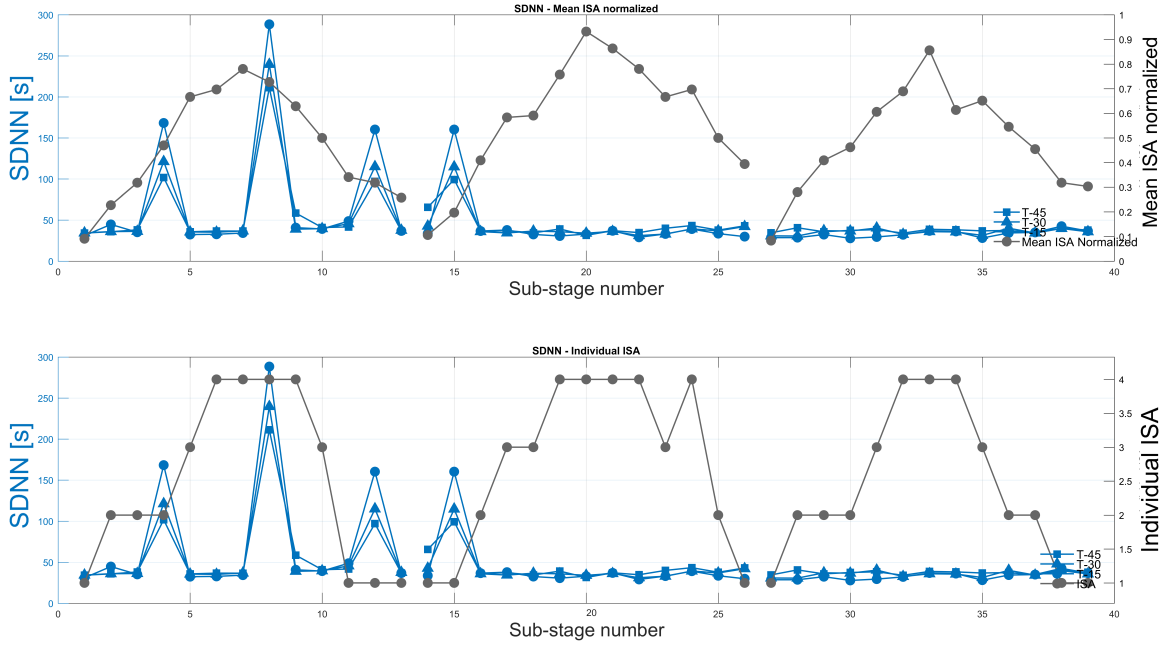

Fig. 7.8 SDNN - mean ISA normalized and individual ISA for Participant 11

SDSD - Standard Deviation of Successive Differences of the R-R Data

For the SDSD[4] measure, when computed over 45s time intervals prior to the ISA sampling, only participants 1 and 3 showed weak to moderate correlations with the subjective data, Table 7.4.

Figure 7.9 shows the SDSD values computed over 45s, 30s and 15s before the ISA sampling point (in blue), the mean ISA normalized and individual ISA for participant 1, that showed the strongest correlation with the subjective measures.

RMSSD - Standard Mean Square of Successive Differences of the R-R Data

For the RMSSD[5] measure, when computed over 45s time intervals prior to the ISA sampling, only participants 1 and 3 showed weak to moderate correlations with the subjective data, Table 7.5.

Figure 7.9 shows the RMSSD values computed over 45s, 30s and 15s before the ISA sampling point (in blue), the mean ISA normalized and individual ISA for participant 1, that showed the strongest correlation with the subjective measures.

[4]Described in 4.2.1
[5]Described in 4.2.1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.34**	**0.116**	**<0.05**	**-0.408**	**0.166**	**<0.01**
2	-0.23	0.053	0.158	0.049	0.002	0.768
3	**-0.378**	**0.143**	**<0.05**	0.069	0.005	0.679
4	-0.195	0.038	0.235	-0.101	0.01	0.542
5	-	-	-	-	-	-
6	-0.104	0.011	0.527	-0.222	0.049	0.175
7	-0.23	0.053	0.159	-0.063	0.004	0.705
8	-0.086	0.007	0.604	-0.123	0.015	0.454
9	-0.134	0.018	0.416	-0.106	0.011	0.523
10	-0.245	0.06	0.133	-0.11	0.012	0.507
11	-0.006	0	0.973	0.009	0	0.958

Table 7.4 Feature SDSD (45s) correlated with subjective ISA reports

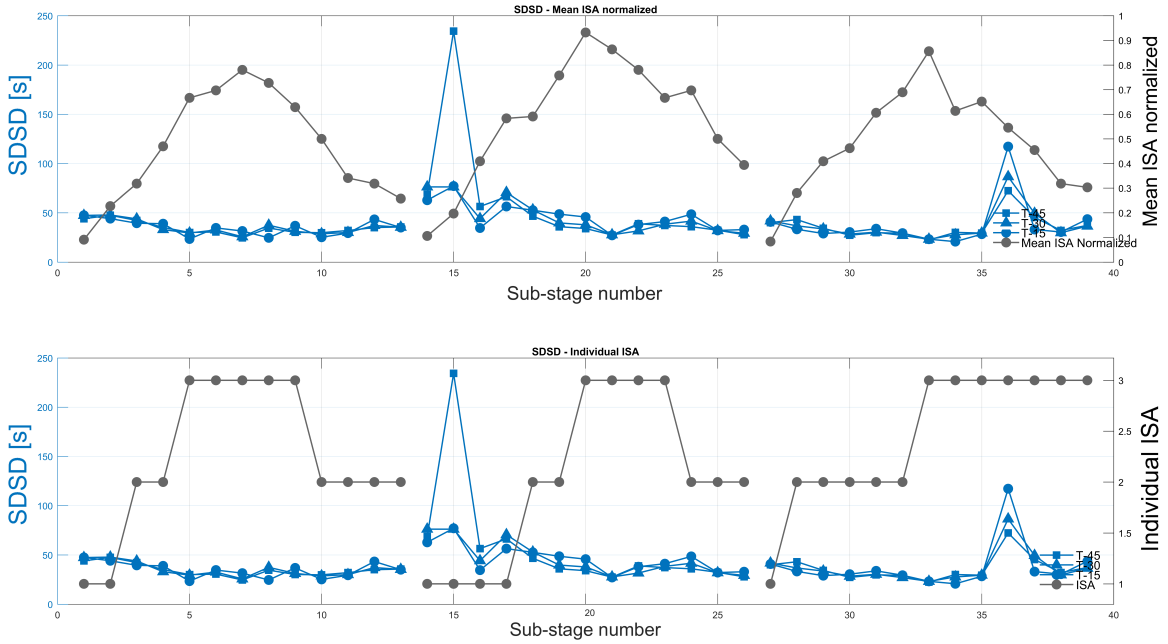

Fig. 7.9 SDSD - mean ISA normalized and individual ISA for Participant 1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.417**	**0.174**	**<0.01**	**-0.503**	**0.253**	**<0.01**
2	-0.226	0.051	0.167	0.051	0.003	0.758
3	**-0.347**	**0.121**	**<0.05**	0.089	0.008	0.589
4	-0.21	0.044	0.199	-0.062	0.004	0.707
5	-	-	-	-	-	-
6	-0.1	0.01	0.544	-0.204	0.042	0.212
7	-0.203	0.041	0.216	0.028	0.001	0.863
8	-0.092	0.008	0.579	-0.117	0.014	0.479
9	-0.122	0.015	0.459	-0.088	0.008	0.593
10	-0.313	0.098	0.052	-0.193	0.037	0.24
11	-0.003	0	0.985	0.016	0	0.924

Table 7.5 Feature RMSSD (45s) correlated with subjective ISA reports

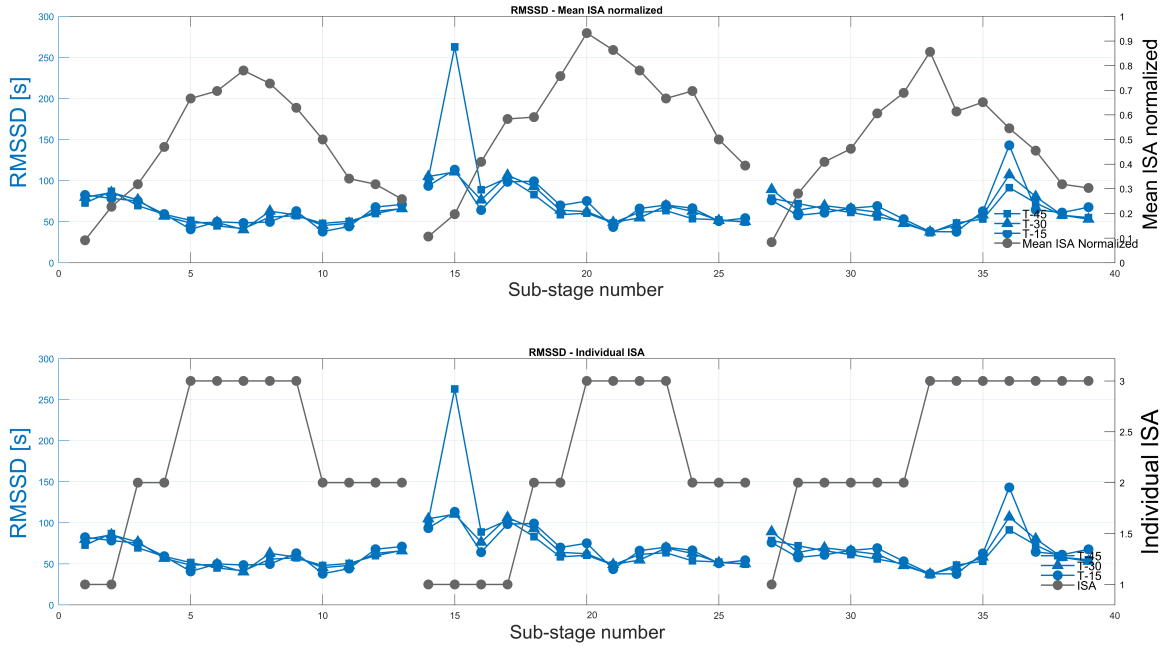

Fig. 7.10 RMSSD - mean ISA normalized and individual ISA for Participant 1

NN50 - Number of Successive R-R Differences that Differ by more than 50 milliseconds

The NN50[6] measure, showed comparable results when computed over 45s and 30s time intervals prior to the ISA sampling. Participants 1, 8 and 10 showed moderate correlations with the subjective data, Table 7.6. It is interesting to observe how the number of successive R-R differences that differ by more than 50 ms varies across the three time intervals it was computed over in Figure 7.11.

Figure 7.12 shows just the NN50 measure computed over the 30s time interval, appearing to respond very well to changes in demand.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**0.487**	**0.237**	**<0.01**	**0.571**	**0.326**	**<0.01**
2	-0.214	0.046	0.191	-0.204	0.042	0.212
3	0.039	0.002	0.812	**-0.422**	**0.178**	**<0.01**
4	0.043	0.002	0.794	-0.139	0.019	0.4
5	-	-	-	-	-	-
6	-0.096	0.009	0.561	0.063	0.004	0.702
7	0.034	0.001	0.838	0.294	0.086	0.07
8	**0.474**	**0.225**	**<0.01**	**0.493**	**0.243**	**<0.01**
9	0.054	0.003	0.744	-0.156	0.024	0.342
10	**0.469**	**0.22**	**<0.01**	**0.356**	**0.127**	**<0.05**
11	0.13	0.017	0.431	0.195	0.038	0.235

Table 7.6 Feature NN50 (45s) correlated with subjective ISA reports

pNN50 - Proportion of the NN50 from the total number of successive differences of the R-R data

The pNN50[7] measure, showed no significant correlation to the subjective data for any time interval it was computed on, except for a weak correlation from participant 7 with the individual ISA measure when pNN50 was computed over a 30s time interval, Table 7.7. Figure 7.13 displays the pNN50 data from participant 7 when computed over a 30s time interval; the data do not display a trend that would indicate there is a correlation to the variation in workload.

[6]Described in 4.2.1
[7]Described in 4.2.1

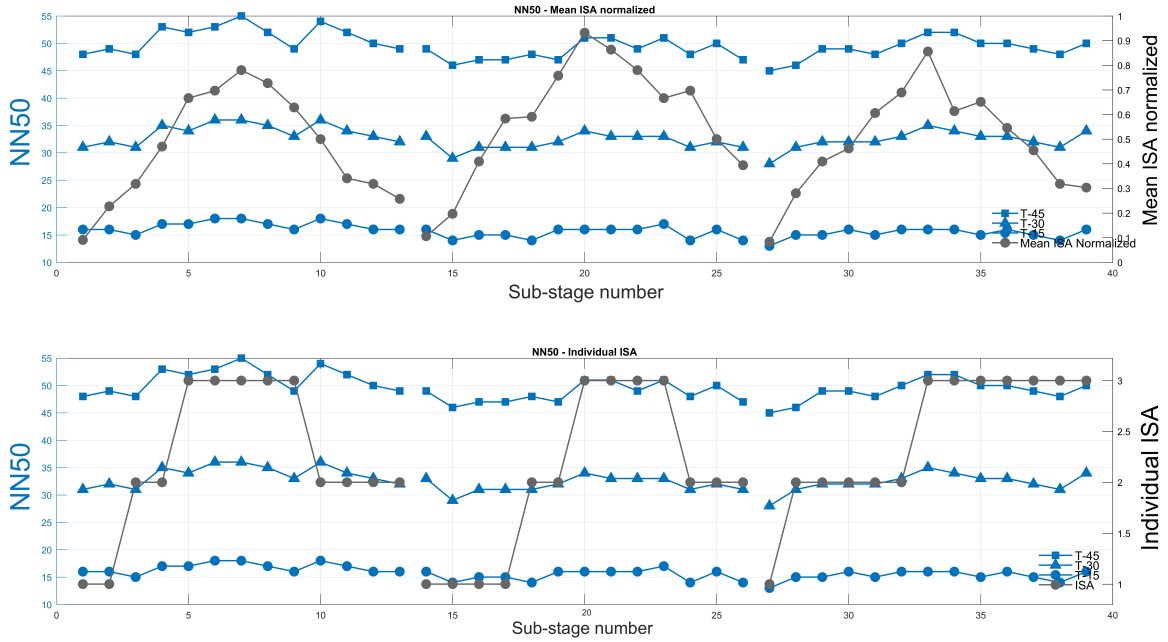

Fig. 7.11 NN50 - mean ISA normalized and individual ISA for Participant 1

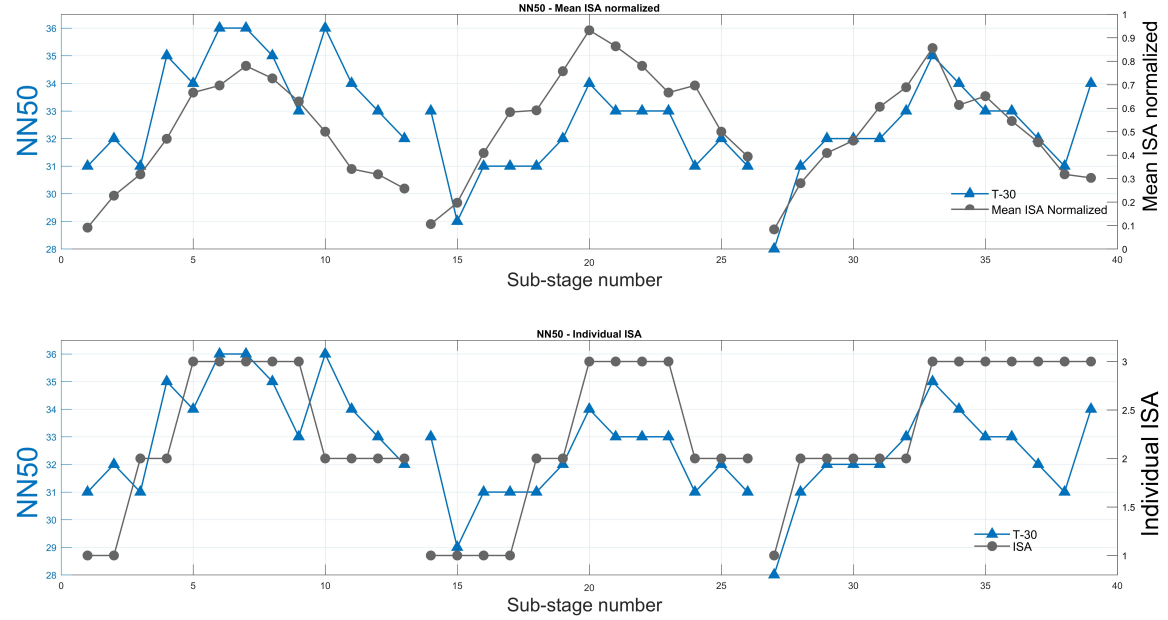

Fig. 7.12 NN50 [30s] - mean ISA normalized and individual ISA for Participant 1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	0.054	0.003	0.744	0.014	0	0.933
2	-0.07	0.005	0.67	0.073	0.005	0.661
3	0.083	0.007	0.614	0.094	0.009	0.569
4	0.024	0.001	0.886	0.236	0.056	0.148
5	-	-	-	-	-	-
6	-0.004	0	0.979	0.049	0.002	0.766
7	0.055	0.003	0.741	**0.357**	**0.128**	**<0.05**
8	0.064	0.004	0.699	0.122	0.015	0.459
9	-0.127	0.016	0.44	-0.113	0.013	0.492
10	0.084	0.007	0.61	0.059	0.004	0.72
11	-0.114	0.013	0.491	-0.104	0.011	0.53

Table 7.7 Feature pNN50 (30s) correlated with subjective ISA reports

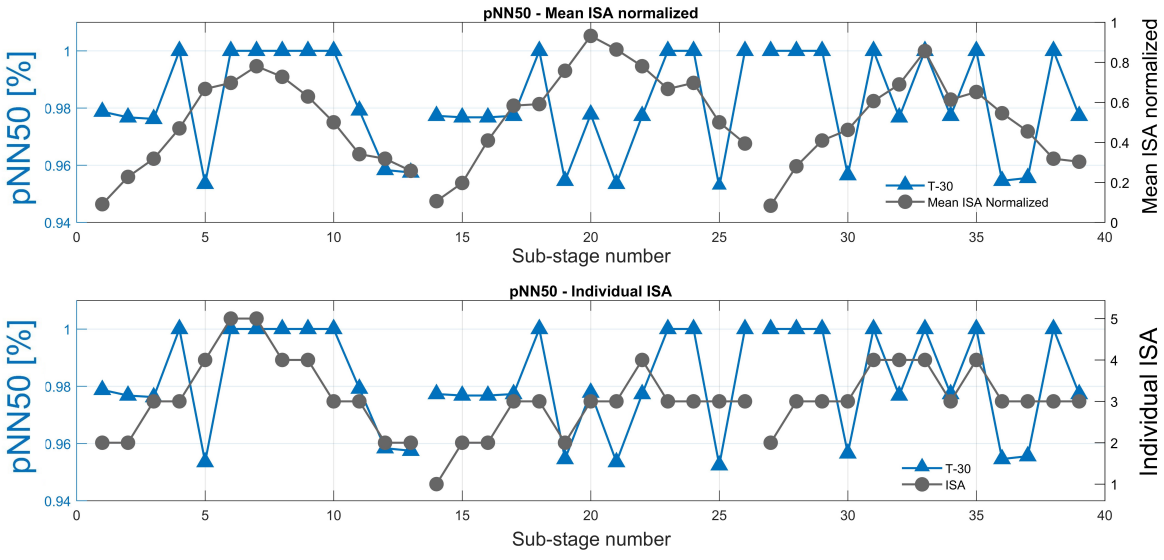

Fig. 7.13 pNN50 [30s] - mean ISA normalized and individual ISA for Participant 7

NN20 - Number of Successive R-R Differences that Differ by more than 20 milliseconds

The NN20[8] measure, showed comparable results when computed over 45s and 30s time intervals prior to the ISA sampling. Participants 1, 8 and 10 showed moderate correlations with the subjective data, Table 7.8. It is interesting to observe how the number of successive R-R differences that differ by more than 20 ms varies across the three time intervals it was computed over Figure 7.14.

Figure 7.15 shows just the NN20 measure computed over the 30s time interval, appearing to respond very well to changes in demand.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**0.487**	**0.237**	**<0.01**	**0.571**	**0.326**	**<0.01**
2	-0.214	0.046	0.191	-0.204	0.042	0.212
3	0.039	0.002	0.812	**-0.422**	**0.178**	**<0.01**
4	0.043	0.002	0.794	-0.139	0.019	0.4
5	-	-	-	-	-	-
6	-0.096	0.009	0.561	0.063	0.004	0.702
7	0.034	0.001	0.838	0.294	0.086	0.07
8	**0.474**	**0.225**	**<0.01**	**0.493**	**0.243**	**<0.01**
9	0.054	0.003	0.744	-0.156	0.024	0.342
10	**0.469**	**0.22**	**<0.01**	**0.356**	**0.127**	**<0.05**
11	0.13	0.017	0.431	0.195	0.038	0.235

Table 7.8 Feature NN20 (45s) correlated with subjective ISA reports

pNN20 - Proportion of the NN20 from the total number of successive differences of the R-R data

The pNN20[9] measure, showed no significant correlation to the subjective data for any time interval it was computed on, except for a weak correlation from participant 7 with the individual ISA measure when pNN20 was computed over a 30s time interval, Table 7.9. Figure 7.16 displays the pNN20 data from participant 7 when computed over a 30s time interval; the data do not display a trend that would indicate there is a correlation to the variation in workload.

[8]Described in 4.2.1
[9]Described in 4.2.1

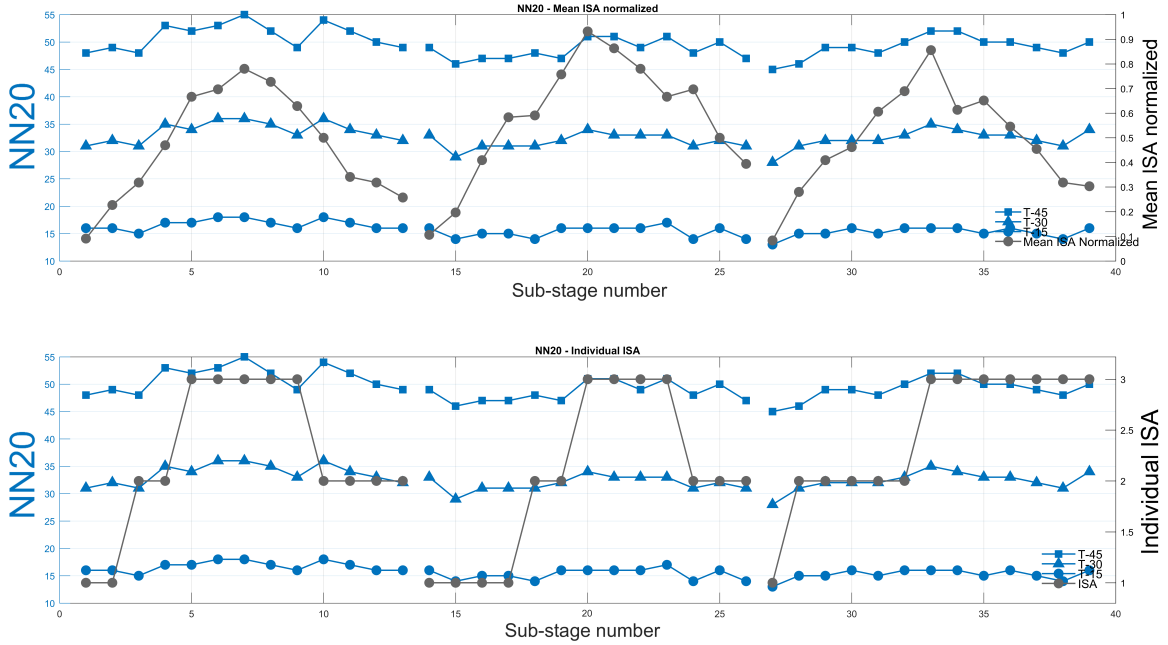

Fig. 7.14 NN20 - mean ISA normalized and individual ISA for Participant 1

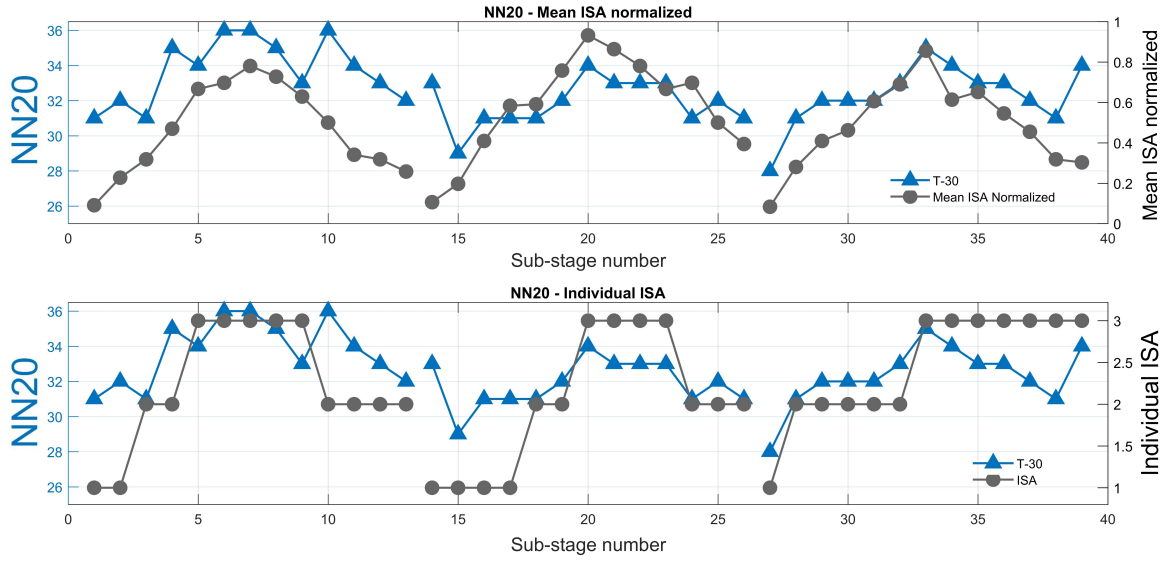

Fig. 7.15 NN20 [30s] - mean ISA normalized and individual ISA for Participant 1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	0.054	0.003	0.744	0.014	0	0.933
2	-0.07	0.005	0.67	0.073	0.005	0.661
3	0.083	0.007	0.614	0.094	0.009	0.569
4	0.024	0.001	0.886	0.236	0.056	0.148
5	-	-	-	-	-	-
6	-0.004	0	0.979	0.049	0.002	0.766
7	0.055	0.003	0.741	**0.357**	**0.128**	**<0.05**
8	0.064	0.004	0.699	0.122	0.015	0.459
9	-0.127	0.016	0.44	-0.113	0.013	0.492
10	0.084	0.007	0.61	0.059	0.004	0.72
11	-0.114	0.013	0.491	-0.104	0.011	0.53

Table 7.9 Feature pNN20 (30s) correlated with subjective ISA reports

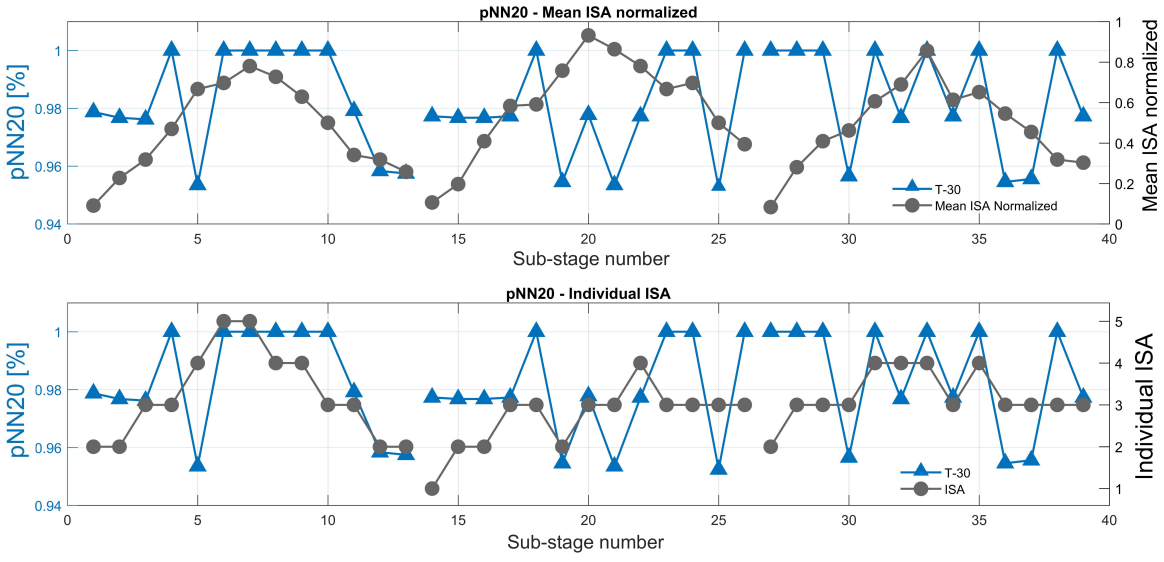

Fig. 7.16 pNN20 [30s] - mean ISA normalized and individual ISA for Participant 7

SD1 and SD2 - The Breadth of the Poincare Plot of the R-R Intervals Across and Along its identity line

The SD1 and SD2[10] measures, showed best correlation results when computed over 45s time intervals prior to the ISA sampling. Participants 1 was the only one to show moderate negative correlations with the subjective data for both SD1 and SD2 measures, Table 7.10 and Table 7.11. Participants 3 and 4 showed weak to moderate negative correlations to the mean ISA normalized data for SD1 respectively SD2 while participant 6 showed a weak negative correlation to the individual ISA measure. Figures 7.17 and 7.18 show the plots for the SD1 and SD2 measures for participant 1.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.416**	**0.173**	**<0.01**	**-0.503**	**0.253**	**<0.01**
2	-0.226	0.051	0.167	0.051	0.003	0.758
3	**-0.348**	**0.121**	**<0.05**	0.09	0.008	0.587
4	-0.21	0.044	0.2	-0.062	0.004	0.707
5	-	-	-	-	-	-
6	-0.101	0.01	0.54	-0.205	0.042	0.21
7	-0.202	0.041	0.217	0.029	0.001	0.862
8	-0.093	0.009	0.575	-0.117	0.014	0.478
9	-0.122	0.015	0.459	-0.088	0.008	0.593
10	-0.312	0.097	0.053	-0.192	0.037	0.242
11	-0.003	0	0.986	0.016	0	0.924

Table 7.10 Feature SD1 (45s) correlated with subjective ISA reports

[10]Described in 4.2.1

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	**-0.54**	**0.292**	**<0.01**	**-0.517**	**0.267**	**<0.01**
2	-0.154	0.024	0.349	0.084	0.007	0.61
3	-0.227	0.052	0.164	0.098	0.01	0.552
4	**-0.416**	**0.173**	**<0.01**	-0.277	0.077	0.088
5	-	-	-	-	-	-
6	-0.172	0.03	0.294	**-0.351**	**0.123**	**<0.05**
7	-0.286	0.082	0.078	-0.182	0.033	0.266
8	0.008	0	0.959	0.024	0.001	0.883
9	-0.13	0.017	0.432	-0.098	0.01	0.555
10	-0.191	0.036	0.245	-0.086	0.007	0.604
11	-0.026	0.001	0.876	-0.004	0	0.979

Table 7.11 Feature SD2 (45s) correlated with subjective ISA reports

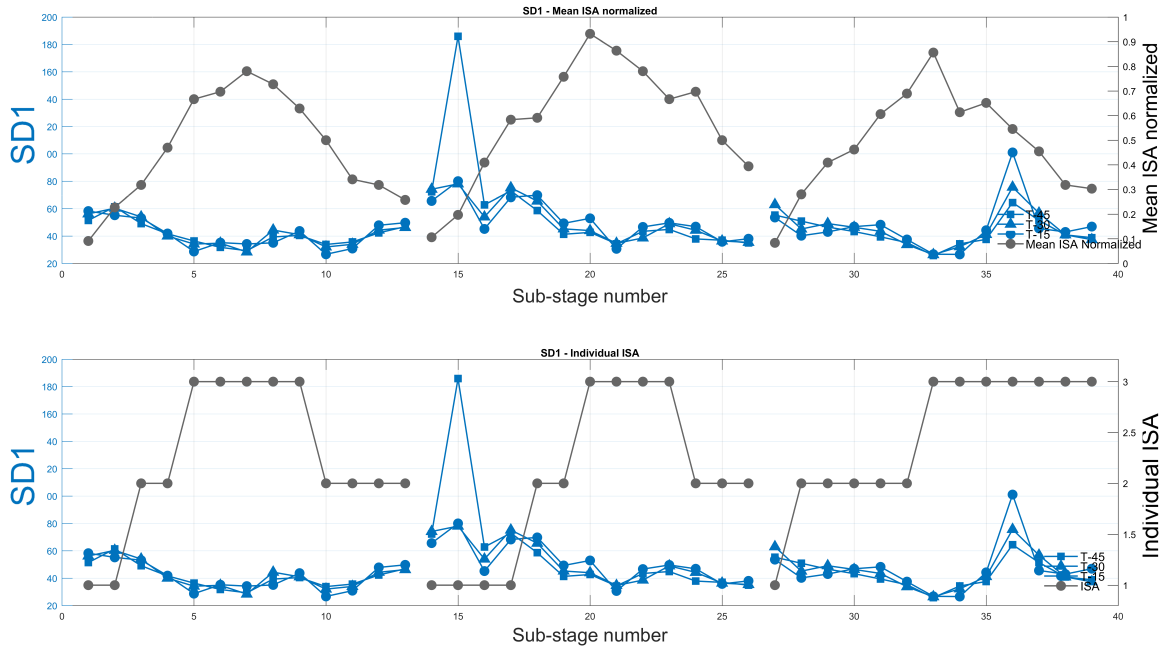

Fig. 7.17 SD1 [45s] - mean ISA normalized and individual ISA for Participant 1

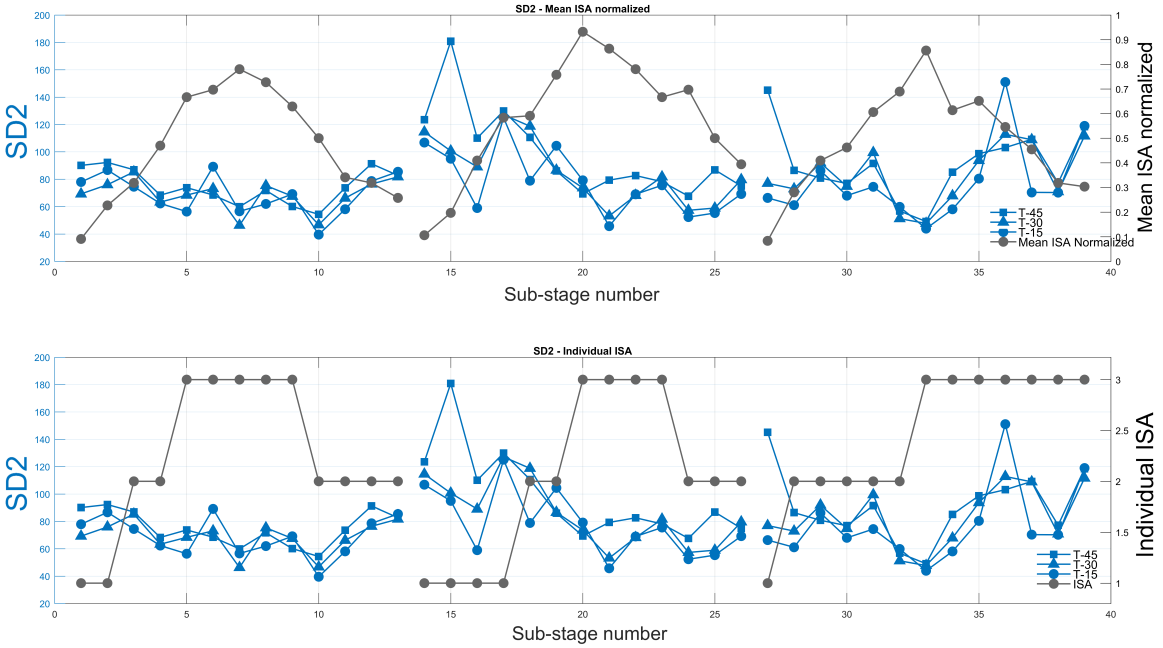

Fig. 7.18 SD2 [45s] - mean ISA normalized and individual ISA for Participant 1

Mean Breathing Rate

The mean breathing rate measures, showed best correlation results when computed over 45s time intervals prior to the ISA sampling. Most participants showed positive correlations, the breathing rate increasing as workload increased. Participant 4 showed a moderate positive correlation with both mean ISA normalized and individual ISA while participant 8 showed a moderate positive correlation with mean ISA normalized and a weak positive correlation with the individual ISA. Table 7.12 lists the correlations for all participants while Figure 7.19 show the plots for the mean breathing rate measure for participant 4.

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	0.195	0.038	0.233	**0.384**	**0.148**	**<0.05**
2	**0.316**	**0.1**	**<0.05**	**0.365**	**0.133**	**<0.05**
3	0.286	0.082	0.077	-0.288	0.083	0.076
4	**0.524**	**0.275**	**<0.01**	**0.513**	**0.264**	**<0.01**
5	-	-	-	-	-	-
6	0.296	0.088	0.067	**0.342**	**0.117**	**<0.05**
7	0.274	0.075	0.092	0.184	0.034	0.263
8	**0.594**	**0.353**	**<0.01**	**0.359**	**0.129**	**<0.05**
9	-0.276	0.076	0.089	-0.227	0.052	0.164
10	-0.234	0.055	0.151	-0.283	0.08	0.081
11	0.024	0.001	0.884	-0.016	0	0.925

Table 7.12 Feature Mean-BR (45s) correlated with subjective ISA reports

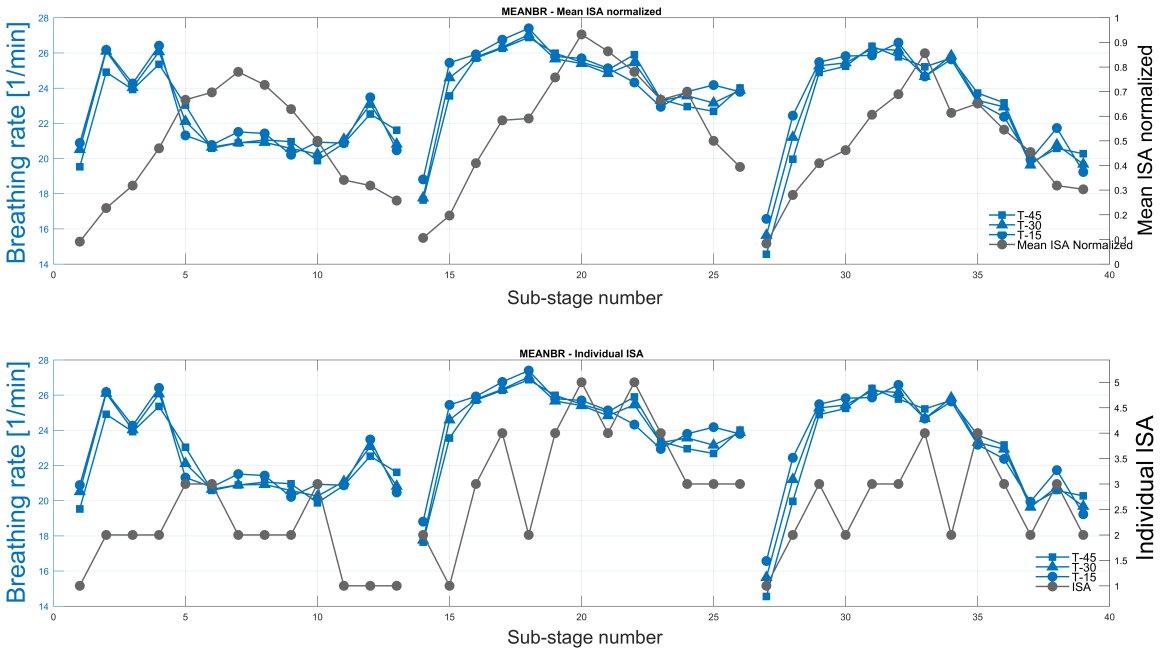

Fig. 7.19 Mean BR [45s] - mean ISA normalized and individual ISA for Participant 4

Facial thermography

This section introduces the results obtained from facial thermography. Some errors in landmark tracking for certain areas of the face led to noisy data that were removed from the analysis. All data presented were averaged over the 45s time intervals before the ISA ratings. Overall, the common face areas (for participants not wearing glasses and wearing glasses) temperature was collected from, was classified in four areas: nose, below nose, lips and cheeks. From each of these areas, temperature was sampled from points, lines and areas and will be presented further in order to extract the most meaningful data that will later be considered for the multiple regression analysis.

Nose area temperatures

From the nose area, temperatures were sampled and analysed from the following regions of interest (the labels can be found in Figure 7.20):

- Points: 22, 23, 24, 25, 26, 27, 28, 29
- Lines: 22-23, 22-25, 22-29, 22-24, 23-24, 23-25, 23-29, 24-25, 24-29, 24-27
- Areas: 22-23-25, 23-24-25, 24-25-26-27, 22-23-29, 23-24-29, 24-29-28-27, 22-25-24-29, 24-25-26-27-28-29

One observation to be made regarding the temperature inside the points is that for 8 out of 11 participants, nose temperature decreased starting from point 22 (top of the nose) to point 24 (nose tip). For 7 out of 11 participants, point 24 offered the best response compared to the other points, while for the other 4 participants, point 22 provided the best response. As expected from the points results, regarding the line temperatures, for most participants, lines 23-24 and 24-27 had the lowest temperature. Line 22-24 characterizes temperature along the entire length of the nose and also provides one of the best responses to variation of demand for all participant. For most participants, all their nose areas responded in a similar way. The one to be further analysed is area 22-25-24-29 as it covers almost the entire area of the nose excepting the nostrils area whose temperature is affected by breathing. The data to be considered for further analysis are those from point 24, line 22-24 and area 22-25-24-29.

Figure 7.21 provides an example of how temperature varied in the nose area, in the three regions of interest mentioned above, for Participant 1. Tables 7.13, 7.14 and 7.15 show the individual correlations for each participant with both mean ISA normalized and the individual ISA values. Three out of eleven participants showed significant medium negative correlations with the mean ISA normalized while one of the participants showed a positive medium correlation with the same measure. Not all participants had significant correlations with the ISA values, nevertheless another pattern emerged, for 8 out of 11, the temperatures

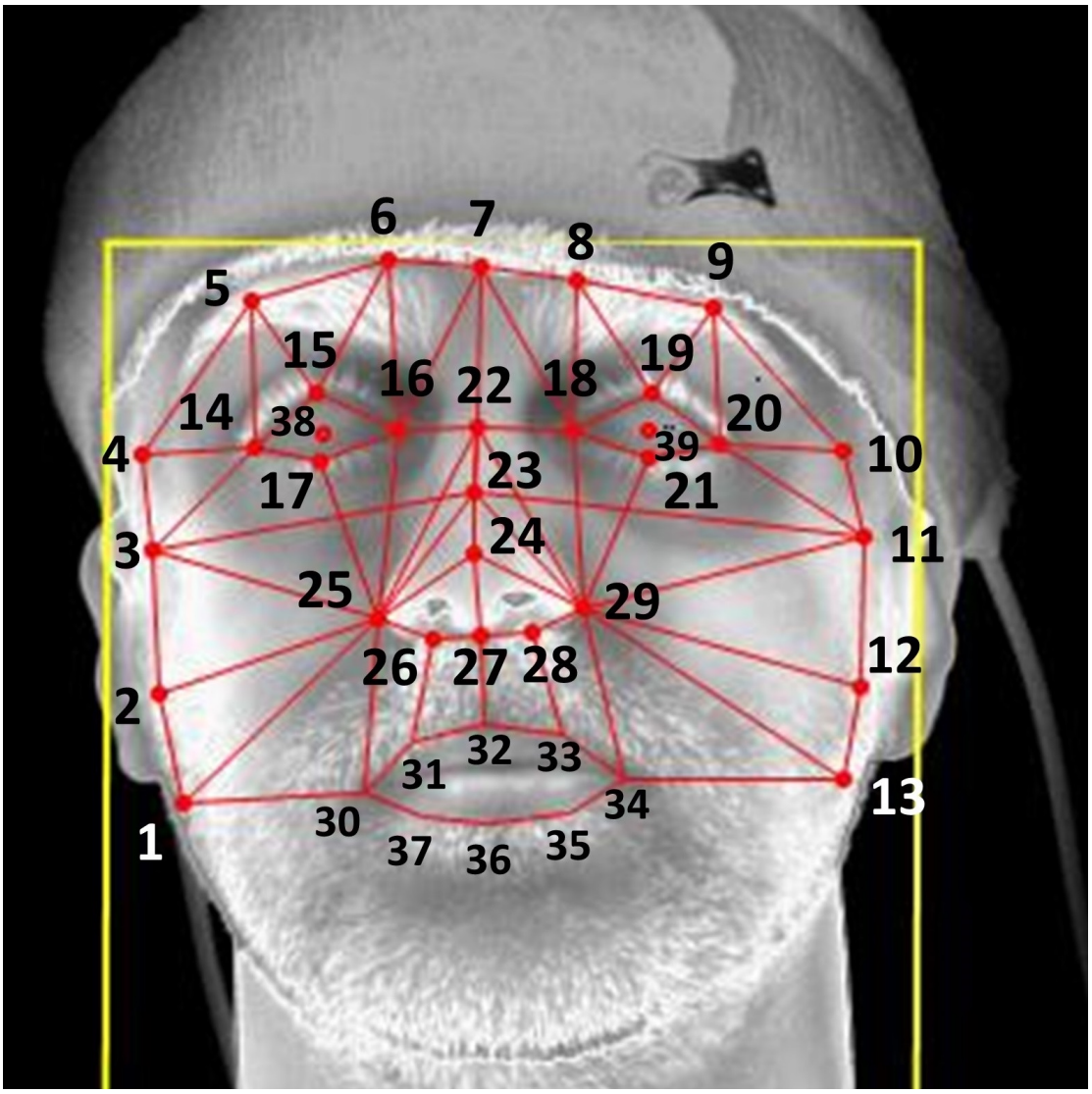

Fig. 7.20 Facial landmarks

in the above mentioned regions of interest decreased during high demand intervals, for 2 out of 11 the temperature remained relatively constant while for 1 out of 11 there was an increase in temperature during higher demand periods. Some participants presented overall slight increases in temperature, over the entire duration of the study, one such example is Participant 1, as it can be seen in Fig. 7.21.

Below nose area temperatures

From the area below the nose, temperatures from the following regions of interest were sampled and analysed (the labels can be found in Figure 7.20):

- Lines: 26-31, 27-32, 28-33

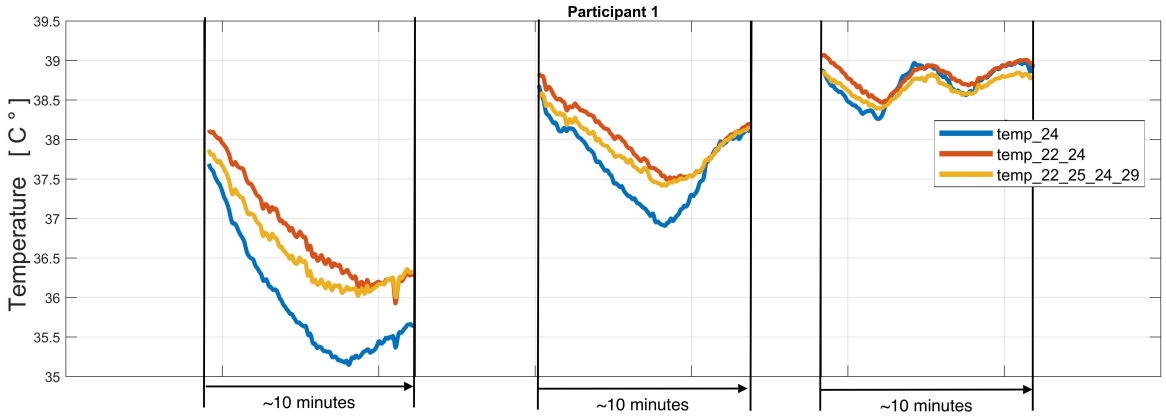

Fig. 7.21 Nose temperature example for Participant 1, sampled from Point 24, Line 22-24 and Area 22-25-24-29. The data show a decrease in temperature with the increase in demand for all three areas and a slight increase in base temperature for each of the stages

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-0.17	0.029	0.299	-0.13	0.017	0.431
2	**-0.345**	**0.119**	**<0.05**	-0.174	0.03	0.289
3	-0.102	0.01	0.537	**0.324**	**0.105**	**<0.05**
4	**-0.373**	**0.139**	**<0.05**	-0.279	0.078	0.085
5	-0.217	0.047	0.185	0.11	0.012	0.504
6	-0.074	0.005	0.656	-0.213	0.045	0.194
7	-0.065	0.004	0.696	-0.009	0	0.957
8	**-0.496**	**0.246**	**<0.01**	-0.166	0.028	0.311
9	-0.258	0.066	0.113	**-0.346**	**0.119**	**<0.05**
10	-0.054	0.003	0.744	0.014	0	0.93
11	**0.421**	**0.177**	**<0.01**	**0.34**	**0.116**	**<0.05**

Table 7.13 Point 24 temperature (45s) correlated with subjective ISA reports

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-0.178	0.032	0.279	-0.147	0.022	0.372
2	**-0.453**	**0.205**	**<0.01**	-0.178	0.032	0.279
3	0.141	0.02	0.391	**0.435**	**0.189**	**<0.01**
4	**-0.46**	**0.211**	**<0.01**	**-0.323**	**0.104**	**<0.05**
5	-0.259	0.067	0.112	0.072	0.005	0.662
6	0.179	0.032	0.275	-0.039	0.002	0.812
7	-0.008	0	0.96	0.009	0	0.958
8	**-0.344**	**0.118**	**<0.05**	-0.128	0.016	0.439
9	-0.162	0.026	0.324	-0.182	0.033	0.267
10	0.289	0.083	0.075	**0.389**	**0.151**	**<0.05**
11	**0.452**	**0.204**	**<0.01**	**0.321**	**0.103**	**<0.05**

Table 7.14 Line 22-24 temperature (45s) correlated with subjective ISA reports

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-0.168	0.028	0.308	-0.134	0.018	0.416
2	**-0.446**	**0.199**	**<0.01**	-0.201	0.04	0.221
3	0.101	0.01	0.542	**0.409**	**0.167**	**<0.01**
4	**-0.504**	**0.254**	**<0.01**	**-0.332**	**0.11**	**<0.05**
5	-0.274	0.075	0.091	0.025	0.001	0.878
6	0.236	0.056	0.148	0.008	0	0.963
7	-0.043	0.002	0.794	-0.004	0	0.979
8	**-0.72**	**0.518**	**<0.01**	**-0.385**	**0.148**	**<0.05**
9	-0.051	0.003	0.757	-0.103	0.011	0.534
10	**0.322**	**0.104**	**<0.05**	**0.351**	**0.123**	**<0.05**
11	0.276	0.076	0.089	0.201	0.04	0.22

Table 7.15 Area 22-25-24-29 temperature (45s) correlated with subjective ISA reports

- Areas: 25-26-31-30, 26-27-32-31, 28-29-34-33, 27-28-33-32, 25-26-27-28-29-34-33-32-31-30

Firstly, the temperature sampled from the above mentioned lines will be presented. Overall, the results showed a decrease in temperature on these lines for 8 out of 11 participants, an increase for one participant and relatively constant temperatures for 2 participants. For 5 out of 11 participants, the temperature on line 27-32 (middle) was lower than the other two lines; also for 5 out of 11 participants the temperatures on lines 26-31 and 28-33 were not similar as it might be expected due to the symmetry. As patterns in temperature variation were similar for all three lines and line 27-32 is the least affected by the temperature variations induced by breathing, it will be chosen to further analysis. Figure 7.22 shows an example of how temperatures on the lines 26-31, 27-32, 28-32 varied during the task for participant 4. Temperature inside the analysed areas responded in a similar manner. Out of the four analysed areas below the nose, area 25-26-27-28-29-34-33-32-31-30 will be analysed further as it covers the entire area and it is less likely to be influenced as much as the others by just the breathing air temperature.

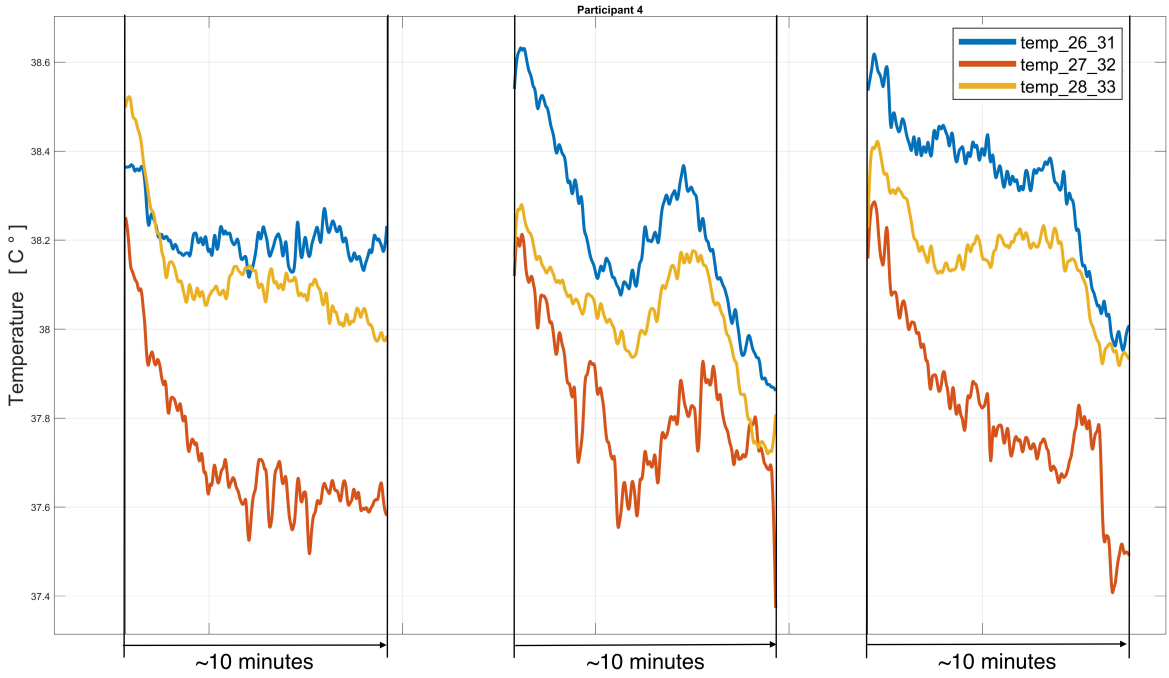

Fig. 7.22 Below nose temperatures example for Participant 4. The data show that the temperature decreased as demand increased; also we can see that after the break in between the stages, the temperature goes back to approximately the initial baseline

Figure 7.23 shows comparison of the temperature response between line 27-32 and the entire area below the nose for participant 4. The drop in temperature is larger for line 27-32

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-0.121	0.015	0.463	0.103	0.011	0.534
2	**-0.537**	**0.288**	**<0.01**	**-0.36**	**0.13**	**<0.05**
3	0.019	0	0.909	**0.544**	**0.296**	**<0.01**
4	**-0.493**	**0.243**	**<0.01**	-0.2	0.04	0.222
5	**-0.492**	**0.242**	**<0.01**	**-0.365**	**0.133**	**<0.05**
6	-0.025	0.001	0.878	0.006	0	0.971
7	-0.153	0.023	0.353	-0.191	0.036	0.245
8	**-0.661**	**0.436**	**<0.01**	**-0.323**	**0.104**	**<0.05**
9	-0.033	0.001	0.842	0.02	0	0.903
10	**-0.437**	**0.191**	**<0.01**	**-0.392**	**0.154**	**<0.05**
11	0.18	0.032	0.273	0.22	0.049	0.177

Table 7.16 Line 27-32 temperature (45s) correlated with subjective ISA reports

but the data collected from the entire area show less noise as they are less sensitive to facial landmark tracking errors.

Lip area temperatures

From the lips area, temperatures from the following regions of interest were sampled and analysed (the labels can be found in Figure 7.20):

- Points: 32, 36
- Lines: 30-31, 31-32, 32-33, 33-34, 34-35, 35-36, 36-37, 37-30

The analysed interest area did not show a clear and consistent pattern, neither across participants or even sometimes within the data from a single participant. Figure 7.24 shows an example of how temperature varied for points 32 and 36 for participant 2. It can be observed that during both the first and last stages temperature increased while for the second stage it decreased. This pattern is not consistent across participants. One other factor influencing the results from this area is the lower quality in landmark tracking. Data from the lip area will not be considered for further analysis.

Cheek area temperatures

From the cheek area, temperatures from the following regions of interest were sampled and analysed (the labels can be found in Figure 7.20):

- Areas: 1-25-30, 1-2-25, 2-3-25, 13-29-34, 12-13-29, 11-12-29

For most participants, the cheek areas that were analysed did not show consistent changes in temperature across the participants as a response to changes in workload. The reactions

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-0.099	0.01	0.549	0.052	0.003	0.752
2	**-0.535**	**0.286**	**<0.01**	**-0.397**	**0.158**	**<0.05**
3	0.044	0.002	0.792	**0.53**	**0.281**	**<0.01**
4	-0.214	0.046	0.19	-0.139	0.019	0.399
5	**-0.331**	**0.11**	**<0.05**	**-0.394**	**0.155**	**<0.05**
6	**-0.342**	**0.117**	**<0.05**	-0.28	0.079	0.084
7	-0.203	0.041	0.216	**-0.317**	**0.101**	**<0.05**
8	**-0.611**	**0.373**	**<0.01**	**-0.329**	**0.108**	**<0.05**
9	0.029	0.001	0.859	0.093	0.009	0.574
10	**-0.359**	**0.129**	**<0.05**	-0.299	0.089	0.065
11	0.18	0.033	0.272	0.201	0.04	0.219

Table 7.17 Area 25-26-27-28-29-34-33-32-31-30 temperature (45s) correlated with subjective ISA reports

ranged from increases in temperature, decreases or relatively constant temperature. Figure 7.25 shows the temperature variation in area 13-29-34 for participant 8, a drop in temperature reaction similar to the nose areas. For two of the participants, the symmetry between the right and left cheeks was not maintained as well. Data from the lower cheek areas were in many cases influenced by the presence of facial hair, inducing noise. For these reasons, the data collected from these regions will not be considered for further analysis.

fNIRS

fNIRS data from two participants (participant 1 and participant 10) were not recorded due to technical difficulties; also not all optodes provided high quality data, most likely due to improper contact with the scalp. Optodes: 1, 2, 7, 9, 11 and 13 recorded data from all the remaining 9 participants. In general, data from all optodes indicated an increase in blood oxygenation with the increase in workload.

Blood oxygenation measured by most optodes showed strong positive correlations with the mean ISA normalized values. Tables 7.18 and 7.19 show the correlations between the oxygenation values for each participant and both mean ISA normalized and the individual ISA values. Both Optode 6 and 15 had strong correlations for most participants with the mean ISA normalized. Figure 7.26 illustrates an example of strong correlation of the blood oxygenation with the mean ISA normalized values for Participant 4, Optode 15.

Figure 7.27 shows an example of strong correlation ($r(37)=0.563, p<0.01$) of the blood oxygenation with the mean ISA normalized values for Participant 2, Optode 13. Apart from being strongly correlated with the subjective ratings, the data from Participant 2, including

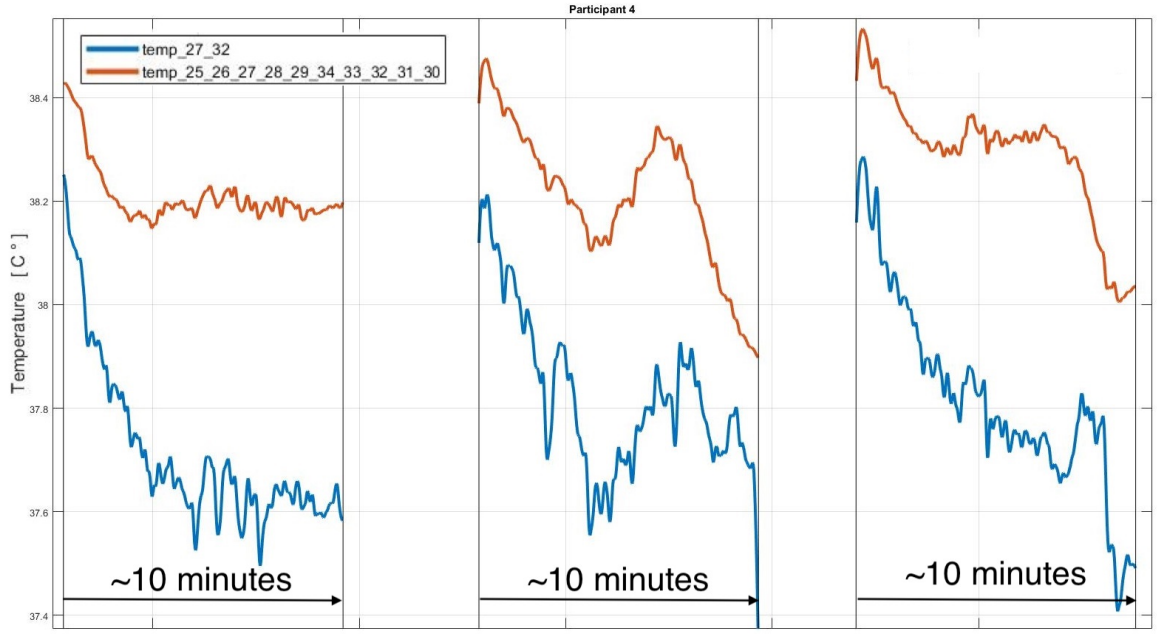

Fig. 7.23 Line 27-32 compared to below nose area for Participant 4. The data show that the decrease in temperature is larger for line 27-32 but it has a similar response to the entire area below the nose

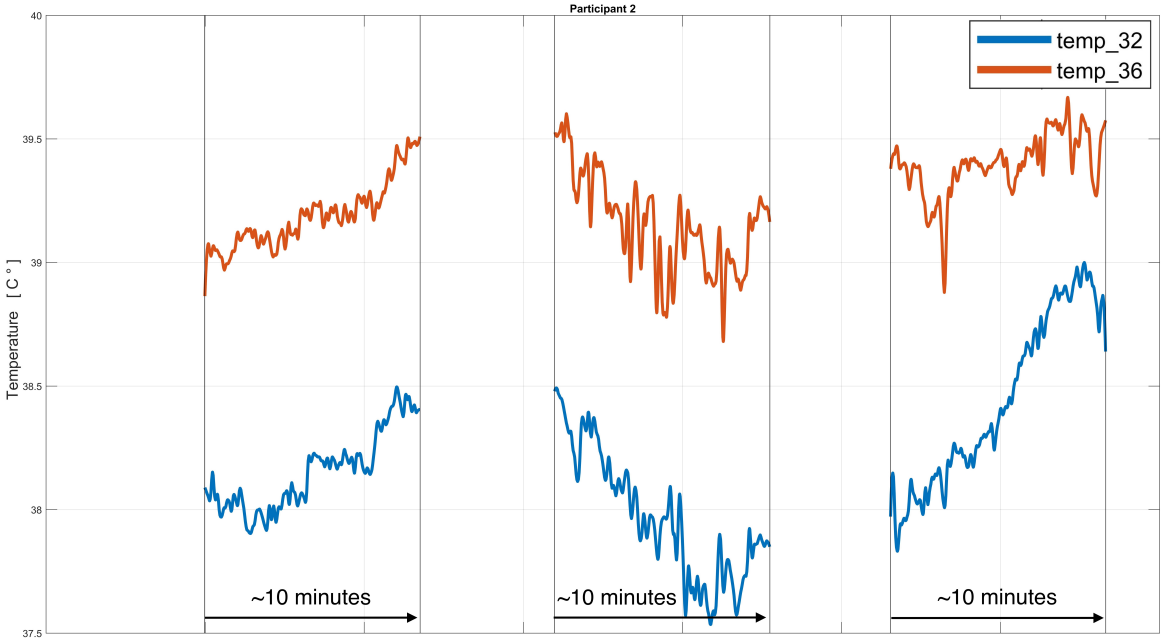

Fig. 7.24 Points 32 and 36 for Participant 2. The data show an increase in temperature for stags 1 and 3 and a decrease for stage 2 that was considered more mentally demanding. The pattern was not consistent with the other participants

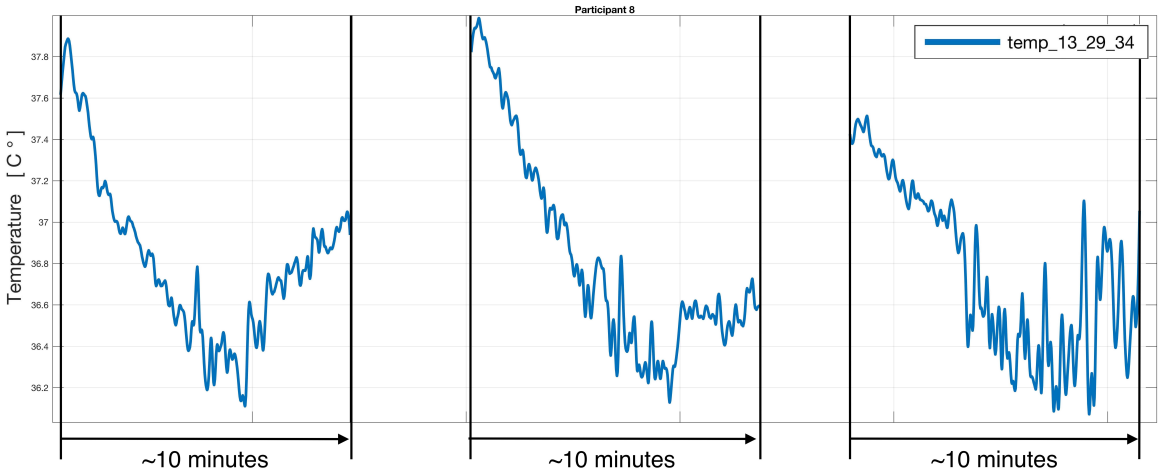

Fig. 7.25 Temperature variation in area 13-29-34 for Participant 8, showing a decrease in temperature with the increase in demand, a response similar to the nose areas; this reaction was not consistent with that of the other participants

optodes 5,7,9,11,13,15 show an increased level of blood oxygenation for stage 2, in most cases almost clearly differentiating it from stages 1 and 3. This indicates that in this case blood oxygenation could differentiate between the stage that was more mentally demanding (the odd number stage) and the stages that were less mentally demanding (the red balls stages).

Figure 7.28 shows an example of a weak correlation (r(37)=-0.074,p=0.65) between the blood oxygenation levels and the mean ISA normalized subjective ratings for Participant 3, optode 7.

Multiple physiological measures as an indication of workload level

The previous sections have looked at the correlation of each of the physiological measures with the subjective measures of workload. Some physiological measures have shown higher and some have shown lower or no correlation at all to the subjective workload measures. This was to be expected, as different measures could show senstivity to different aspects of workload or may have a different bandwidth in which they are sensitive.

This section explores how the different physiological measures can be combined to produce a more accurate measure of workload. Some of the measures presented above show promising correlations to the subjective ISA measure of mental workload. A multiple linear

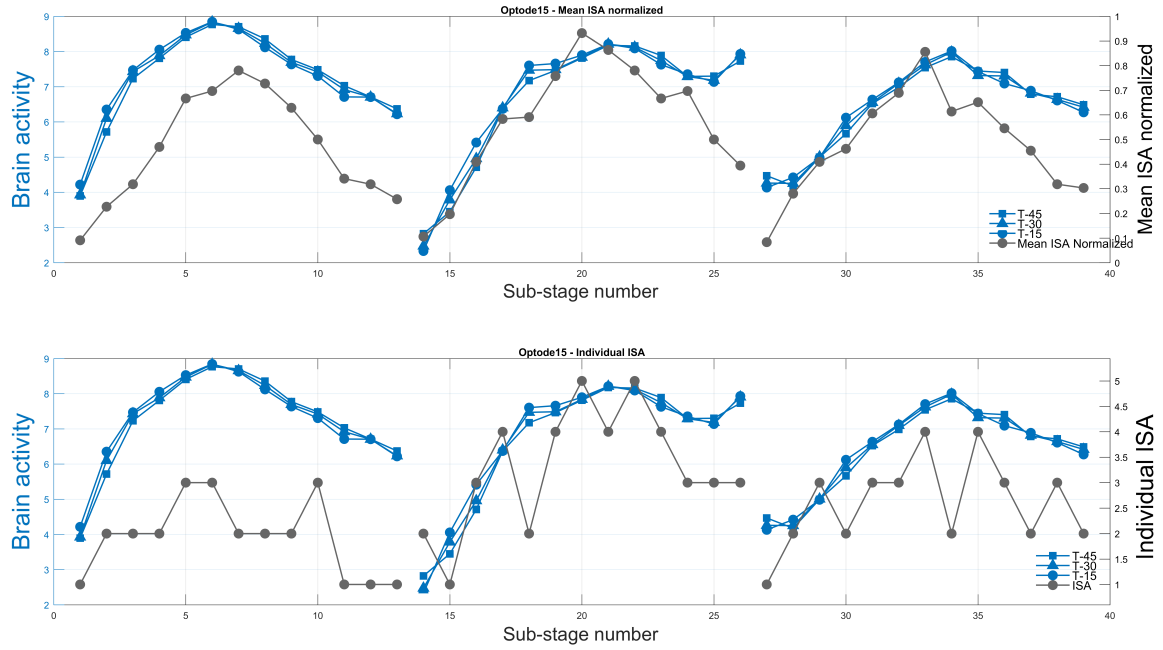

Fig. 7.26 Example of change in blood oxygenation for Participant 4, Optode 15; the data present a strong correlation of the blood oxygenation with the mean ISA normalized values ($r(37)=0.779, p<0.01$)

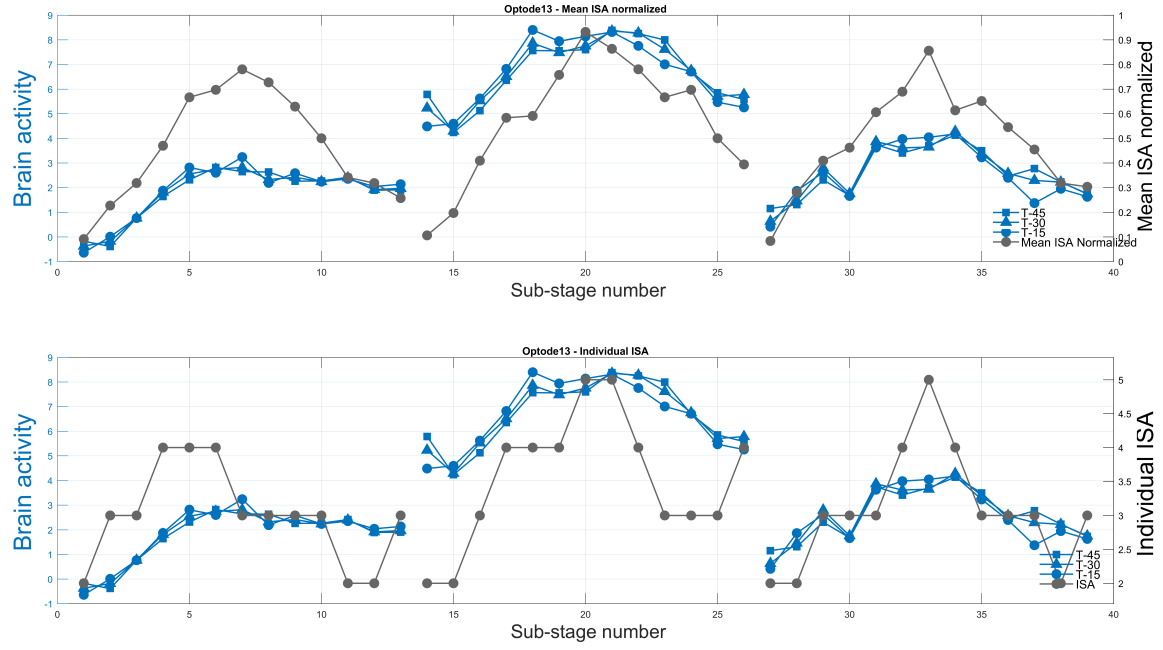

Fig. 7.27 Example of change in blood oxygenation for Participant 2, Optode 13, showing a strong correlation ($r(37)=0.563, p<0.01$) with the mean ISA normalized values; also the data show a clear difference in blood oxygenation for stage 2 as compared to stages 1 and 3

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-	-	-	-	-	-
2	-	-	-	-	-	-
3	**0.586**	**0.343**	**<0.01**	**0.433**	**0.187**	**<0.01**
4	**0.691**	**0.478**	**<0.01**	**0.418**	**0.175**	**<0.01**
5	-	-	-	-	-	-
6	**0.542**	**0.294**	**<0.01**	0.223	0.05	0.172
7	**0.685**	**0.469**	**<0.01**	**0.582**	**0.339**	**<0.01**
8	**0.575**	**0.33**	**<0.01**	**0.36**	**0.13**	**<0.05**
9	**0.665**	**0.442**	**<0.01**	**0.719**	**0.517**	**<0.01**
10	-	-	-	-	-	-
11	**0.58**	**0.336**	**<0.01**	**0.489**	**0.239**	**<0.01**

Table 7.18 fNIRS: Optode 6 (45s) correlated with subjective ISA reports

Participant No.	Mean ISA Normalized			Individual ISA		
	r(37)	r^2	p	r(37)	r^2	p
1	-	-	-	-	-	-
2	**0.693**	**0.481**	**<0.01**	**0.59**	**0.348**	**<0.01**
3	**0.372**	**0.138**	**<0.05**	0.077	0.006	0.643
4	**0.779**	**0.607**	**<0.01**	**0.452**	**0.204**	**<0.01**
5	-	-	-	-	-	-
6	**0.503**	**0.253**	**<0.01**	**0.481**	**0.232**	**<0.01**
7	**0.746**	**0.556**	**<0.01**	**0.547**	**0.299**	**<0.01**
8	**0.759**	**0.576**	**<0.01**	**0.335**	**0.112**	**<0.05**
9	-	-	-	-	-	-
10	-	-	-	-	-	-
11	**0.684**	**0.468**	**<0.01**	**0.581**	**0.338**	**<0.01**

Table 7.19 fNIRS: Optode 15 (45s) correlated with subjective ISA reports

regression was performed for each participant individually on more combinations of the predictor variables to test which one explains more of the variability in the response variable and how different physiological parameters can be combined for more reliable and valid capture of workload. Seven combinations of the predictor variables were chosen:

1. Heart (R-R interval) and Breathing Rate data (Mean RR, Mean BR)
2. Thermal data
3. fNIRS data - blood oxygenation
4. Heart (R-R interval), Breathing Rate and Thermal data

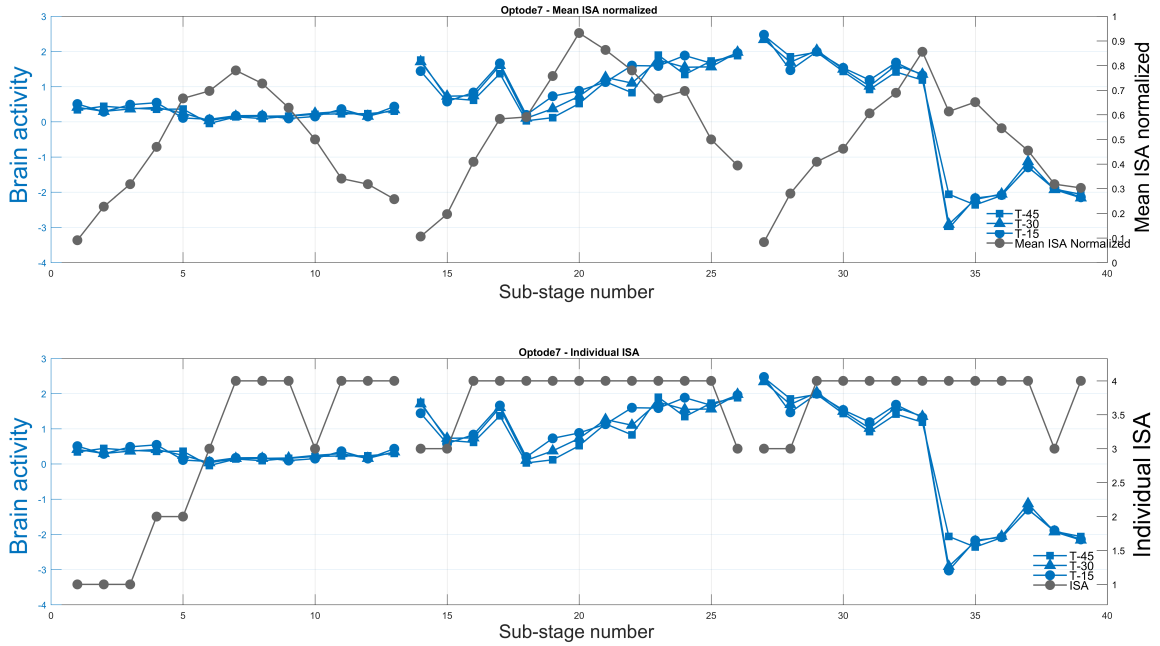

Fig. 7.28 Example of change in blood oxygenation for Participant 3, Optode 7, showing an example of a weak correlation (r(37)=-0.074,p=0.65) between the blood oxygenation levels and the mean ISA normalized subjective ratings

5. Heart (R-R interval), Breathing Rate and fNIRS data - blood oxygenation
6. Thermal data and fNIRS data - blood oxygenation
7. Heart (R-R interval), Breathing Rate, Thermal data fNIRS data - blood oxygenation

The reason behind the choice of the predictor variables combinations was to start with features from only one of the sensor and gradually add the others; category 1 contains just the features produced by the Zephyr sensor, category 2 contains just facial thermography data, category 3 contains just blood oxygenation data from the fNIRS sensor, category 4 combines categories 1 and 2, category 5 combines categories 1 and 3, category 6 combines categories 2 and 3 and category 7 combines data from all the sensors.

It has to be stated that performance measures are not expected to correlate to either subjective of physiological measures, as performance can be protected and therefore maintained at the same level through investing more resources in accomplishing the task. However, in the particular case of the task used for this study, performance showed a high correlation to subjective measures, being in a way almost a measure of workload in itslef; this should not be generalized on any task. For this reason game performance, rather than ISA ratings, was selected as the response variable for this analysis. Game performance was used as it strongly correlates with the subjective ISA ratings and it is also a continuous variable that was needed

for the analysis; game performance is represented by the height the participants managed to maintain the yellow line on the screen.

Some of the predictor variables for some of the participants were highly correlated to each other. Inter-variable correlation influences the ability of multiple linear regression to distinguish between the predictive ability of each individual variable. Our approach to this limitation was to systematically add and remove predictors based on the F-statistic; the tool used for this was stepwise regression in Matlab. The algorithm starts with a constant model and iteratively adds and removes predictors until the model can no longer be improved substantially. One table will be presented for each of the combinations, each of the sections in Tables 7.20, 7.21, 7.22, 7.23, 7.24, 7.25, 7.26 shows the multiple linear regression results for each of the described groups of predictors for each participant. The adjusted r^2 column, contains the proportion of variability of the dependent variable accounted for by the regression model. Because the r^2 value increases by adding more predictor variables in the model, the adjusted r^2 value was reported in order to make the comparison between models more meaningful. The table also displays the F statistic of the linear fit versus the constant model, testing the statistical significance of the model; the predictors column contains the names of the predictors selected by the algorithm for each of the regressions. The Beta column contains the estimate standardized coefficients of the terms in the regression, indicating how many standard deviations the dependent variable will change with the change of one standard deviation in the predictor variable, allowing for a comparison of the relative contribution of each of the predictors. The t-statistic test for the significance of each term given the other terms in the model is used to test the null hypothesis that the term is equal to zero (versus the alternate hypothesis that the coefficient is different from zero). The associated p values are also reported in the table.

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0.17	0.91	8.84	<0.01	Mean RR	0.44	2.97	<0.01
2	0	-	-	-	-	-	-	-
3	0	-	-	-	-	-	-	-
4	0.27	0.85	15.03	<0.01	Mean BR	-0.54	-3.88	<0.01
5	0	-	-	-	-	-	-	-
6	0.19	0.9	10.2	<0.01	Mean BR	-0.46	-3.19	<0.01
7	0.16	0.92	8.33	<0.01	Mean BR	-0.43	-2.89	<0.01
8	0.19	0.9	10.17	<0.01	Mean RR	0.46	3.19	<0.01
9	0	-	-	-	-	-	-	-
10	0.16	0.92	8.37	<0.01	Mean RR	0.43	2.89	<0.01
11	0	-	-	-	-	-	-	-

Table 7.20 Proportion of the variability accounted for by the regression model in the response variable using combination 1 predictors (Heart Rate, Breathing rate)

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0	-	-	-	-	-	-	-
2	0.51	0.7	21.11	<0.01	Line 22-24	0.48	3.72	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.38	2.92	<0.01
3	0	-	-	-	-	-	-	-
4	0.54	0.68	22.99	<0.01	Point 24	-2.31	-4.95	<0.01
					Area 22-25-24-29	2.76	5.9	<0.01
5	0.32	0.83	18.53	<0.01	Line 27-32	0.58	4.3	<0.01
6	0.64	0.6	23.18	<0.01	Line 22-24	2.37	4.08	<0.01
					Area 22-25-24-29	-2.66	-4.62	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.89	7.56	<0.01
7	0.75	0.5	39.46	<0.01	Line 22-24	-15.58	-9.23	<0.01
					Area 22-25-24-29	15.06	8.91	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.79	5.54	<0.01
8	0.91	0.31	71.53	<0.01	Line 22-24	-0.64	-4.39	<0.01
					Area 22-25-24-29	0.84	2.19	<0.01
					Line 27-32	0.6	1.82	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.27	1.83	<0.01
9	0	-	-	-	-	-	-	-
10	0.46	0.73	16.46	<0.01	Line 27-32	5.62	5.1	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	-4.63	-4.69	<0.01
11	0.71	0.54	47.91	<0.01	Line 22-24	-1.53	-6.99	<0.01
					Area 22-25-24-29	0.8	3.68	<0.01

Table 7.21 Proportion of the variability accounted for by the regression model in the response variable using combination 2 predictors (Facial Thermography)

Figure 7.29 presents a box plot summary of the results in the regression tables, the left side indicating the adjusted r^2 value while the right side showing the RMSE for each of the combinations of predictors. For a better visualisation, overlaid on each of the boxes are the individual values (as black circles).

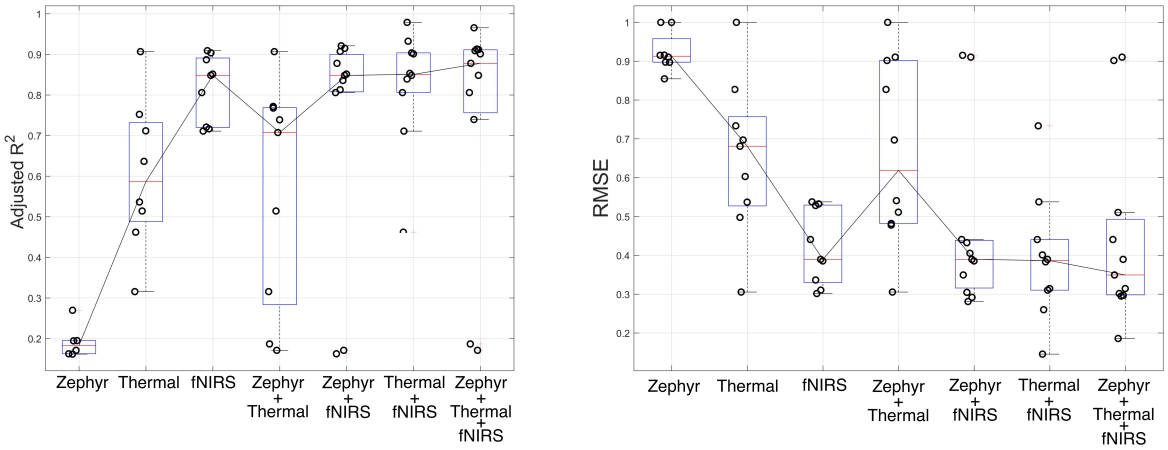

Fig. 7.29 Adjusted R^2 and RMSE for each of the four combinations of predictors

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0	-	-	-	-	-	-	-
2	0.72	0.53	33.03	<0.01	Optode9	-0.92	-3.15	<0.01
					Optode13	1.28	4.75	<0.01
					Optode15	-1.06	-4.65	<0.01
3	0.71	0.54	47.75	<0.01	Optode4	0.4	4.31	<0.01
					Optode6	-0.91	-9.73	<0.01
4	0.89	0.34	100.3	<0.01	Optode1	-0.28	-3.12	<0.01
					Optode13	1.11	8.94	<0.01
					Optode15	-1.53	-9.93	<0.01
5	0.81	0.44	53.59	<0.01	Optode1	2.04	10.13	<0.01
					Optode11	-0.5	-2.26	<0.01
					Optode12	-1.52	-9.57	<0.01
6	0.9	0.31	72.36	<0.01	Optode3	-1.41	-6.08	<0.01
					Optode5	1.72	5.79	<0.01
					Optode6	-0.65	-5.55	<0.01
					Optode12	0.44	2.43	<0.01
					Optode13	-1.04	-5.34	<0.01
7	0.72	0.53	25.53	<0.01	Optode2	0.66	2.42	<0.05
					Optode12	1.09	2.95	<0.05
					Optode15	-1.5	-4.53	<0.05
					Optode16	-0.94	-3.72	<0.05
8	0.91	0.3	77	<0.01	Optode1	-0.81	-3.93	<0.01
					Optode2	1.27	5.1	<0.01
					Optode4	-1.12	-4.48	<0.01
					Optode9	-0.57	-3.68	<0.01
					Optode16	0.31	2.55	<0.01
9	0.85	0.39	71.85	<0.01	Optode2	-0.83	-3.96	<0.01
					Optode4	1.31	4.36	<0.01
					Optode8	-1.38	-5.48	<0.01
10	0	-	-	-	-	-	-	-
11	0.85	0.39	55.46	<0.01	Optode9	0.47	5.68	<0.01
					Optode11	-1.14	-5.83	<0.01
					Optode14	1.05	4.55	<0.01
					Optode16	-0.82	-3.57	<0.01

Table 7.22 Proportion of the variability accounted for by the regression model in the response variable using combination 3 predictors (Prefrontal cortex blood oxygenation)

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0.17	0.91	8.84	<0.01	Mean RR	0.44	2.97	<0.01
2	0.51	0.7	21.11	<0.01	Line 22-24	0.48	3.72	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.38	2.92	<0.01
3	0	-	-	-	-	-	-	-
4	0.77	0.48	32.42	<0.01	Mean BR	-0.61	-7.53	<0.01
					Point 24	-2.15	-6.92	<0.01
					Line 22-24	2.38	7.76	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.24	2.65	<0.01
5	0.32	0.83	18.53	<0.01	Line 27-32	0.58	4.3	<0.01
6	0.71	0.54	23.99	<0.01	Mean BR	-0.38	-3.08	<0.01
					Line 22-24	2.11	4.01	<0.01
					Area 22-25-24-29	-2.55	-4.93	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.65	4.9	<0.01
7	0.77	0.48	43.82	<0.01	Line 22-24	-15.3	-9.45	<0.01
					Area 22-25-24-29	14.24	8.71	<0.01
					Line 27-32	1.28	6.03	<0.01
8	0.91	0.31	71.53	<0.01	Line 22-24	-0.64	-4.39	<0.01
					Area 22-25-24-29	0.84	2.19	<0.01
					Line 27-32	0.6	1.82	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.27	1.83	<0.01
9	0	-	-	-	-	-	-	-
10	0.19	0.9	9.27	<0.01	Mean RR	0.45	3.04	<0.01
11	0.74	0.51	36.84	<0.01	Mean BR	-0.18	-2.18	<0.05
					Line 22-24	-1.55	-7.45	<0.05
					Area 22-25-24-29	0.79	3.81	<0.05

Table 7.23 Proportion of the variability accounted for by the regression model in the response variable using combination 4 predictors (Heart Rate, Breathing rate, Facial Thermography)

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0.17	0.91	8.84	<0.01	Mean RR	0.44	2.97	<0.01
2	0.84	0.41	39.7	<0.01	Mean BR	-0.37	-5.15	<0.01
					Optode1	0.28	2.43	<0.01
					Optode9	-1.17	-4.48	<0.01
					Optode13	1.58	7.3	<0.01
					Optode15	-1.3	-7.22	<0.01
3	0.88	0.35	46.58	<0.01	Mean RR	-0.22	-2.28	<0.05
					Mean BR	-0.43	-6.01	<0.05
					Optode2	-0.3	-4.12	<0.05
					Optode4	1.06	6.93	<0.05
					Optode6	-0.82	-12.37	<0.05
					Optode14	-0.79	-4.1	<0.05
4	0.91	0.3	94	<0.01	Mean BR	-0.15	-2.56	<0.05
					Optode6	-0.42	-5.68	<0.05
					Optode15	-1.35	-9.41	<0.05
					Optode16	1.05	8.33	<0.05
5	0.81	0.44	53.59	<0.01	Optode1	2.04	10.13	<0.01
					Optode11	-0.5	-2.26	<0.01
					Optode12	-1.52	-9.57	<0.01
6	0.91	0.29	69.11	<0.01	Mean BR	-0.17	-2.31	<0.05
					Optode3	-0.94	-3.17	<0.05
					Optode5	1.48	4.98	<0.05
					Optode6	-0.68	-6.18	<0.05
					Optode12	0.35	2.01	<0.05
					Optode13	-1.1	-5.96	<0.05
7	0.81	0.43	28.54	<0.01	Mean RR	-0.39	-3.76	<0.01
					Optode2	1.68	5.18	<0.01
					Optode5	-0.8	-2.38	<0.01
					Optode12	0.97	2.55	<0.01
					Optode15	-1.15	-3.67	<0.01
					Optode16	-1.47	-5.6	<0.01
8	0.92	0.28	74.96	<0.01	Mean RR	0.13	2.46	<0.05
					Optode1	-0.82	-4.27	<0.05
					Optode2	1.11	4.56	<0.05
					Optode4	-0.91	-3.64	<0.05
					Optode9	-0.59	-4.11	<0.05
					Optode16	0.33	2.96	<0.05
9	0.85	0.39	71.85	<0.01	Optode2	-0.83	-3.96	<0.01
					Optode4	1.31	4.36	<0.01
					Optode8	-1.38	-5.48	<0.01
10	0.16	0.92	8.37	<0.01	Mean RR	0.43	2.89	<0.01
11	0.85	0.39	55.46	<0.01	Optode9	0.47	5.68	<0.01
					Optode11	-1.14	-5.83	<0.01
					Optode14	1.05	4.55	<0.01
					Optode16	-0.82	-3.57	<0.01

Table 7.24 Proportion of the variability accounted for by the regression model in the response variable using combination 5 predictors (Heart Rate, Breathing rate, Prefrontal cortex blood oxygenation)

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0	-	-	-	-	-	-	-
2	0.85	0.38	45.18	<0.01	Area 25-26-27-28-29-34-33-32-31-30	0.52	5.66	<0.01
					Optode1	-0.64	-4.34	<0.01
					Optode2	0.61	4.8	<0.01
					Optode11	0.55	3.19	<0.01
					Optode15	-0.55	-3.02	<0.01
3	0.71	0.54	47.75	<0.01	Optode4	0.4	4.31	<0.01
					Optode6	-0.91	-9.73	<0.01
4	0.98	0.15	160.99	<0.01	Line 22-24	-1.16	-10.53	<0.01
					Line 27-32	0.31	4.28	<0.01
					Optode1	-0.49	-3.72	<0.01
					Optode4	0.3	1.84	<0.01
					Optode6	-0.99	-7.1	<0.01
					Optode9	0.45	2.73	<0.01
					Optode11	-0.92	-5.07	<0.01
					Optode12	0.82	3.75	<0.01
					Optode13	1	3.6	<0.01
					Optode15	-0.94	-6.35	<0.01
					Optode16	-0.4	-1.84	<0.01
5	0.81	0.44	53.59	<0.01	Optode1	2.04	10.13	<0.01
					Optode11	-0.5	-2.26	<0.01
					Optode12	-1.52	-9.57	<0.01
6	0.9	0.31	72.36	<0.01	Optode3	-1.41	-6.08	<0.01
					Optode5	1.72	5.79	<0.01
					Optode6	-0.65	-5.55	<0.01
					Optode12	0.44	2.43	<0.01
					Optode13	-1.04	-5.34	<0.01
7	0.84	0.4	40.68	<0.01	Line 22-24	-9.73	-4.91	<0.01
					Area 22-25-24-29	9.08	4.73	<0.01
					Line 27-32	0.91	4.52	<0.01
					Optode13	0.76	2.55	<0.01
					Optode15	-1.1	-3.39	<0.01
8	0.93	0.26	101.12	<0.01	Line 22-24	-0.36	-2.99	<0.01
					Line 27-32	0.8	5.8	<0.01
					Optode2	0.42	2.48	<0.01
					Optode13	-0.9	-5.28	<0.01
9	0.85	0.39	71.85	<0.01	Optode2	-0.83	-3.96	<0.01
					Optode4	1.31	4.36	<0.01
					Optode8	-1.38	-5.48	<0.01
10	0.46	0.73	16.46	<0.01	Line 27-32	4.56	5.1	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	-4.19	-4.69	<0.01
11	0.9	0.31	70.37	<0.01	Line 22-24	-0.68	-7.4	<0.01
					Optode4	0.59	4.25	<0.01
					Optode10	0.43	2.1	<0.01
					Optode11	-0.72	-2.97	<0.01
					Optode14	-0.66	-2.89	<0.01

Table 7.25 Proportion of the variability accounted for by the regression model in the response variable using combination 6 predictors (Facial Thermography, Prefrontal cortex blood oxygenation)

Participant	Adjusted R^2	RMSE	F statistic	p-Value	Predictors	Beta	t-statistic	p-Value
1	0.17	0.91	8.84	<0.01	Mean RR	0.44	2.97	<0.01
2	0.91	0.29	67.5	<0.01	Mean BR	-0.35	-6.91	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.5	6.91	<0.01
					Optode2	0.29	4.49	<0.01
					Optode7	-0.36	-2.13	<0.01
					Optode13	1.17	5.86	<0.01
					Optode15	-1.24	-9.66	<0.01
3	0.88	0.35	46.58	<0.01	Mean RR	-0.22	-2.28	<0.05
					Mean BR	-0.43	-6.01	<0.05
					Optode2	-0.3	-4.12	<0.05
					Optode4	1.06	6.93	<0.05
					Optode6	-0.82	-12.37	<0.05
					Optode14	-0.79	-4.1	<0.05
4	0.97	0.19	152.97	<0.01	Mean BR	-0.23	-6.52	<0.01
					Line 22-24	-0.92	-8.44	<0.01
					Line 27-32	0.31	3.57	<0.01
					Optode1	-0.32	-4.29	<0.01
					Optode6	-0.72	-6.09	<0.01
					Optode14	0.89	11.74	<0.01
					Optode15	-0.85	-10.64	<0.01
5	0.81	0.44	53.59	<0.01	Optode1	2.04	10.13	<0.01
					Optode11	-0.5	-2.26	<0.01
					Optode12	-1.52	-9.57	<0.01
6	0.91	0.3	66.56	<0.01	Mean BR	-0.27	-4.51	<0.01
					Optode1	-0.31	-2.39	<0.01
					Optode5	1.02	4.36	<0.01
					Optode6	-0.86	-6.7	<0.01
					Optode13	-1.16	-5.62	<0.01
					Optode14	0.35	2.23	<0.01
7	0.74	0.51	36.97	<0.01	Mean RR	-0.34	-3.69	<0.01
					Line 27-32	0.53	5.84	<0.01
					Optode15	-0.86	-9.8	<0.01
8	0.91	0.3	73.49	<0.01	Mean RR	0.26	4.02	<0.01
					Area 25-26-27-28-29-34-33-32-31-30	0.31	2.48	<0.01
					Optode13	-0.82	-5.04	<0.01
					Optode15	0.29	2.32	<0.01
9	0.85	0.39	71.85	<0.01	Optode2	-0.83	-3.96	<0.01
					Optode4	1.31	4.36	<0.01
					Optode8	-1.38	-5.48	<0.01
10	0.19	0.9	9.27	<0.01	Mean RR	0.46	3.04	<0.01
11	0.9	0.31	70.37	<0.01	Line 22-24	-0.68	-7.4	<0.01
					Optode4	0.59	4.25	<0.01
					Optode10	0.43	2.1	<0.01
					Optode11	-0.72	-2.97	<0.01
					Optode14	-0.66	-2.89	<0.01

Table 7.26 Proportion of the variability accounted for by the regression model in the response variable using combination 7 predictors (Heart Rate, Breathing rate, Facial Thermography, Prefrontal cortex blood oxygenation)

7.4 Discussion

The results presented in this chapter are based on an extension of the study presented in Chapter 5. The main change in this study was the use of the fNIRS sensor which provides a higher face validity. The results are similar to the ones presented in Chapter 5, physiological measures, especially facial thermography and fNIRS proving to be good candidates for non-invasive real-time measurement of workload. One aspect to be noted is that models have to be appropriately trained on previously recorded data from the user population.

For most participants, fNIRS provided the best response to changes in workload, followed by facial thermography. Heart and breathing rate data did not prove to be very sensitive. Facial thermography is one of the least intrusive methods, and while it does not provide the sensitivity of fNIRS, it could benefit from it in future studies as a ground truth measure. Combining cardiac measures with fNIRS and facial thermography with fNIRS greatly improved the amount of variability explained in the performance response variable.

As in the case of the study presented in Chapter 5, the extension study also demonstrates the importance of identifying whether an individual is one who demonstrates a strong relationship between physiological measures and experienced workload measures before these methods can be applied uniformly. The data from participants 1, 3 and 9 support the case that not all participants show changes in physiological parameters when experiencing higher levels of workload. It is interesting to observe that even though participants 1 and 3 showed almost no changes in terms of facial thermography or cardiac measures, they did show significant changes in terms of prefrontal cortex blood oxygenation.

This research presents novel insights into the relative value of physiological and subjective techniques for assessment of workload and human performance. The main novelty lies in the fact that multiple continuous physiological measures were recorded and synchronized with task performance and subjective ratings. The hypotheses explored in this study were:

1. There will be a measurable difference in subjective workload between the two levels of task difficulty This hypothesis was rejected: The mental demand measured using NASA-TLX showed no significant difference between the two levels of difficulty (stage 2 as compared to stages 1 and 3).

2. The subjective ratings of workload will be associated with changes in physiological measures Hypothesis 2 was partially supported: The study explored which physiological measures showed a change in accordance to the change in workload as measured subjectively on the ISA scale. It was found that for some of the participants, the mean normalized ISA ratings showed a stronger correlation with some of their physiological measures than it did with the individual ISA rating. Table 7.27 summarizes the results

by displaying the number of participants that showed moderate to strong correlations with mean ISA normalized or individual ISA ratings for each of the physiological measures presented above.

Measure	No. of participants showing moderate to strong correlations	
	Mean ISA Normalized	Individual ISA
R-R Intervals	3/10	4/10
Breathing Rate	3/10	5/10
Point 24	4/11	3/11
Line 22-24	4/11	4/11
Area 22-25-24-29	4/11	4/11
Line 27-32	5/11	5/11
Area 25-26-27-28-29-34-33-32-31-30	5/11	5/11
Optode 6	7/7	6/7
Optode 15	7/7	6/7

Table 7.27 No. of participants showing moderate to strong correlations with the ISA rating

3. Multiple physiological measures can be used in combination to analyse workload Hypothesis 3 was tested by using a multiple linear regression on the data from each of the participants. Facial thermography data improve the predictive model based on just heart rate and breathing rate by 40% on average while blood oxygenation in the prefrontal cortex improves model by about 71%. Already the fNIRS based model preforms very well, but some improvements can be seen when combining fNIRS and heart and breathing rate data as well as thermal data and fNIRS. Although the average in the explained variability does not improve by much, the data shown in boxplot Figure 7.29 shows visibly improved distribution of results. As mean performance across the participants was strongly correlated with the mean ISA normalized, it is an indication that these physiological measures could also provide good prediction results for the level of subjectively experienced mental workload.

One of the limitations of the study was the small number of participants; for the limited number of participants (11), the only physiological measure that proved to work best at predicting mental workload or performance levels across all participants was blood oxygenation in the prefontal cortex. Although from a physiological point of view people responded differently when being subjected to the type of demand induced by the task, some of the other physiological measures, especially nose temperature proved to be consistent indicators of the level of performance (and implicitly the level of demand) for about half of the participants.

7.5 Chapter Summary

This chapter presented the results obtained in the third study performed during my research and representing an extension of the first study. An additional physiological measurement sensor, the fNIRS was used for measuring blood oxygenation in the prefrontal cortex. Although the fNIRS results do not represent my contribution to the thesis, they were shown as a comparison and as a measure that has higher face validity when measuring mental workload. The fNIRS measure also shows similar trends to other physiological measures indicating concurrent validity. The results demonstrate that physiological measures, especially face temperature and blood oxygenation in the prefrontal cortex, can be used for non-invasive real-time measurement of workload when combined with a facial landmark tracking algorithm, assuming models have been appropriately trained on previously recorded data from the user population.

This chapter has addressed research questions 1, 2 and 3:

1. How do human physiological responses change in response to variations in task demand and task performance? This was explored by recording multiple channels of physiological data while the participant was exposed to a task eliciting a predefined pattern of demand.

2. How are physiological responses associated with variations in subjective reports of mental workload? This was tested by comparing the response of multiple physiological data pieces with the subjective ratings of mental workload.

3. Can multiple combined physiological parameters explain more of the variability in mental workload or performance than individual parameters? This was explored by performing a multiple linear regression on the data from each of the participants, showing that using facial thermography data improves the predictive model.

The next chapter will present the conclusions of the thesis, a brief account of the main lessons learned and ideas for future work.

Chapter 8

Discussion and Conclusions

8.1 Contributions to research

The research behind this thesis was aimed at exploring objective physiological measures for estimating the level of workload a person is subjected to during their work. The project was mainly focused on safety critical environments, such as aeroplane/helicopter pilots but this type of measurements can be extended to other workplaces where the subject is seated.

The contributions of this research are:

- This thesis contributes to the measurement and assessment of workload using facial thermography in controlled laboratory conditions using a task designed specifically for inducing a controlled pattern of variation in demand.
- This thesis further contributes to the assessment of workload using multiple physiological measures and by presenting the relative contribution of each of them.
- The last contribution of this thesis is the use of facial thermography measures coupled with heart and breathing rate measures in a highly realistic helicopter simulator considered an ecologically valid environment, where active pilots performed the task. This study revealed how highly trained individuals react to variations in demand from a physiological point of view.

One of the main reasons for the chosen physiological measures was minimal intrusiveness. Some of them were more widely used in workload research, such as heart rate, breathing rate [30],[30],[32],[26],[33],[34],[6],[25],[36],[37] and pupil diameter [38], [39], [40], [30], [41],[42],[43]. Apart from these measures, facial thermography was the main measure explored in this research. As mentioned in the literature review, facial thermography was not

used extensively in workload research [5], [8], [46], [47], [48],[9]; the most likely reason for this is the high prices of the thermal cameras. As thermal cameras are beginning to be more accessible and as recent advances in image processing make real time tracking of facial landmarks easier to achieve, it is feasible for this technology to be used for real time physiological monitoring.

The following research questions were addressed during this research:

1. How do human physiological responses change in response to variations in task demand and task performance?

   This research question has been explored throughout the three studies performed during this research. The physiological reaction in terms of heart rate, breathing rate, facial thermography and pupil diameter has been analysed. Most findings confirm previous results obtained by other researchers.

2. How are physiological responses associated with variations in subjective reports of mental workload?

   This research was focused on tasks that have a predominant mental component, hence the emphasis on mental workload. Mental workload subjective measures (ISA and NASA-TLX) have been used and suggested a strong association with the variation in physiological parameters. While cardiac and breathing measures did not show strong correlations with either subjective or performance measures, pupil diameter and nose temperature proved to be better predictors of workload and performance.

3. Can multiple combined physiological parameters explain more of the variability in mental workload or performance than individual parameters?

   This hypothesis was tested for the laboratory studies, presented in Chapters 5 and 7, showing that combining multiple physiological measures allows a multiple linear regression model to explain more of the variability in the performance measure. As the performance measure was strongly correlated with the subjective workload measure, this is an indication that multiple physiological measures can also provide a good prediction of the experienced level of mental workload.

4. How do highly trained individuals respond to variations of task demand in an ecologically valid aircraft simulator?

   This hypothesis was tested during the helicopter flight simulator study presented in Chapter 6, showing that highly trained individuals present changes in physiological measures when experiencing high workload situations.

8.2 Laboratory studies

Two laboratory studies have been performed during this research, involving the collection of physiological data from participants playing a custom built computer game. These studies have identified which physiological measures provide better responses to the change in workload as well as which areas of the face respond by changing temperature. This research presents novel insights into the relative value of physiological and subjective techniques for assessment of workload and human performance. The main novelty lies in the fact that multiple continuous physiological measures were recorded and synchronized with task performance and subjective ratings. These studies have demonstrated that the use of physiological monitoring is a feasible proposition for work environments where the user is seated, such as aircraft cockpits or air traffic control rooms.

8.3 Flight simulator study

The flight simulator study performed during this research was novel as it allowed for the collection of physiological data from highly trained helicopter pilots while performing flight scenarios of various levels of difficulty in a very realistic simulator. This study provided insights into the challenges of deploying such measurements in a real work environment such as a helicopter cockpit as well as important data on the pilot's physiological reaction to changes in demand.

8.4 Lessons learned

The research presented in this thesis involved preparing studies on human participants and processing of large sets of data collected from multiple sensors. In this sub-section I will share a few of the challenges that I have faced in the hope they will be useful for researchers undertaking similar studies in the future.

The task that will be performed by the participants is one of the most important parts of the study. This has to be carefully chosen or designed in order for the induced effects to be relevant and in line with the aims of the study. The task should ideally be tested beforehand on a number of participants. In the case of my first laboratory study, using the coloured balls task, I have initially tested it with a few friends but not in the laboratory setting and only on one participant in the final environment before starting the study; everything seemed to work well. It was only after starting the study, that I noticed the first two participants were at times choosing a strategy of abandoning the balls that were too low on the screen and focusing on

the top ones that they could more easily reach. The problem here was that in this manner not all participants would have been exposed to the same level of demand depending on how many balls they chose to ignore. It was at this moment that I decided to change the task and introduced the yellow horizontal line which would be dragged down by the balls; this gave the participants a reason to destroy the lowest balls on the screen as for them not to reach the yellow line and drag it down. One reason I did not notice this before was that my first test participant was very good at the task and did not need to abandon the lower targets on the screen. In hindsight, I should have tested the task on more participants before running the study. Choosing the tasks for the helicopter flight simulator was different. Due to the high price and low availability of pilots, I had less opportunity of testing it beforehand. I overcame this challenge by designing the task together with the flight instructor, who had great experience with what the pilots would consider demanding.

The data collection strategy will probably depend on the task, sensors and environment. In dealing with physiological signals which will results in multiple time-series of data, synchronizing all devices is very important. In my case, for the laboratory studies, before collecting the data, I would synchronize the clocks on all computers/devices, this makes the data processing much easier during the later stages. One thing I would improve if I had to do it again would be building a user friendly software interface for importing multiple sets of data, cropping the time intervals that are of interest, filtering the data and generating the features. If someone is planning to conduct multiple studies of this type and have more participants, I think it is worth using more time at the beginning to streamline the data collection and analysis as it will save more time in later stages.

Collecting facial thermography data presented challenges from the beginning. The first thermal camera that I used was a FLIR T-300 which did not allow for an easy way to extract the radiometric data. The second thermal camera was the FLIR SC7000 described in Chapter 4. This is a professional thermal camera that is able to collect high resolution images with high thermal sensitivity. The downside of it was that the generated files were very large (about 50 GB for each participant); the next challenge was reading the files into Matlab for processing the images and efficiently converting the digital level generated by the camera to temperatures using the calibration curve. While in the end all these challenges were overcome, the time it takes should not be underestimated.

The image processing techniques applied on the thermal images in order to track landmarks and to extract temperatures presented another challenge. I have initially tried an image registration approach, of fusing a thermal image with a visual image. The aim was to perform the landmark tracking in the visual image and extract the temperatures from the resulting coordinates in the thermal images. Even with using face markers, this approach turned out

not to be successful and in the end I settled for performing the landmark tracking directly on the thermal images. This approach required generating a training set of positive landmarks on samples from each participant which also proved to be time consuming. I now believe that the image registration approach would have been more robust and if I had to run multiple studies using thermal imaging, I would put more effort into overcoming that challenge.

8.5 Future work

This research has shown that it is feasible to use physiological data to infer the state of workload perceived by a person performing a task. The focus was on safety critical environments such as pilots flying a helicopter. One weakness of the research was the reduced number of participants; further research is needed to confirm the physiological reactions on a larger number of participants. This research has also shown that not all participants show an expected physiological reaction in terms of cardiac and facial thermography measures while they do show changes in the blood oxygenation of the prefrontal cortex.

While this thesis uses a multiple linear regression approach to showing that it is feasible to estimate the levels of demand based on physiological data, in some cases the relationships between demand and physiological parameters are more complex than that. A more accurate approach would benefit from machine learning algorithms. As an example, a support vector machine classifier was trained using a Gaussian kernel for classifying the subjective levels of mental workload experienced by the participants based on the data from the first study, described in Chapter 5. For this application, the data were split into 30 seconds intervals with 90% overlap. Table 8.1 below shows the results of the classifier. The mean accuracy on a new participant column contains the results of the classifier when it was trained and validated using 5 fold cross-validation on the data from 9 out of the 10 participants and then tested on the left out data of one participant. This process was repeated for each of the participants and the values in the table represent the mean accuracy of these tests. The "Accuracy on 40% holdout[1] column" shows the results of the classifier when trained on a 60% partition of the data containing roughly equal class proportions from each of the participants and validated using 5 fold cross-validation. The accuracy of the classifier was tested on the remaining 40% of the data. This second case implies that in order for the algorithm to obtain these high accuracies it had to be be trained on data from each of the participants.

[1]The holdout method partitions the data into two mutually exclusive subsets called training and test set, or holdout set. The data in the holdout set are not used for training purposes, they are just used to test the accuracy of the classifier [82]

Predictors	Mean accuracy on new participant	Accuracy on 40% holdout
Heart and Breathing measures	30.89%	70.53%
Heart, Breathing and Pupil diameter measures	35.08%	80.2%
Thermal Measures	37.59%	88.25%
Heart, Breathing, Pupil diameter and Thermal measures	37.59%	90.82%

Table 8.1 Classifier Accuracy

There are of course shortcomings in these results as well, the main one being the small dataset. Further research on larger datasets and using different machine learning algorithms should be conducted to confirm these results and to also test if a classifier of this type, once trained, would maintain accuracy in time.

I believe that one other possible area of research would be in the automotive domain for monitoring driver state and situation awareness. Using a near-infra red camera (for night vision capabilities and more robust facial landmark tracking) coupled with a thermal camera and an eye tracker could provide the following data: facial expression, blink frequency, facial temperature, breathing rate (extracted from the thermal images as described in the next paragraph) and area of gaze estimation. One interesting research question would be if based on these pieces of data, the system could estimate how ready the driver is to take over control from the car.

Having mentioned the extraction of breathing rate from thermal images, exploring the data collected during my research I noticed that the nostril areas could provide breathing rate information. Figure 8.1 shows the an example of temperature variation in area 24-25-26-27 (marked with blue on the participant's face) for participant 1 during the first stage of the task; the bottom part of the figure shows a zoomed in area of the top part of the figure. Overall, a temperature drop can be observed as the task was becoming more difficult; apart from this temperature drop it can be observed that the temperature in the specified area varies as the participant breathes in and out. The observed variation ranges from 0.3°C to 0.55°C with an average of 0.4°C.

A continuous wavelet transform was applied on the temperature signal in order to extract both time and frequency estimates of the breathing rate. Figure 8.2 shows the breathing rate as recorded by the Zephyr sensor in red, overlaid over the magnitude of various frequencies at a certain time. It can be observed that the patterns are similar. This was just an example and further work should be done for the estimate to be improved.

The benefit of this method would be that breathing rate could be extracted in a non-intrusive manner by using a thermal camera pointed at the user.

I believe that another important area of study for future research on physiological monitoring as a measure for workload has to do with the physiological response over larger time

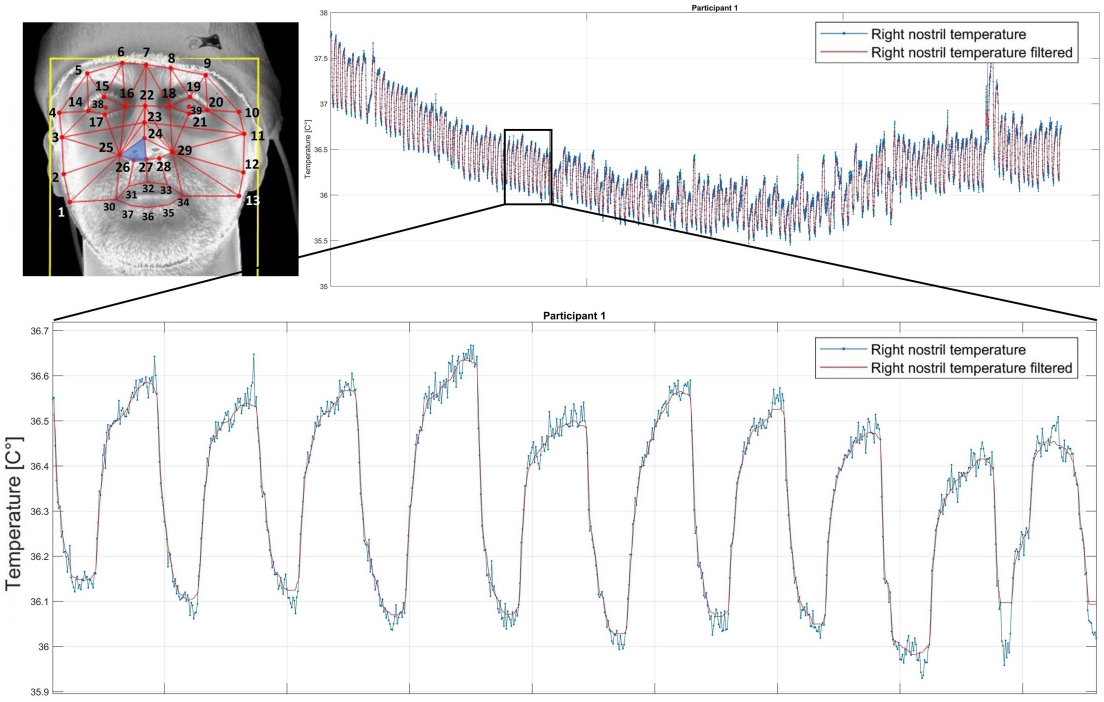

Fig. 8.1 Right nostril temperature variation indicating breathing rate for Participant 1

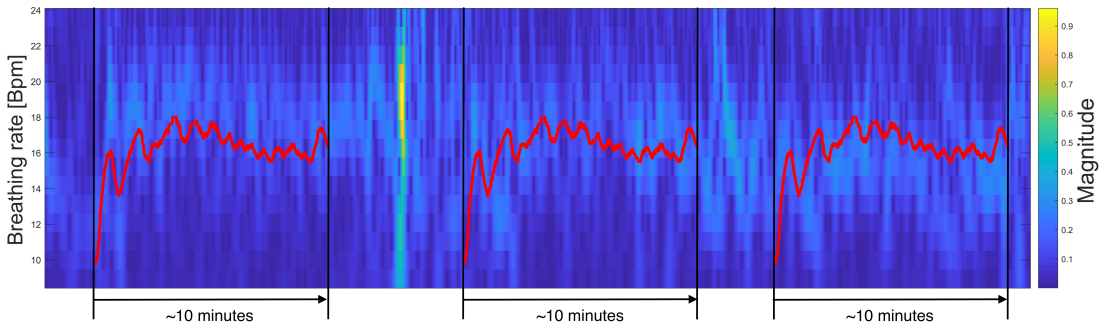

Fig. 8.2 Continuous wavelet transform visual comparison to Zephyr recorded breathing rate

periods. Will the measures be influenced by other phenomena such as fatigue? A possible observation on this came as I was trying to visualise the data from the first study; a projection of all physiological features to a lower dimensional space using PCA (Principal Component Analysis) revealed the result shown in Fig.8.3 for Participant 5. The colours of the markers (green, blue and red) indicate the performance achieved by the participant while the shape of the markers indicates the stage of the task. It can be seen that the data clusters into three groups, symbolising the stage while at the same time showing a curvature towards the low performance instances (those are the periods characterized by higher levels of workload). While this is just a qualitative visualisation, another interesting aspect to be observed is that the clusters seem to be ordered in the direction of time passage. Could this be an indication of fatigue? Similar trends but not as clear have been observed for a few other participants.

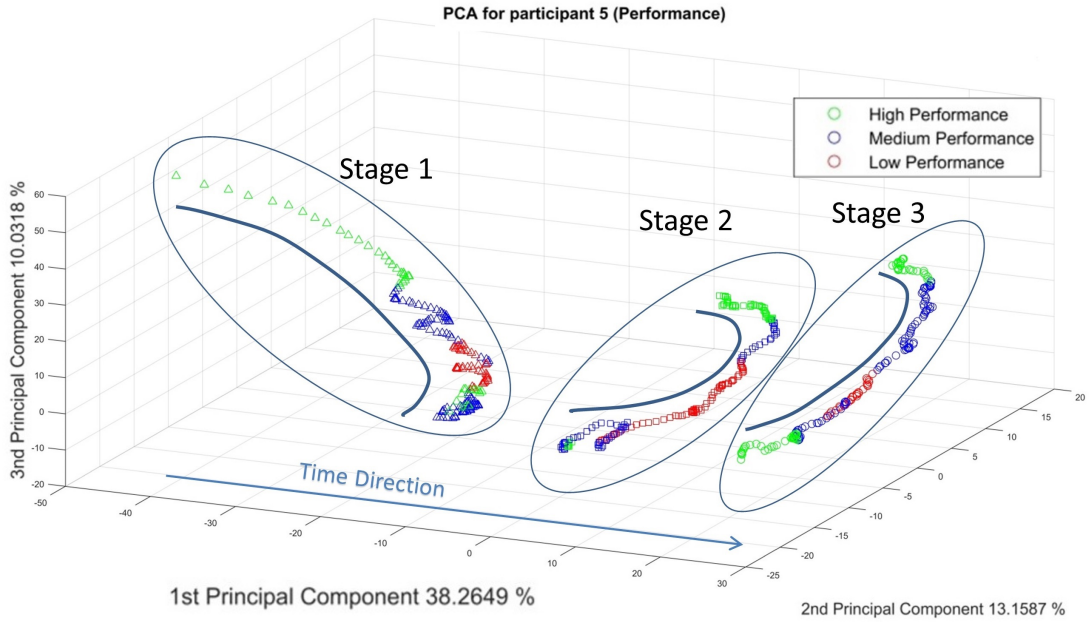

Fig. 8.3 PCA example for Participant 5 - Study 1 explaining about 62% of the variability in the initial data

The use of physiological measures appears to be the closest to the ideal of real time, non-invasive workload measurement technique, being able to reflect the way the task is perceived without interfering with the task. The research presented in this thesis indicates to a wide range of possible uses for this approach; nevertheless, more research on larger data sets including machine learning approaches will be necessary.

8.6 Summary of results

As a summary of results, I will attempt to compare the findings presented in this thesis with the results obtained by other researchers. I used the verb "attempt" as I believe this to be a challenging task; the main reason for this is that the differences between the study designs of the researchers are difficult to account for: the tasks to be solved are different and difficult to compare, the sensors used, as well as the data processing approaches are different.

Overall, in the laboratory studies, the R-R inter-beat intervals did not show a consistent, strongly correlated response with the changes in demand, agreeing with [30], [31] and [32]. The helicopter flight simulator study was slightly different, heart rate means during the difficult scenarios were significantly different from the heart rate means recorded during the easy and medium scenarios, agreeing with [26], [33], [34] and [36]. It is interesting to note that three of the four previously mentioned studies ([26], [33] and [34]) were performed in either a flight simulator or a real plane. As mentioned before this could also be induced by the physical type of work taking place during flying.

Pupil diameter was recorded only during the first laboratory study presented in Chapter 4 and the results are in agreement with most of the literature findings [38], [39], [40], [30], [41],[42], except for [43] which found a decrease in pupil diameter with the increase in demand. While this measure showed a strong correlation to the subjective ratings of workload, it is a more difficult measure to record outside well controlled laboratory conditions.

Facial thermography, the main measure explored in this research, revealed mostly similar results with regards to the nose area temperature variation with the increase in demand [5], [8], [46], [47], [48]. In this research, temperatures in various areas of the face were not reported as relative to the temperature in another area of the face. Forehead temperature has been reported as constant by [5], [44] and used as a reference for the nose temperature by [8], [46].

Most studies to date have examined how individual physiological measures respond to changes in mental demand while a few have looked at multiple physiological measures (e.g.: [30],[8]). The studies presented in Chapters 5 and 7 explore the response of multiple physiological parameters to workload by using a task that varies demand in a gradual and well controlled manner across all participants. Besides the carefully controlled variation of demand, one novelty lies in the fact that multiple areas, covering most of the face were examined for changes in temperature as a response to the variation of demand; another novel aspect is related to quantifying the added value of each of the measures when estimating the level of demand. The study presented in Chapter 7 is one of the few studies exploring the facial thermography measure in a high fidelity flight simulator operated by active helicopter pilots. The data from the helicopter flight simulator study revealed that while in most

situations forehead temperature does not change by much, there are certain changes that appear especially in high demand situations for some of the participants; the changes involved temperature decreases and have been observed in 5 out of 10 participants. This should be further examined, especially if forehead temperature is to be considered a reference for nose temperature as in [8]. Another novelty of this research was the continuous tracking of facial landmarks in thermal videos without the need of facial markers. This allowed for less intrusive measurements and the posibility of using this measure in the helicopter flight simulator. By using this techniques, it was possible to continuously extract temperature data from various areas of the face and one observation that was made on the data from the flight simulator study revealed that while the nose area shows the strongest response to changes in demand (temperature decrease with the increase in demand), in some situations the temperature on the cheeks also decreases in a pattern similar to the nose but at a smaller scale. This should also be further analysed as it might be related to the hypothesis that the decrease in nose temperature is caused by vasoconstriction. Another important observation made possible by the accurate facial landmark tracking technique revealed that, in some cases, asymmetries in temperature patterns were observed between the right and left sides, especially in the cheek areas and in the area between the nose and the lips. For future studies, this should be observed carefuly and it should be checked if they maintain over longer time spans for the same individual.

Facial thermography, especially nose temperature, has been reported to be sensitive to the presence of loud noises [44], to either a positive (laughter) or negative (watching other people suffer) empathic response [49] and to the participant lying [50]. Further research should also focus on examining the selectivity of the facial thermography measure.

As a conclusion to the use of facial thermography as a tool for assessing workload, I will make a few comments with regards to the criteria for workload measures presented in section 2.3:

1. **Validity:** is the method measuring workload?

    - *Face Validity: would the measure be accepted as measuring workload by all involved stakeholders?* The evidence does show that the method responds to changes in workload as measured by subjective methods but, further research should be undertaken as to establish it as an accepted measure of workload.
    - *Construct Validity: does the method measure all aspects associated with workload?* The method does not measure all aspects associated with workload.
    - *Concurrent/Convergent Validity: do multiple workload measurement methods show the same trends?* Multiple workload measurements methods show the

same trends, including subjective methods, heart rate measures (especially in the helicopter simulator case), pupil diameter and most important of all the fNIRS measure.

2. **Reliability:** *if applied multiple times, are the results provided by the method consistent?* There is not enough evidence based on the studies performed that if applied multiple times the results would be consistent; it is also not known how, for example the temperature in the nose area would react over long periods of time while the participant is exposed to varying levels of demand, as they would be in a workplace.

3. **Generalisability:** *can the method be applied in multiple domains?* The method is especially suitable for workplaces where the person is seated so that the thermal camera can have a view of their face and also so that there are little effects from physical workload. Another challenge would be performing the measurement in an environment where the temperature would be either too low, too high or the face of the person would be under direct sunlight. It is not known whether the sensitivity would be affected in those cases.

4. **Sensitivity:** *how small are the changes in task demand or performance that can be detected by the method?* Based on the laboratory studies where the demand was varied in small increments, it appears that the method of facial thermography is sensitive to small changes in demands, but not for all participants.

5. **Interference:** *does the method technique interfere with the primary task?* The method does not interfere in any way with the primary task.

6. **Diagnosticity:** *can the method distinguish between demands incurred by different resources (as described above in the Multiple Resource Model)?* There is no evidence, based on the studies performed, to support the fact that the method can distinguish between demands incurred by different resources.

7. **Selectivity:** *can the method distinguish mental workload from other factors such as emotional stress or physical workload?* This aspect has not been tested in this research but there are reasons to believe, based on other research, that nose temperature decreases in the presence of loud noises [44], as a result of either a positive (laughter) or negative (watching other people suffer) empathic response [49] and when the participants are lying [50]. These aspects should be tested in future research; multiple physiological measures might offer a solution to this challenge.

8. **Granularity/bandwidth:** *what is the time resolution of the method?* Depending on the thermal camera, the data collection can be performed up to at least 100 frames per second. In terms of the physiological response, meaningful temperature changes can be observed over time periods of 45 seconds as tested in the laboratory study and this could possibly be lower.

9. **Feasibility of use:** *how easy would it be to use the method?* If proven to reliably work by more studies, the method would be easy to use and non-invasive provided more work is done on the facial landmark tracking algorithm and machine learning classifier.

10. **Acceptability and Ethics:** *would the participant and those around him accept the measurement applied?* I believe it is an easy to accept measure as it has no secondary effects on the human body; it is measuring the thermal radiation emitted by the human body, from this point of view being completely passive. From a privacy point of view it could raise concerns of collecting physiological data that could in the near future be used to diagnose various conditions.

11. **Resources:** *how practical is the method in terms of required financial and time resources?* If proven to work in terms of reliably classifying multiple levels of workload, the method would be very easy to use, requiring very little time, this however would require more research and development. In terms of costs, the thermal cameras are not very easily affordable, with starting prices in the order of a few thousand pounds, to which computational resources should be added. I am hopeful that this technology could be implemented using the new FLIR One modules with price ranges between £200 and £400 and potentially using the computational power of a mobile phone.

Overall, the results obtained reveal similar findings to the literature. The main contributions come from the assessment of multiple physiological measures for workload and their relative contribution, as well as by more accurately continuously tracking facial landmarks during a specially designed task for the laboratory study, which allowed for a more accurate control of the variation of demand. The final contribution was the assessment of the physiological reaction of highly trained individuals in a helicopter flight simulator.

The demonstration of feasibility of physiological measures as a method for assessing workload presented within this thesis allows the identification of guidance for how this approach can be used in the future, and requirements for further research. The methods presented in this thesis, with current technological capabilities, are better suited for workplaces where the subject is seated as the methods can cope with a limited amount of head movement. Continuous real time non-invasive workload measurement techniques is now a

realistic proposition that will allow for improved design of human-machine systems, operating procedures and operations scheduling in ways that will bring us closer to the goal of optimizing human well-being and overall system performance.

References

[1] Airbus. Flying by numbers, August 2015.

[2] Raja Parasuraman and Ranjana Mehta. Neuroergonomic methods for the evaluation of physical and cognitive work. In John R Wilson and Sarah Sharples, editors, *Evaluation of human work*, chapter 22. CRC press, 2015.

[3] Sarah Sharples and Ted Megaw. Definition and measurement of human workload. In John R Wilson and Sarah Sharples, editors, *Evaluation of human work*, chapter 18. CRC press, 2015.

[4] Mark S Young, Karel A Brookhuis, Christopher D Wickens, and Peter A Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.

[5] Calvin KL Or and Vincent G Duffy. Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, 7(2):83–94, 2007.

[6] Paul Lehrer, Maria Karavidas, Shou-En Lu, Evgeny Vaschillo, Bronya Vaschillo, and Andrew Cheng. Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study. *International Journal of Psychophysiology*, 76(2):80–87, 2010.

[7] Ulf Ahlstrom and Ferne J Friedman-Berg. Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, 36(7):623–636, 2006.

[8] Koji Murai, Yuji Hayashi, Tadatsugi Okazaki, Laurie C Stone, and Nobuo Mitomo. Evaluation of ship navigator's mental workload using nasal temperature and heart rate variability. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 1528–1533. IEEE, 2008.

[9] John Stemberger, Robert S Allison, and Thomas Schnell. Thermal imaging as a way to classify cognitive workload. In *Computer and Robot Vision (CRV), 2010 Canadian Conference on*, pages 231–238. IEEE, 2010.

[10] Paul M Linton, Brian D Plamondon, AO Dick, Alvah C Bittner Jr, and Richard E Christ. Operator workload for military system acquisition. In *Applications of human performance models to system design*, pages 21–45. Springer, 1989.

[11] Mark S Young and Neville A Stanton. It's all relative: defining mental workload in the light of annett's paper. *Ergonomics*, 45(14):1018–1020, 2002.

[12] Christopher D Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[13] Neville Moray. *Mental workload: Its theory and measurement*, volume 8. Springer Science & Business Media, 2013.

[14] Laura Pickup, John R Wilson, Sarah Sharpies, Beverley Norris, Theresa Clarke, and Mark S Young. Fundamental examination of mental workload in the rail industry. *Theoretical issues in ergonomics science*, 6(6):463–482, 2005.

[15] Christopher D Wickens, Justin G Hollands, Simon Banbury, and Raja Parasuraman. *Engineering psychology & human performance*. Psychology Press, 2015.

[16] Chris Eccleston and Geert Crombez. Pain demands attention: A cognitive–affective model of the interruptive function of pain. *Psychological bulletin*, 125(3):356, 1999.

[17] Malcolm H Johnson, Guy Breakwell, Wanda Douglas, and Steven Humphries. The effects of imagery and sensory detection distractors on different measures of pain: how does distraction work? *British Journal of Clinical Psychology*, 37(2):141–154, 1998.

[18] David Navon. Resources—a theoretical soup stone? *Psychological review*, 91(2):216, 1984.

[19] Neville A Stanton. *Advances in human aspects of road and rail transportation*. CRC Press, 2012.

[20] T Edwards, S Sharples, JR Wilson, and B Kirwan. The need for a multi-factorial approach to safe human performance in air traffic control. In *Proceedings of the 4th AHFE International Conference*, 2012.

[21] Robert D O'Donnell and F Thomas Eggemeier. *Workload assessment methodology*. John Wiley & Sons, 1986.

[22] SD Brennan. An experimental report on rating scale descriptor sets for the instantaneous self assessment (isa) recorder. *Portsmouth: DRA Maritime Command and Control Division. DRA Technical Memorandum (CAD5)*, 92017, 1992.

[23] SG Hart. California mf, staveland le. *Development of NASA-TLX (Task Load Index): Results of emprical and theoretical research. Adv Psychol*, 52:139–83, 1988.

[24] Stefan Schneegass, Bastian Pfleging, Nora Broy, Frederik Heinrich, and Albrecht Schmidt. A data set of real world driving to assess driver workload. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, pages 150–157. ACM, 2013.

[25] Karel A Brookhuis and Dick de Waard. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, 42(3):898–903, 2010.

[26] Malcom A. Bonner and Glenn F. Wilson. Heart rate measures of flight test and evaluation. *The international journal of aviation psychology*, 12(1):49–61, 2002.

[27] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. Classifying driver workload using physiological and driving performance data: Two field studies. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 4057–4066. ACM, 2014.

[28] Thibault Gateau, Gautier Durantin, Francois Lancelot, Sebastien Scannella, and Frederic Dehais. Real-time state estimation in a flight simulator using fnirs. *PloS one*, 10(3):e0121279, 2015.

[29] Mark A Staal. Stress, cognition, and human performance: A literature review and conceptual framework. *NASA*, 2004.

[30] John G Casali and Walter W Wierwille. A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Human Factors*, 25(6):623–641, 1983.

[31] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377, 1996.

[32] Willem B Verwey and Hans A Veltman. Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of experimental psychology: Applied*, 2(3):270, 1996.

[33] Erland A.I. Svensson and Glenn F. Wilson. Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *The international journal of aviation psychology*, 12(1):49–61, 2002.

[34] Glenn F Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1):3–18, 2002.

[35] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. Psychophysiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 301–310. ACM, 2010.

[36] Arjan Stuiver, Karel A Brookhuis, Dick de Waard, and Ben Mulder. Short-term cardiovascular measures for driver support: increasing sensitivity for detecting changes in mental workload. *International Journal of Psychophysiology*, 92(1):35–41, 2014.

[37] DA Spyker, SP Stackhouse, AS Khalafalla, and RC McLane. Development of techniques for measuring pilot workload. *NASA*, 1971.

[38] Eckhard H Hess and James M Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[39] Daniel Kahneman and Jackson Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966.

[40] Marjorie Jane Krebs, James W Wingert, and Thomas Cunningham. Exploration of an oculometer-based model of pilot workload. *NASA*, 1977.

[41] Miguel Ángel Recarte, Elisa Pérez, Ángela Conchillo, and Luis Miguel Nunes. Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish journal of psychology*, 11(2):374–385, 2008.

[42] Tjerk de Greef, Harmen Lafeber, Herre van Oostendorp, and Jasper Lindenberg. Eye movement as indicators of mental workload to trigger adaptive automation. *Foundations of augmented cognition. Neuroergonomics and operational neuroscience*, pages 219–228, 2009.

[43] Leandro Luigi Di Stasi, Adoración Antolí, and José Juan Cañas. Evaluating mental workload while interacting with computer-generated artificial environments. *Entertainment Computing*, 4(1):63–69, 2013.

[44] A Naemura, K Tsuda, and N Suzuki. Effects of loud noise on nasal skin temperature. *Shinrigaku kenkyu: The Japanese journal of psychology*, 64(1):51–54, 1993.

[45] Hirokazu Genno, Keiko Ishikawa, Osamu Kanbara, Makoto Kikumoto, Yoshihisa Fujiwara, Ryuuzi Suzuki, and Masato Osumi. Using facial skin temperature to objectively evaluate sensations. *International Journal of Industrial Ergonomics*, 19(2):161–171, 1997.

[46] Koji Murai, Kenichi Kitamura, and Yuji Hayashi. Study of a port coordinator's mental workload based on facial temperature. *Procedia Computer Science*, 60:1668–1675, 2015.

[47] Jihun Kang and Kari Babski-Reeves. Detecting mental workload fluctuation during learning of a novel task using thermography. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 1527–1531. SAGE Publications Sage CA: Los Angeles, CA, 2008.

[48] Michelle L Reyes, John D Lee, Yulan Liang, Joshua D Hoffman, and Ritchie W Huang. Capturing driver response to in-vehicle human-machine interface technologies using facial thermography. In *Proceedings of the International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, volume 5, pages 536–542, 2009.

[49] E Salazar-López, E Domínguez, V Juárez Ramos, J de la Fuente, A Meins, O Iborra, G Gálvez, MA Rodríguez-Artacho, and E Gómez-Milán. The mental and subjective skin: Emotion, empathy, feelings and thermography. *Consciousness and cognition*, 34:149–162, 2015.

[50] A Moliné, G Gálvez-García, J Fernández-Gómez, J De la Fuente, O Iborra, F Tornay, JL Mata Martín, M Puertollano, and E Gómez Milán. The pinocchio effect and the cold stress test: Lies and thermography. *Psychophysiology*, 2017.

[51] Frans F Jobsis. Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267, 1977.

[52] Kurtulus Izzetoglu, Scott Bunce, Banu Onaral, Kambiz Pourrezaei, and Britton Chance. Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of human-computer interaction*, 17(2):211–227, 2004.

[53] Gautier Durantin, J-F Gagnon, Sébastien Tremblay, and Frédéric Dehais. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural brain research*, 259:16–23, 2014.

[54] Scott Bunce, Kurtulus Izzetoglu, Hasan Ayaz, Patricia Shewokis, Meltem Izzetoglu, Kambiz Pourrezaei, and Banu Onaral. Implementation of fnirs for monitoring levels of expertise and mental workload. *Foundations of augmented cognition. Directing the future of adaptive systems*, pages 13–22, 2011.

[55] Hasan Ayaz, Patricia A Shewokis, Scott Bunce, Kurtulus Izzetoglu, Ben Willems, and Banu Onaral. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47, 2012.

[56] Horia A Maior, Matthew Pike, Sarah Sharples, and Max L Wilson. Examining the reliability of using fnirs in realistic hci settings for spatial and verbal tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3039–3042. ACM, 2015.

[57] Glenn F. Wilson. Similarities and differences in psychophysiological reactions between simulated and real air-to-ground missions. *The international journal of aviation psychology*, 12(1):49–61, 2002.

[58] Neurosky eeg sensors. http://neurosky.com/biosensors/eeg-sensor/biosensors/.

[59] Emotiv eeg research and education. https://www.emotiv.com/neuroscience-research-education-solutions/.

[60] Haskins eeg lab. https://haskinslabs.org/research/research-infratructure/eeg-lab.

[61] Zephyr. Bioharness 3.0 user manual, September 2012. [online] https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf.

[62] P. Tarvainen Mika, Lipponen Jukka, Niskanen Juha-Pekka, and O. Ranta-aho Perttu. Kubios hrv. http://www.kubios.com/downloads/Kubios_HRV_Users_Guide.pdf.

[63] Heart Rate Variability. Standards of measurement, physiological interpretation, and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*, 93(5):1043–1065, 1996.

[64] Sensomotoric Instruments. iview x system manual, March 2011.

[65] Vérane Faure, Regis Lobjois, and Nicolas Benguigui. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation research part F: traffic psychology and behaviour*, 40:78–90, 2016.

[66] FLIR. The ultimate infrared handbook for rd professionals, 2012. [online] http://www.flir.com/science/display/?id=69528.

[67] FLIR. Flir sc7000 series brochure, March 2011. [online] http://www.flirmedia.com/MMC/THG/Brochures/RND_017/RND_017_US.pdf.

[68] FLIR. User's manual flir ax5 series, March 2011. [online] http://www.flir.com.

[69] Biopac. Biopac website. https://www.biopac.com/knowledge-base/fnir-faq/.

[70] George E Billman. Heart rate variability–a historical perspective. *Frontiers in physiology*, 2, 2011.

[71] Gaetano Valenza, Antonio Lanata, and Enzo Pasquale Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE transactions on affective computing*, 3(2):237–249, 2012.

[72] Constantino A Garcıa, Abraham Otero, Xosé Vila, Arturo Méndez, Leandro Rodrıguez-Linares, and Marıa José Lado. Getting started with rhrv, 2013.

[73] Gary G Berntson, J Thomas Bigger, Dwain L Eckberg, Paul Grossman, Peter G Kaufmann, Marek Malik, Haikady N Nagaraja, Stephen W Porges, J Philip Saul, Peter H Stone, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648, 1997.

[74] J Piskorski and P Guzik. Geometry of the poincaré plot of rr intervals and its asymmetry in healthy adults. *Physiological measurement*, 28(3):287, 2007.

[75] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[76] Carlos Adrián Vargas Aguilera. Extrema2. https://uk.mathworks.com/matlabcentral/fileexchange/12275-extrema-m--extrema2-m?focused=6267318&tab=function, 2007. Version: 2007-04-09.

[77] S Sharples, T Edwards, and N Balfe. Inferring cognitive state from observed interaction. In *Proceedings of the 4th AHFE International Conference*, 2012.

[78] James D Evans. *Straightforward statistics for the behavioral sciences*. Brooks/Cole, 1996.

[79] Airbus Helicopters. Airbus helicopters. http://www.airbushelicopters.co.uk/website/en/press/Eurocopter%E2%80%99s%20EC225%20helicopter%20Full%20Flight%20Simulator_20.html, 2011.

[80] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4, 2013.

[81] Michael J Kane and Randall W Engle. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review*, 9(4):637–671, 2002.

[82] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.