Formalising Human Mental Workload as a Defeasible Computational Concept

Luca Longo

A Dissertation submitted to the University of Dublin, Trinity College

in fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

October 2014

Declaration

I, the undersigned, declare that this work has not previously been submitted to this or any other University, and that unless otherwise stated, it is entirely my own work.

Luca Longo

Dated: October , 2014

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this Dissertation upon request.

Luca Longo

Dated: October , 2014

Abstract

Human mental workload has gained importance, in the last few decades, as a fundamental design concept in human-computer interaction. It can be intuitively defined as the amount of mental work necessary for a person to complete a task over a given period of time. For people interacting with interfaces, computers and technological devices in general, the construct plays an important role. At a low level, while processing information, often people feel annoyed and frustrated; at higher level, mental workload is critical and dangerous as it leads to confusion, it decreases the performance of information processing and it increases the chances of errors and mistakes. It is extensively documented that either mental overload or underload negatively affect performance. Hence, designers and practitioners who are ultimately interested in system or human performance need answers about operator workload at all stages of system design and operation. At an early system design phase, designers require some explicit model to predict the mental workload imposed by their technologies on end-users so that alternative system designs can be evaluated. However, human mental workload is a multifaceted and complex construct mainly applied in cognitive sciences. A plethora of ad-hoc definitions can be found in the literature. Generally, it is not an elementary property, rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed and the skills, behaviours and perceptions of the operator. Although measuring mental workload has advantages in interaction and interface design, its formalisation as an operational and computational construct has not sufficiently been addressed. Many researchers agree that too many ad-hoc models are present in the literature and that they are applied subjectively by mental workload designers thereby limiting their application in different contexts and making comparison across different models difficult.

This thesis introduces a novel computational framework for representing and assessing human mental workload based on defeasible reasoning. The starting point is the investigation of the nature of human mental workload that appears to be a defeasible phenomenon. A defeasible concept is a concept built upon a set of arguments that can be defeated by adding additional arguments. The word ‘defeasible’ is inherited from defeasible reasoning, a form of reasoning built upon reasons that can be defeated. It is also known as non-monotonic reasoning because of the technical property (non-monotonicity) of the logical formalisms that are aimed at modelling defeasible reasoning activity. Here, a conclusion or claim, derived from the application of previous knowledge, can be retracted in the light of new evidence. Formally, state-of-the-art defeasible reasoning models are implemented employing argumentation theory, a multi-disciplinary paradigm that incorporates elements of philosophy, psychology and sociology. It systematically studies how arguments can be built, sustained or discarded in a reasoning process, and it investigates the validity of their conclusions.

Since mental workload can be seen as a defeasible phenomenon, formal defeasible argumentation theory may have a positive impact in its representation and assessment. Mental workload can be captured, analysed, and measured in ways that increase its understanding allowing its use for practical activities. The research question investigated here is whether *defeasible argumentation theory can enhance the representation of the construct of mental workload and improve the quality of its assessment in the field of human-computer interaction*.

In order to answer this question, recurrent knowledge and evidence employed in state-of-the-art mental workload measurement techniques have been reviewed in the first place as well as their defeasible and non-monotonic properties. Secondly, an investigation of the state-of-the-art computational techniques for implementing defeasible reasoning has been carried out. This allowed the design of a modular framework for mental workload representation and assessment. The proposed solution has been evaluated by comparing the properties of sensitivity, diagnosticity and validity of the assessments produced by two instances of the framework against the ones produced by two well known subjective mental workload assessments techniques (the Nasa Task Load Index and the Workload Profile) in the context of human-web interaction. In detail, through an empirical user study, it has been firstly demonstrated how these two state-of-the-art techniques can be translated into two particular instances of the framework while still maintaining the same validity. In other words, the indexes of mental workload inferred by the two original instruments, and the ones generated by their corresponding translations (instances of the framework) showed a positive and nearly perfect statistical correlation. Additionally, a new defeasible instance built with the framework showed a better sensitivity and a higher diagnosticity capacity than the two selected state-of-the art techniques. The former showed a higher convergent validity with the latter techniques, but a better concurrent validity with performance measures. The new defeasible instance generated indexes of mental workload that better correlated with the objective time for task completion compared to the two selected instruments. These findings support the research question thereby demonstrating how defeasible argumentation theory can be successfully adopted to support the representation of mental workload and to enhance the quality of its assessments.

The main contribution of this thesis is the presentation of a methodology, developed as a formal modular framework, to represent mental workload as a defeasible computational concept and to assess it as a numerical usable index. This research contributes to the body of knowledge by providing a modular framework built upon defeasible reasoning and formalised through argumentation theory in which workload can be optimally measured, analysed, explained and applied in different contexts.

Keywords: Mental Workload, Defeasible Reasoning, Argumentation Theory, Human-Computer Interaction

Acknowledgements

First and foremost I wish to thank my family who formed part of my vision, taught me about hard work, self-respect, persistence and independency. Thanks for your unconditional and perpetual support and love.

I would like to express my special appreciation to my advisor at Trinity College Dublin, prof. Stephen Barrett for guiding me through the process, for the freedom given in my research studies and for letting me explore and follow my intuitions.

I would like to thank the members of the Distributed System Group who have contributed to my personal and professional growth at Trinity College Dublin. In particular some has been a source of real friendship as well as good advice and collaboration. A special thank to dr. Mithileash Mohan, dr. Yang Guoxian, dr. Christin Groba.

I am really grateful to the members of the knowledge and Data Engineering Group for the constructive feedback in the experimental studies. Also my appreciation goes to the group that partially but unconditionally supported me financially. A special thanks to dr. Killian Levacher, dr. Rami Ghorab, dr. Kevin Koidl, dr. Catherine Mulwa, prof. Seamus Lawless and prof. Vincent Wade.

I want to say how much I appreciate the members of the Centre for Health Informatics who gave me the opportunity to investigate some of the techniques developed in this thesis also in the field of health-care. A special thanks to prof. Lucy Hederman and dr. Bridget Kane.

I want to express my gratitude to prof. Pierpaolo Dondio for the enriching, fruitful discussions about the theoretical aspects of this study and for reminding me the importance of innovation and originality.

I would like to thank everyone who is somehow involved in the development of this thesis for their critical opinions on many aspects of the research.

Luca Longo
University of Dublin, Trinity College
October 2014

Contents

List of Tables

List of Figures

Acronyms

AAT Abstract Argumentation Theory.

AF Argumentation Framework.

AI Artificial Intelligence.

AT Argumentation Theory.

BCI Brain-Computer Interface.

BPV Blood Pressure Variability.

BS Bedford Scale.

CH Cooper-Harper Scale.

DR Defeasible Reasoning.

EEG Electroencephalography.

EHR Electronic Health Records.

EKG Electrocardiogram.

EMG Electromyography.

EOG Electrooculography.

ER Errors Rate.

ET Estimation Time.

FAF Fuzzy Argumentation Framework.

FAF Fuzzy Preference-based Argumentation Framework.

FL Fuzzy Logic.

FS Fuzzy Sets.

FST Fuzzy Set Theory.

HCI Human-Computer Interaction.

HRV Heart Rate Variability.

KR Knowledge Representation.

MCH Modified Cooper-Harper scale.

ML Machine Learning.

MRT Multiple Resource Theory.

MWL Mental Workload.

NASA-TLX Nasa Task Load Index.

NMR Non-Monotonic Reasoning.

NPP Nuclear Power Plants.

PAF Preference-based Argumentation Framework.

PD Pupil Dilation.

PRAF Probabilistic Argumentation Framework.

QT Queueing Theory.

RRV Respiration Rate Variability.

RSME Rating Scale Mental Effort.

RT Response and Reaction Time.

SWAT Subjective Workload Assessment Technique.

SWORD Subjective Workload Dominance technique.

TR Tapping Regularity.

VAF Value-based Argumentation Framework.

VSAAF Varied-Strength Attack Argumentation Framework.

W/INDEX Workload Index.

WP Workload Profile.

WWW World Wide Web.

Publications related to this thesis

Longo, L. and Barrett, S. (2010a). Cognitive effort for multi-agent systems. In *Brain Informatics, International Conference, BI 2010, Toronto, ON, Canada, August 28-30, 2010. Proceedings*, pages 55–66

Longo, L. and Barrett, S. (2010b). A computational analysis of cognitive effort. In *Intelligent Information and Database Systems, Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part II*, pages 65–74

Longo, L. and Kane, B. (2011). A novel methodology for evaluating user interfaces in health care. In *Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems, 27-30 June, 2011, Bristol, United Kingdom*, pages 1–6

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *INTERACT (4)*, pages 402–405

Longo, L., Rusconi, F., Noce, L., and Barrett, S. (2012b). The importance of human mental workload in web design. In *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, 18 - 21 April, 2012*, pages 403–409

Longo, L., Kane, B., and Hederman, L. (2012a). Argumentation theory in health care. In *Proceedings of CBMS 2012, The 25th IEEE International Symposium on Computer-Based Medical Systems, June 20-22, 2012, Rome, Italy*, pages 1–6

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *UMAP*, pages 369–373

Longo, L. and Hederman, L. (2013). Argumentation theory for decision support in health-care: A comparison with machine learning. In *Brain and Health Informatics - International Conference, BHI 2013, Maebashi, Japan, October 29-31, 2013. Proceedings*, pages 168–180

Longo, L. and Dondio, P. (2014). Defeasible reasoning and argument-based systems in medical fields: An informal overview. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, USA, May 27-29, 2014*, pages 376–381

Chapter 1

Introduction

The increase in use of computers and web-based technologies has led to human activity becoming more cognitively focused. As a result, the objective measurement of Mental Workload (MWL) has become increasingly important, for instance, in the development of digital interfaces, in the design of activities and in understanding human performance within complex systems. If people could accomplish everything they are expected to do quickly, accurately, and reliably using the available resources, this concept would be of little practical importance. However, this is not the case, as other factors such as system and task complexity, multitasking contexts, external influences and inputs, as well as individual fatigue, stress, motivation, skills and knowledge, all influence operator performance and make it difficult to maintain such performance at an optimal level. It has been extensively documented that either overload or underload negatively affect performance (Xie and Salvendy, 2000b). Hence, designers and practitioners who are ultimately interested in system or human performance need solutions to issues that may arise in matters relating to operator workload at all stages of system design and operation (Hart, 2006). In particular, at an early system design phase, designers need some explicit model to predict the mental workload imposed by their technologies on end-users so that system design alternatives can be evaluated. Modern computer-based technologies may impose severe or non-optimal requirements on information-processing capabilities. For instance, modern web-based systems are becoming more complex and interactive, adapting themselves to end-users' preferences and providing them with personalised visual interfaces and contents (Steichen et al., 2011). Designers are becoming increasingly concerned with reducing complexity and maintaining the mental workload imposed on end-users by their systems at an optimal level; evidence of this can be seen in some of the ergonomic principles which have been developed relating to the design of work systems (Nachreiner, 1995). This will, in turn, improve operator satisfaction, promote work efficiency, increase perceived usability and will ultimately lead to increased performance reliability and system success. Additionally, the ability to assess mental workload has a great impact on the investigation of user experience in interacting with computer-based systems. Although MWL has been mainly applied in the automobile and aviation industries, and has been investigated in particular by psychologists, ergonomists and neuroscientists, it can be applied across a much broader domain, where its future impact is likely to be highly significant. This thesis focuses on the challenge of viewing MWL as a defeasible phenomenon and modelling it by building a framework for its representation and assessment with the goal of being applicable in the multi-disciplinary field of human-computer interaction.

1.1 The construct of human mental workload

Despite over 40 years of research, there is still no clearly defined, universally accepted definition of human mental workload (Cain, 2007). The lack of a formal theory of mental workload has led to a proliferation of disparate methods of measurement, with little chance of reconciliation. The operational definitions of MWL from various fields are not consonant on such matters as its sources, methods, consequences as well as measurements. Yet in spite of this lack of agreement about its nature and definition, it remains an important, practical, relevant and measurable entity. The principal reason for measuring mental workload is to quantify the mental cost of performing a task in order to predict operator and system performance (Cain, 2007). Defining human mental workload is a non-trivial problem: the literature suggests it is hard to define due to its multifaceted and multidimensional nature which is dependent on the capabilities and effort of the operators in the context of specific situations. In the mid-1980s Gopher and Dochin noted that no single, representative measure of mental workload existed or was likely to be of general use (Gopher and Donchin, 1986). Nowadays, this point of view seems to still be valid (Cain, 2007). A general intuitive definition is that mental workload is the amount of mental or cognitive work necessary for a person to complete a task over a period of time. Unfortunately, this is a simplistic view of the concept. Additionally, the construct of mental workload is often described by terms such as 'mental strain', 'mental effort', 'cognitive load', 'cognitive effort' and 'emotional strain', making its definition even more confusing.

Despite the plethora of definitions, the literature suggests that mental workload involves the interaction of two principal components: a task and a person. However, this interaction might be mediated by several other elements such as available cognitive resources, the ability and skill of a person, the effort exerted as well as time, context and external factors. Gopher and Dochin consider MWL to be an intervening variable rather than a hypothetical construct. The former is a theoretical concept that is simply a quantity inferred by aggregating the values carried by empirical variables. The latter involves terms that are not reducible to empirical terms and that are not directly observable (Gopher and Donchin, 1986). Xie and Salvendy also believe that mental workload is an intervening variable because it can be measured by other means, such as indicators of performance, psychophysiological measures and subjective ratings which show high correlation to MWL. Therefore, the general consensus among scholars in the field of MWL is that there is no definitive absolute truth in designing, measuring and predicting workload; rather there is only the perception of truth (Xie and Salvendy, 2000b). It is unlikely that anyone would take issue with this stance, in particular because MWL is frequently experienced by humans. However, because mental workload means different things to different researchers, this generates problems for applied research. Unless a universal definition, or at least a general structure, is proposed, each field and perhaps each investigator will continue with their culturally preferred definition of human mental workload.

1.2 Issues in modelling and formalising human mental workload

According to state-of-the-art research on mental workload, developing a general model is a multifaceted problem that must take stock of a broad range of situations, time scales, influences and applications. Researchers

have attempted to represent the construct in several ways, influenced by their knowledge-bases and the context of application. This has led to proposals computational models with different workload attributes and features manipulated and aggregated in various ways. Some of these models fail to consider individual differences, while others do not take into account external factors. In addition, some workload attributes might be affected by other attributes or other variables that could be expressed by a hierarchical structure or a graph (Xie and Salvendy, 2000b). Empirical models are the type which tend to be most frequently proposed: these gather subjective psycho-physiological measures from users that are aggregated in distinct ways. In the Nasa Task Load Index (NASA-TLX), for instance, six attributes are gathered after the execution of a task, and a weighted average is computed considering the subjective preference of attributes provided by the user. In the Subjective Workload Assessment Technique (SWAT), three dimensions are described by three discrete values and users are required to sort each of the 27 possible combinations from the one representing the lowest mental workload to the one that represents the highest (Reid and Nygren, 1988). Correlation coefficients are subsequently used to aggregate partial results to obtain a final workload scale. The complexity of this procedure has prompted other researchers to focus on ways of simplifying it (Luximon and Goonetilleke, 2001) by introducing a pair-wise comparison procedure among the dimensions, with both discrete and continuous scales. In a more recent multidimensional subjective assessment instrument - the Workload Profile, based on multiple resource theory (originally proposed by Wickens and recently reviewed in (Wickens, 2002)) - a simple computational mechanism adds up workload attributes to provide an overall workload score (Tsang and Velazquez, 1996). Another class of computational models have been developed from analytical techniques that do not involve end-users, but rather require inputs from experts. Examples are mathematical models, task analyses and computer simulations mainly based on information, control or queuing theories (Xie and Salvendy, 2000b). However, the limitations of these models are that they need a sophisticated design and great understanding because usually they require well-defined input parameters to fully operate. Furthermore, the compulsory inputs which are very often required for these models, can be partial or not available at all, as in the case of human-computer interactive systems, thereby reducing the applicability of analytical models.

According to (Young and Stanton, 2002b) there are other issues associated with the representation, formalisation and assessment of mental workload. These include subjectivity in the interpretation, the perception as well as the measurement of the construct. Young and Stanton, along with many other researchers, have proposed their own definition of mental workload according to their interpretation of the available literature: their own knowledge-bases, background and experience served for the construction of an ad-hoc computational model. The second issue is the introduction of subjectivity into the definition of mental workload by allowing such a definition to be influenced by personal goals that might vary between individuals and across situations. This issue makes mental workload a context-specific and user-centred construct. The third issue is the complex problem of the measurement of MWL. Several authors, notably (Hart and Staveland, 1988), maintain that subjective ratings represent the only index of 'true' human mental workload and are preferred over physiological or performance measures. Subjective scores have been proved sensitive to perceived difficulty, demand for multiple resources and changes in effort. However, despite the extensive use of subjective ratings, there is the problem of validating them against objective demands or by correlation with other measures.

1.3 Assuming human mental workload as a defeasible phenomenon

The literature suggests that modelling the construct of mental workload not only involves a consideration of the context of use, the state of the user, as well as the complexity and demands of the task under consideration. It also involves the interpretation of the available workload literature for the selection of those attributes which are believed to influence mental workload, and for the analysis of their interaction in a given context. In addition, each attribute can be vaguely defined, introducing a form of uncertainty. Eventually, these attributes can be aggregated in various ways and each can have a different impact on the workload prediction. To clarify these difficulties, consider the following illustrative reasoning that a designer might follow to represent and assess the workload imposed by a web-based interface on a skilled user after interaction:

> *The mental demand of the task to which the user is exposed has been rated as low; thus the designer can infer a low workload. If the 'mental demand' attribute is the only evidence available, the majority of mental workload designers would be likely to infer the same conclusion. However, if it is also known that the user interrupted the execution of the task a number of times, then the previous conclusion could be retracted, inferring a higher workload. Since new evidence has entered the reasoning process, the conclusion is now different. Yet, if it is also known that the user was highly knowledgeable on the task, additional evidence is available from which a lower degree of workload could be inferred, retracting again the previous prediction. However, if the overall performance on the task was perceived as being poor, an inconsistency now arises and the conclusion could be revised upwards to indicate a higher workload. Although the task was not demanding and the user was skilled, external distractions might play a role in increasing the task completion time, minimising performance. The designer might eventually infer a relatively high degree of mental workload because the attributes 'time' and 'distractions' are preferred over 'skill' and 'task complexity'.*

The above example shows how mental workload can be seen as a defeasible phenomenon, starting with the reasonable assumption that it is a complex construct built upon a network of pieces of evidence. The second assumption is that the understanding of the interactions among these pieces of evidence is essential in defining and assessing it. These assumptions are the key components of a defeasible phenomenon: a concept built upon a set of arguments that can be defeated by adding additional arguments. The word 'defeasible' is inherited from Defeasible Reasoning (DR), a form of reasoning built upon reasons that can be defeated. Here, a conclusion or claim, derived from the application of previous knowledge, can be retracted in the light of new evidence. Defeasible reasoning is also known as Non-Monotonic Reasoning (NMR) because of the technical property (non-monotonicity) of the logical formalisms that are aimed at modelling defeasible reasoning activity (Baroni et al., 1997). NMR differs from standard deductive reasoning because in the former, a conclusion can be retracted in light of new evidence while in the latter, a conclusion follows from a set of strictly true premises. Clearly, the previous illustrative reasoning has the property of non-monotonicity as the tentative inference of the degree of mental workload is retracted several times in the light of new information. The pieces of evidence considered in the example are heterogeneous and evidently characterised by uncertainty and vagueness. Each attribute (demand, distraction, skill, time) might be interpreted and defined subjectively by two different designers, according to their backgrounds and knowledge-bases. Interaction of attributes might generate contradictions that have to be considered for a final inference of mental workload.

1.4 Defeasible argumentation theory

State-of-the-art defeasible and non-monotonic reasoning models are formally implemented by Argumentation Theory (AT), an important topic in the field of Artificial Intelligence (AI). It is a multi-disciplinary paradigm that incorporates elements of philosophy, psychology and sociology. It systematically studies how arguments can be built, maintained or discarded in a reasoning process, and the validity of the conclusions reached. That is to say, it studies how people reason and express their arguments. Argumentation theory has gained importance in computer science, with the introduction of formal and computable models of human-like reasoning. These models have extended classical reasoning models based on deductive logic that become increasingly inadequate for tackling many knowledge representation problems. In particular, the study of non-monotonic reasoning, commonly used by humans, has generated several formal models of logical systems. Examples include default logic and explanatory reasoning, successfully applied in IT applications. Argumentation theory has proven useful for modelling defeasible reasoning as proposed by (Dung, 1995) and (Bondarenko et al., 1997). Furthermore, argumentation can be seen as a particularly useful and intuitive paradigm for doing non-monotonic reasoning, with the advantage that the reasoning process is composed of modular and quite intuitive steps, in contrast to the monolithic approach of many traditional logics for defeasible reasoning. The hypothesis behind this thesis is that mental workload can be reasonably viewed as a defeasible phenomenon, and it can be represented and shaped using defeasible reasoning and formally modelled with argumentation theory.

1.5 Problem statement and research question

Since mental workload may be referred to as a defeasible phenomenon, formal defeasible argumentation theory may have a positive impact on its representation and assessment. Mental workload can be captured, analysed, and measured in ways that increase its understanding, enabling it to be used for practical activities. In this thesis it is argued that the concept could be optimally defined from a different perspective using a framework that permits defeasible reasoning over multiple workload attributes, the vagueness associated with their definition, as well as the conflicts and contradictions that might arise from their interaction. The framework allows different designers to define mental workload according to their expertise, the influence of their various fields of research, as well as their knowledge-bases, intuitions, assumptions, beliefs and contexts of application.

This thesis aims to investigate how non-monotonic techniques, developed in the field of defeasible argumentation theory by AI researchers, could be effective in modelling and assessing mental workload. We contend that there is a gap between theoretical cognitive models and the current landscape of computational models of human mental workload. Theoretical approaches have been introduced over the last 40 years, but their intrinsic high-level and ad-hoc view of mental workload has often been criticised due to their non-extensibility and comparability. These approaches have been mainly used by psychologists and ergonomists in the aviation and automobile industry, with sporadic applications in the more general field of human-computer interaction. For this reason the aim is to design a more extensible framework for representing and predict-

ing mental workload in the field of Human-Computer Interaction (HCI). We reject any reductionist approach in which rigid formulas are encoded in dedicated infrastructures; instead we present a modular defeasible reasoning framework in which a workload designer's knowledge-base can be formally translated into interactive arguments and employed for representing and assessing mental workload. This solution is modular in the sense that it is built upon layers aimed at guiding designers in mental workload representation and assessment. The framework should offer a practical alternative solution to those scholars interested in engaging with the multi-disciplinary field of mental workload and human-computer interaction. This proposal was firstly acknowledged by the reviewing committee at a doctoral consortium of a top-tier conference in the field of HCI (Longo, 2011) and was encouraged at another top-tier doctoral symposium in the field of user modelling (Longo, 2012). The rationale behind the proposal of such a qualitative framework is:

- to provide a lightweight methodology, based upon defeasible argumentation theory, to facilitate the development of future computational models for representing and assessing mental workload in HCI;
- to encompass and abstract existing and future mental workload assessment models in order to facilitate their comparison and evaluation;
- to move the community towards a more robust definition of human mental workload;
- to account for the uncertainty and vagueness in defining and assessing mental workload;
- to allow a workload designer to deal with the potential conflicts that might arise from the interaction of those pieces of evidence believed to influence mental workload;
- to promote the proposal of technologies and solutions that take stock of mental workload in the field of HCI.

The research question investigated is:

Can defeasible argumentation theory enhance the representation of the construct of mental workload and improve the quality of its assessment in the field of human-computer interaction?

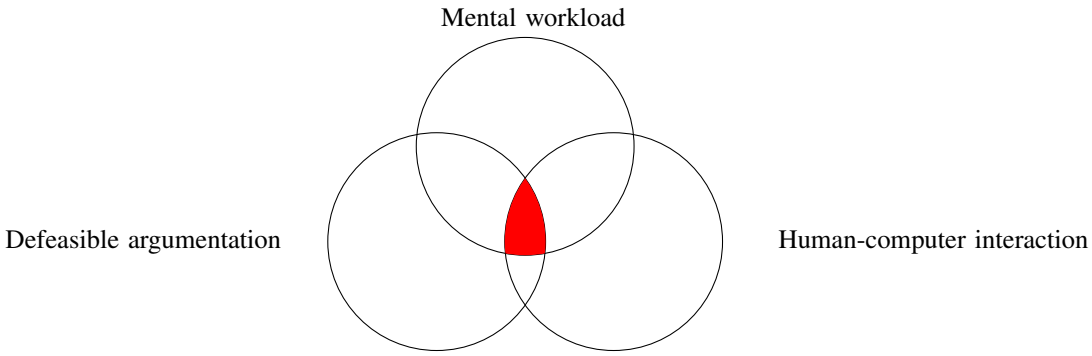


Fig. 1.1: The scope of the research

1.6 Research methodology and contribution

Figure 1.1 shows the scope of the thesis and in order to answer the research question, the following objectives are set:

1. to define the recurring concepts employed in mental workload measurement techniques;
2. to define the notions of non-monotonic, defeasible reasoning and formal argumentation theory;
3. to design a modular framework for representing the complex construct of mental workload as a defeasible phenomenon by employing argumentation theory and argument-based computations;
4. to investigate the capacity of the designed framework to reproduce and abstract state-of-the-art workload assessment techniques and to show how these are particular instances of the framework;
5. to investigate the quality of the assessments produced by a brand new instance of the defeasible framework in the field of HCI.

The research methodology adopted is mixed. Firstly, there is a literature review to identify theoretical knowledge in mental workload modelling. The output of the review has led the author to the formulation of the research question and the design of the framework. The instantiation of the framework follows an inductive theoretical approach: the subjective theoretical knowledge-base and expertise of a MWL designer is required and then translated into a particular instance of the framework. This qualitative approach is followed by a more quantitative method: different instances built with the framework are subsequently quantitatively evaluated. An empirical user study is designed to accomplish objectives 4 and 5 involving 40 participants who are required to fill in questionnaires, providing the numerical inputs to the previously constructed instances. Statistical methods are then employed to analyse the quality of the mental workload assessments produced by these instances. In detail, analysis of variance, multinomial logistic regression and correlation coefficients are adopted to investigate different properties of the mental workload assessments.

The main *contribution* of this thesis is the introduction of a methodology, developed as a formal framework, to represent mental workload as a defeasible computational concept and to assess it as a numerical usable index. This research contributes to the body of knowledge by providing a modular framework built upon defeasible argumentation theory in which mental workload can be robustly modelled, defined, measured and applied in different contexts. The thesis concretely tries to put together the field of Human-Centred Computing, where Human-Computer Interaction is a central part, and Computing Methodologies, where Artificial Intelligence is a major paradigm, and non-monotonic reasoning a part of it[1]. The thesis aims to be appreciated both by scholars familiar in the applications of non-monotonic reasoning as well as HCI informaticians, in particular end-user design experts.

[1] According to the Computing Classification System, 2012 Revision by the Association for Computing Machinery.

1.7 Thesis outline

State-of-the-art: human mental workload - Chapter 2 is devoted to accomplish objective 1, with a literature review of state-of-the-art approaches for modelling, defining and measuring mental workload as well as computational techniques for assessing it, with applications in HCI. This chapter highlights the issues in modelling human mental workload and it introduces why it can be seen as a defeasible phenomenon.

State-of-the-art: defeasible argumentation theory - Chapter 3 is aimed at achieving objective 2 by reviewing state-of-the-art solutions for modelling defeasible reasoning activities and introducing the basic building blocks of non-monotonic logics, notions that stand at the core of this thesis. Subsequently, formal Argumentation theory (AT), based upon these notions, is described with a particular emphasis on its role for knowledge representation.

Design - Chapter 4 is aimed at accomplishing objective 3 by designing a computational framework, based on AT, for mental workload representation and assessment. This framework is built considering the issues and properties of state-of-the-art approaches for modelling and assessing mental workload, emerged in chapter 2, and state-of-the-art computational techniques of argumentation theory, as emerged in chapter 3.

Implementation and instantiation - Chapter 5 describes how the previously designed framework has been implemented in practice and can be actually employed by a MWL designer. The aim is to accomplish objective 4 by instantiating the framework. Specifically, the chapter firstly shows how two state-of-the-art subjective mental workload assessment techniques - namely the NASA Task Load Index and the Workload Profile - can be replicated and translated into two computational instances of the framework. Secondly, it describes how a brand new instance of the framework can be developed from scratch.

Evaluation - Chapter 6 addresses objective 5 by evaluating the assessment capacity of previously developed instances in the field of human-computer interaction. An experimental study is set requiring human participants to perform a set of web-based tasks. The evaluation strategy includes a comparison of the degree of sensitivity, diagnosticity and validity of the mental workload assessments produced by the brand new computational instance, against the two selected state-of-the-art subjective assessment techniques.

Discussion and application - Chapter 7 is devoted to a critical examination and discussion of findings, highlighting advantages and limitations of the application of defeasible argumentation theory for modelling and assessing human mental workload. In addition, it illustrates, through examples, how the mental workload assessments produced by the brand new instance evaluated in chapter 6, can be practically employed for supporting and enhancing the design of HCI-based systems and applications.

Conclusions Chapter 8 summarises this thesis underlying the achievements, the major and minor contributions to the body of knowledge, strengths and weaknesses, as well as illustrating open research issues and delineating future directions.

Chapter 2

Literature review of mental workload

This chapter is a review of the construct of mental workload, starting with a presentation of the relevant theories, definitions and reasons for its measurement. A list of criteria for developing and selecting workload assessment techniques is introduced, followed by a description of the available typologies of measures. A detailed section subsequently introduces the computational strategies employed by state-of-the-art techniques, for the assessment of an index of mental workload. Fields of application of workload-based models are briefly presented with particular emphasis on human-computer and human-web interactive domains, these being the contexts of application of this thesis. The goal of the review is to identify those recurrent dimensions and arguments employed by state-of-the-art representation and assessment techniques, as well as the computational techniques for aggregating these towards a numerical representation of workload. A critical discussion highlights the gaps and limitations of the current state-of-the-art research in mental workload. In turn, a list of properties of an ideal framework for workload representation and assessments is presented. This supports the research question, which shows how the construct of mental workload can be seen as a defeasible phenomenon, which is the core premise of this thesis.

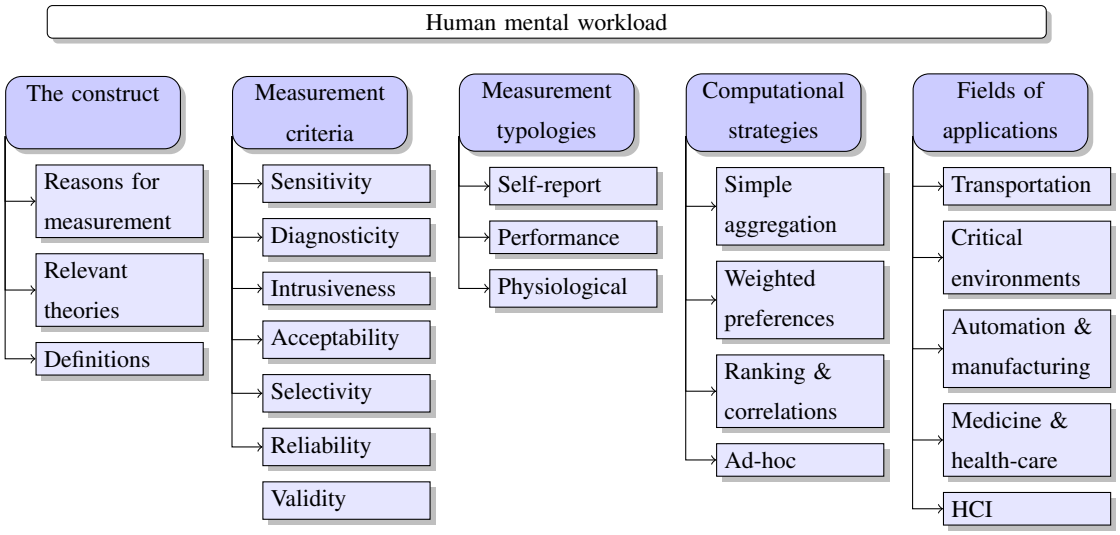


Fig. 2.1: Structure of the literature review of human mental workload

2.1 The construct of human mental workload

The concept of human Mental Workload (MWL) has a long history in the fields of ergonomics and psychology, with several applications in the aviation and automobile industry. Although the concept has been under investigation for the last four decades, there is no clear definition of mental workload that has a general validity, and that is universally accepted. Most of the work concerning mental workload was done in the seventies and eighties when the proliferation of computer-based systems was not as extended as it is nowadays. Until the early nineties, definitions of mental workload in different research fields seemed to conflict with each other in relation to their sources and mechanisms, as well as their consequences and measurements (Huey and Wickens, 1993). Unfortunately, the situation today, is little different , and although several practical applications have been created in the last few years, these are still based on earlier findings and theories, without a proper re-investigation into emerging fields, such as the multidisciplinary domain of human-computer interaction HCI. This state-of-the-art research is also justified by the fact that defining human mental workload is a non-trivial problem. This complexity was earlier acknowledged by Gopher and Dochin (Gopher and Donchin, 1986) who felt that no representative measure of mental workload exists, or is likely to have a general use. In fact, in their contributions to the field, the authors were not capable of indicating how many workload dimensions are necessary or sufficient for a strong assessment. This complexity is also acknowledged in a more recent review confirming that mental workload is difficult to be uniquely defined, due to its multi-faceted and multi-dimensional nature, as it depends on 'the capabilities and effort of the operators in the context of specific situations' (Cain, 2007).

2.1.1 Reasons for measuring MWL

The main reason for assessing mental workload, is to measure the mental cost associated with performing a certain task, with the objective of predicting operator and system performance (Cain, 2007). Modern technologies and human-computer interactive systems, have become increasingly complex, with augmentations in the degree of increased workload on operators. In turn, this has increased the likelihood of exceeding the limitations of the information-processing capacity of human operators, increasing the need for reliable computational models to assess the mental workload, resulting from alternative design options (Eggemeier and O'Donnell, 1998). According to Xie and Salvendy, these computational models need to be built for the prediction of workload levels, mainly when a system is at an early design phase, when it is conceptualised and early operationalised. At this stage, the system can not only be optimised in relation to workload, but can also be a form of guidance for designers, who can better interpret predictions and assessments, thus making suitable structural system changes to benefit the user (Xie and Salvendy, 2000b). Assessing mental workload is also aimed at improving user engagement and satisfaction by designing, for instance, more intuitive interfaces, or creating more effective procedures. Yet, the construct can be applied for legal reasons: it can be adopted to enhance the usability assessment of user interfaces that need to be legally certified.

It has been proved extensively that both underload and overload can degrade performance (Lysaght et al., 1989; Young and Stanton, 2002a), and as a consequence, affect the efficiency of a system as a whole (Xie and Salvendy, 2000b), as well as having a negative impact on human performance (Huey and

Wickens, 1993, Ch. 2). For instance, in present human-computer interactive domains, such as the World Wide Web (WWW), situations of underload or overload can cause websurfers to leave the website, with evident repercussions on the success of the website itself. The assumption in design approaches is that as task difficulty increases, perhaps due to a complex interface, workload increases, and performance usually decreases. In turn, errors are more frequent, response times increase, and fewer tasks can be completed within a unit of time, with changes in performance strategy, and with a smaller mental residual capacity for dealing with other tasks (Huey and Wickens, 1993). On the other hand, when task difficulty is negligible, as in monitoring technologies, interfaces can impose a low amount of workload on operators: this situation should also be avoided, as it leads to difficulties in maintaining attention and increases reaction time (Cain, 2007). Human factors and ergonomic scholars, as well as engineers and researchers, in human-computer interaction, agree that physical workload should be minimal, but mental workload should be optimised. It is generally accepted that human performance achieves its highest level when the demands of a task are matched to the mental capacities available to the operator. In addition, designers of complex systems agree that removing as many tasks as possible from operators, by automation of tasks, does resolve the problem of decreasing mental workload, but it generates other effects such as underload. Rather than implementing automation, and decreasing imposed demands on operators, mental workload practitioners should design tasks, and use available systems and technologies in a way that allows exploitation of the unique characteristics of individuals, such as their skills, knowledge and flexibility (Young and Stanton, 2006).

In summary, mental workload is an important factor to take into consideration in system design (Longo et al., 2012b), and its formalisation as a computational concept is useful for optimising system performance with manipulable numerical values (Diane Kuhl, 2000). This optimisation can support and increase productivity as well as operator satisfaction and user engagement, whilst minimising human errors, and also improving system safety. Figure 2.2 shows the disadvantages behind underload and overload situations, as well as the advantages offered by optimal workload.

Underload	*Optimal Workload*	*Overload*
low sustained attention high reaction time low performance	high user satisfaction high system success low error rate high productivity/safety	high response time/error rate small mental residual capacity low performance

Fig. 2.2: Disadvantages associated with low and high mental workload levels and advantages of optimal workload

2.1.2 Relevant theories

In order to understand the construct of human mental workload, it is essential to briefly introduce a few related concepts and relevant theories, most notably the concepts of *limited cognitive processing capacity* and *performance*. Kahneman, in the early seventies, referred to this using the metaphor of a 'single undifferentiated *capacity*, the modal view', from which a limited pool of *resources* are available to humans to perform tasks (Kahneman, 1973). His theory, the so-called *single resource theory*, is in contrast to the *Multiple Resource Theory (MRT)* proposed in the early nineties, by professor Wickens. MRT suggests that *capacity* is the upper or maximum limit of the cognitive processing capability and *resource* represents the mental effort exerted for the improvement of processing efficiency (Wickens, 1991). This view has been refined, in the last decade, by the same author, and nowadays it is the most prevalent theory in the field of mental workload (Wickens, 2002). In more detail, *single resource theory* assumes that the capacity of the information-processing system, and the availability of resources are not static concepts, rather they have a certain degree of elasticity. To support this, resources might be increasingly allocated and employed as a consequence of an increase in processing load. A linear relationship is supposed to exist between their allocation and task performance,up to the moment that all resources are allocated and employed, and performance remains stable (Kahneman, 1973). For instance, the additions of a secondary visual task to a primary auditory task can leave performance at the same level, even if time-sharing between the two tasks is very successful. Although the theory can be applied across different scenarios, it is not capable of explaining the reasons that lead to unaffected performance, even if time-sharing is effective, for instance: why trained operators can manage time sharing of multiple different tasks. The limitations of the single resource theory are overwhelmed by the *multiple resource theory*, which is based on the assumption that a pool of resources exists for different modalities (Wickens, 1984). According to Wickens, human operators do not only have one single source of information processing that can be adopted, but rather different pools of resources that can be tapped into concurrently. His model, depicted in figure 2.3 (each box corresponds to a cognitive resource) is built on 4 dichotomies [1] of information processing:

- the stages of processing (perceptual, central and response)
- the codes of processing (verbal and spatial)
- the modalities of the input (visual and auditory), and output (speech and manual)
- responses (manual, spatial, vocal, verbal)

The *stages of processing* dimension refers to those perceptual and cognitive activities (for example involving working memory), that use different resources than those activities that denote the selection and execution of actions (response). An operational example from aviation can clarify this stage. Consider an air traffic controller that is required to manually or vocally acknowledge each change of the aircraft state. That is a response demand. This activity would not disrupt or mitigate the capability to maintain an accurate view of the airspace, which is a perceptual/cognitive demand. The *codes of processing* dimension denotes that spatial activity employs a different pool of resources than verbal or linguistic activity. It underlines the distinction between analogue/spatial processes and categorical/symbolic processes (usually linguistic or verbal). For

[1] A dichotomy is any splitting of an element or set into exactly two non-overlapping parts. It is a partition of an element, or a set divided into two sub-parts or subsets that are both jointly extensive and mutually exclusive. The former refers to the idea that everything must belong to either one part or another, while the latter refers to the property that nothing can belong to both parts simultaneously. A dichotomy is frequently called a bipartition.

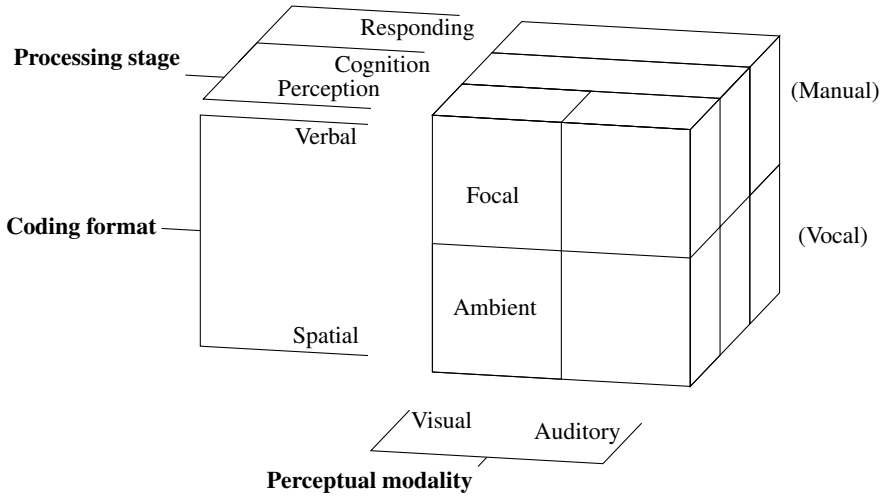


Fig. 2.3: The 4-D Wickens Multiple-resource model

instance, while driving a car, it can be useful to predict the potential dangers of manually dialling mobile phones, considering the visual, spatial and manual demands that are required to drive a vehicle, which would conflict with those similar demands required to dial a mobile phone and thus suggesting the benefits of using voice dialling. This dichotomy also refers, for instance, to the larger disruption of background music when it has words, as opposed to when it does not, in a typical office environment where verbal processing frequently occurs. The *modalities* dimension, included within perception but not within cognition or response, denotes that auditory perception employs a different pool of resources than the one used by visual perception. In other words, this suggests that a human can split their attention between eyes and ears better than between two auditory or two visual channels: cross-modal time-sharing is easier than intra-modal time sharing. In other words, listening to someone whilst simultaneously watching something cause less interference to one another than listening to two conversations simultaneously do. The fourth dimension, the *visual channel*, is aimed at distinguishing between focal and ambient vision, and is a nested dimension within visual resources. Focal vision is needed for pattern recognition, such as identifying small objects, or reading text and fine details. Ambient vision is required for sensing ego motion and orientation, that is, the direction and the speed employed to move through the environment. For example, walking in a corridor while reading a book, represents the distinction between ambient vision (walking) and focal vision (reading). Similarly, driving a car in the middle of a lane (ambient vision) while reading a road sign, or an unusual object in the middle of the road (focal vision). Wickens proposed that each dimension has two discrete levels, and as long as two tasks use different levels along each of the proposed dimension, time-sharing is better. In other words, maintaining few things equal, for instance,the same resource demand or the same task difficulty, two tasks that both require one level of a given dimension will have a higher interference to each other than two tasks that require separate levels on the same dimension. For further details, as well as scientific and empirical explanations of the choices of the four dichotomies, we refer the reader to (Wickens, 1991, 2002). Task demand, allocation policy and resource overlap represent mental workload determinants within the scope of the MRT. In addition, this theory is relevant to the scope of this thesis, human-computer interaction, this being a multi-tasking environment that often requires the elicitation of different cognitive resources by human operators. For instance, an operator interacting with a web-based system might concurrently watch a video or listen to streaming music,read

news or interact with e-mails systems. In turn, the four dichotomies of the MRT might all be elicited, and mental workload might be influenced accordingly. This suggests that, for the problem of formalising mental workload, as a computational concept, within HCI, these workload determinants should be taken into account.

2.1.3 Definitions

A general intuitive definition is that 'mental workload (MWL) is the amount of mental work necessary for an individual, or group of people, to complete a task over a period of time'. Unfortunately, this is a simple view of the concept and, as described in a recent review with Cain, several more complex and complete definitions of mental workload exist (Cain, 2007). Gopher and Dochin's definition states that 'mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time'. This definition supports previous theories regarding the limitation of the human information processing system, which cannot be fully used in the execution of a target task. It also refers to mental workload as a construct used to describe the aspects of an interaction between a person and an assigned task (Gopher and Donchin, 1986). The authors initially referred to MWL as a 'mental construct, a latent variable, or perhaps an intervening variable, reflecting the interaction of mental demands on operators by task they attend to'. However, they finally suggested to see MWL as a hypothetical construct rather than an intervening variable, because the concept cannot be described and fully summarised in empirical terms [2] (Gopher and Donchin, 1986). A similar point of view is supported by O' Donnell and Eggemeier: 'workload refers to that portion of the operator's limited capacity actually required to perform a particular task' (Eggemeier et al., 1991; O' Donnel and Eggemeier, 1986). Again, the authors' assumption was that a person has a limited capacity to process information and respond to it. In this case, the task response, and the processing demands exceed the personÕs limited available capacity; the overload that results can be manifested in decrements in terms of performance. For O' Donnell and Eggemeier the principal objective of assessing and measuring mental workload was to specify the amount used by this limited capacity (Eggemeier et al., 1991).

Always in line with the reasonable assumption of the limited capacity of the information processing system, Kramer, Sirevaag, and Braune have described 'mental workload as the cost of performing a task in terms of a reduction in the capacity to perform additional tasks that use the same processing resource' (Kramer et al., 1987). Their definition was driven by their studies in the aviation industry, where dual-task experiments were conducted in order to provide information concerning the mental workload of subjects and their mental residual capacity. Similarly, Lysaght et al. referred to mental workload as, 'the relative capacity to respond, the emphasis is on predicting what the operator will be able to accomplish in the future'. Their working and operational definition was aimed at being general, and not at explaining individual factors that influence the performance of individuals or their perception of the workload of a task. It implies the amount of spare capacity, and also the ability of a person to use that spare capacity in a specific context, and in specific personal situations (Lysaght et al., 1989).

[2] In scientific theories, particularly in the field of psychology, a hypothetical construct is an explanatory variable that is not directly observable, and it differs from an intervening variable because it has properties and implications that have not been demonstrated in empirical terms. On the other hand, an intervening variable can be summarised by findings empirically observed.

Hart suggested that 'mental workload is not an inherent property, but rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviours and perceptions of the operator' (Hart and Staveland, 1988). This definition does not explicitly refer to the limitation of the cognitive processing system, rather it highlights the multi-faceted nature of the construct of mental workload. It embeds notions related to the task involved by the interaction with a person, its complexity, the situation and the conditions under which the user performs it, as well as individual differences such as skills, background, subjective perception and behaviour. This multi-faceted view is also supported in (Cain, 2007) where 'workload can be characterised as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, considering capabilities of the operator to respond to those demands'. Yet, in line with the assumption that mental workload is a multi-dimensional construct, Young and Stanton suggested that 'the mental workload of a task represents the level of attentional resources required to meet both objective and subjective performance, which may be mediated by task demands, external support and past experience' (Young and Stanton, 2002b, 2006). In this definition, the authors assumed that the level of attentional resources has a finite capacity, thus beyond it, any increase in terms of task demand is reflected in terms of performance decrease. They also introduced external factors and individual features as moderators of human mental workload. This multidimensionality is also acknowledged by Vidulich and Tsang (Tsang and Vidulich, 2006) who separated influencing workload factors into two main categories: exogenous task demands, and endogenous supply of attentional resources. The former refers to those factors such as task difficulty, situational contingencies and task priority, while the latter refers to those processing resources aimed at supporting information processing, such as planning, decision making, perceiving, updating memory and response processing. In addition, this supply is mitigated and affected by individual differences such as expertise, knowledge, background and skills (Tsang and Vidulich, 2006).

In summary, a general and commonly accepted definition is not present in literature on mental workload, as agreed in different reviews (Cain, 2007; Gopher and Donchin, 1986; Xie and Salvendy, 2000b). However, according to the aforementioned definitions provided by several practitioners, it is reasonable to view mental workload as a multi-dimensional construct that can be influenced by various factors. Some of these factors can be found in Huey and Wickens (Huey and Wickens, 1993) who provided an overview of many tasks and external variables which contribute to mental workload. Similarly, Xie and Salvendy presented an analysis of factors affecting mental workload, both in single and multi-tasking environments (Xie and Salvendy, 2000b). While in (Tsang and Vidulich, 2006), the authors investigated mental workload against the construct of situational awareness. The multi-dimensional interpretation of mental workload seems to include:

- *task* (exogenous factors - those that are inherent in the situation such as task demands, situation complexity and uncertainty);
- *operator* (endogenous factors - those that are inherent in a person's ability and skill);
- *context and situation* (exogenous factors);

2.2 Criteria for measurement methods

Methodologies for measuring mental workload can be performed either in experimental or operational settings. In the former, there are generally more options than in the latter, and they are usually adopted. However, there exist concerns regarding the application of workload measures, which are practically conducted and based on laboratory studies. For instance, users performing web-based tasks in a laboratory can behave differently than performing the same task in their familiar environments (home, office or other contexts), with assessments of workload differing. This suggests that the context is an important factor that should be accounted for in any measurement method, above all, if applied in the field of human-computer interaction where users can be physically located in different heterogeneous environments. According to this, several criteria exist and have been proposed as a guidance for selecting and developing measurement techniques (O' Donnel and Eggemeier, 1986):

- *sensitivity*: the methodology must have a high reliability in terms of sensitivity to changes in resource demand or task difficulty and in terms of discrimination capacity between significant variations in workload;
- *diagnosticity*: the method should be highly diagnostic, that means being capable of indicating the sources that cause variations in workload and to quantify the contributions by the type or resource demand;
- *intrusiveness*: the methodology should not be intrusive and interfere with the performance of the task of the operator, becoming an important source of workload itself; (this property is referred to as *obtrusiveness* by Wickens in (Wickens and Hollands, 1999, Ch. 11));
- *requirements*: the methodology should require equipment as minimal as possible to avoid impact on operator's performance. (Muckler and Seven, 1992) refers to this as *resource requirements*;
- *acceptability*: the method should have high operator acceptance showing at least face validity [3], without being onerous. (Muckler and Seven, 1992) refers to this as *relative simplicity*.

Wickens et al. (Wickens and Hollands, 1999, Ch. 11) extended these criteria with two further categories:

- *selectivity*: the method should be selectively sensitive only to differences in resource demand, and not to changes in other factors unrelated to mental workload;
- *bandwidth and reliability*: the assessment procedure should be reliable both within and across tests, and it should be capable of rapidly detecting transient changes in workload levels. (Muckler and Seven, 1992; Wierwille and Eggemeier, 1993) respectively refers to this as *transferability*, and *sufficient reliability* highlights the importance of the capability of a technique to be used in different applications.

Other criteria are worth mentioning:

- *construct validity*: a property aimed at really assessing whether an instrument is measuring mental workload. This is inherently difficult to be verified because of the complexity of the construct itself,

[3] Face validity refers to what a concept superficially appears to measure, mainly testing if it looks valid. It is in contrast with content validity that is a more strict property that requires the use of recognised tests or subject experts for evaluating whether evaluated items assess defined content. This includes statistical tests which are in general more rigorous than methodologies applied in face validity tests.

thus two variations have been proposed: *concurrent validity* and *predicting validity* (also referred to as *convergent validity*). The former is when the metrics correlate with operational measures for validation, while the latter is when different workload metrics correlate to each other (Tsang, 2006).

- *generalisability*: a measurement technique should meet *formal constraints* (Cain, 2007) and it should allow comparisons with different techniques, aimed at increasing the understanding of the construct of mental workload (Muckler and Seven, 1992).

The above criteria are all important for the development and evaluation of an assessment technique. However, some of them, such as acceptability, are difficult to be achieved at the beginning of the development of a technique, and can only be obtained after several experiments and user studies. For this reason, some of them are carefully applied to evaluate the proposed defeasible computational framework for mental workload representation and assessment: sensitivity, diagnosticity, concurrent and convergent validity. In the following subsections details of each individual criterium are provided.

2.2.1 Sensitivity

Sensitivity refers to the capability of an assessment technique to reflect changes in mental workload imposed by task performance. O' Donnell and Eggemeier suggest that the theoretical relationship between operator performance and workload can guide the evaluation of the sensitivity of an assessment procedure. This hypothetical relationship is depicted in figure 2.4, where three regions are identified by the relative workload levels imposed on the operator (O' Donnel and Eggemeier, 1986). In region A, low to moderate levels of operator workload are reflected in high and adequate operator performance. In this area, increments of workload do not influence the performance of the operator, who has enough spare information processing capacity to handle increments in workload. In region B, higher levels of workload reflect changes in operator performance because they exceed the capability of the operator to compensate. In this area, a monotonic relationship between the workload and the performance of the operator exists, with primary-task performance decrementing with increments in workload levels. In region C workload is extremely, high and as a consequence the performance of the operator is exceptionally low.

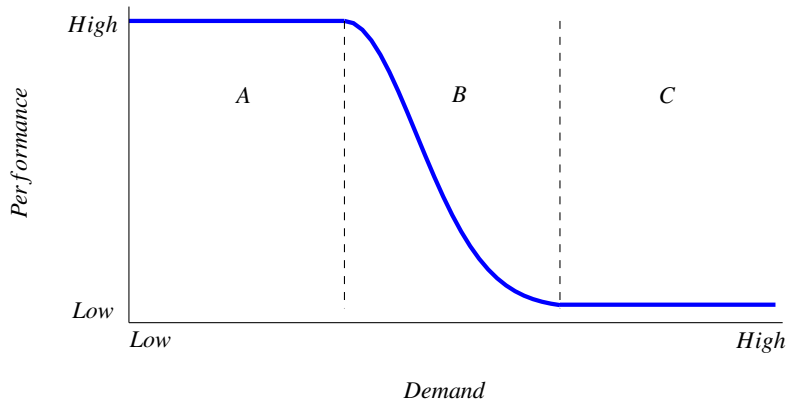


Fig. 2.4: Hypothetical relationship between demand and performance

Formally, sensitivity has been assessed in different ways. (Tsang and Velazquez, 1996) et al. used *multiple regression* to predict performance objective measures through combinations of subjective ratings related to a set of tasks executed by humans. (Rubio et al., 2004) used *analysis of variance* (ANOVA) to investigate the sensitivity of different workload assessment instruments. The aim was to investigate to what extent mental workload indexes, computed by different assessment instruments, varied as a function of objective changes of a set of tasks executed by humans. These changes included single/dual-tasks and manipulation of difficulty. However, a set of tasks might be designed in different ways, manipulating mental demands, required cognitive resources, temporal constraints or even the context of execution. If there are two different tasks, T-tests are usually used to find a significant difference between the two. If more tasks are designed, ANOVA is preferred and Post-hoc comparisons of workload means, such as Tukey or Duncan, tests are used. An assessment procedure with high sensitivity should be able to spot statistically significant differences between the different tasks under examination (Zhang and Luximon, 2005).

2.2.2 Diagnosticity

The property of diagnosticity refers to the capacity of a workload assessment procedure in discriminating between demands on specific resources. In other words, this criterion specifies the ability to discern the cause or typology of workload, and or the capability to attribute it to a specific aspect, or a characteristic of the task performed by an operator (O' Donnel and Eggemeier, 1986). This characteristic relies upon the multiple-resource theory (MRT described in section 2.1.2, page 12) that considers the capacity of the information-processing system being limited and built upon as a set of independent resources, or capacities, which are not interchangeable (Wickens, 1991, 2002, 2008). The approach of multiple-resources holds that the processing capacity devoted to task performance is not unitary, rather it is drawn from a pool of resources that cannot be exchanged with other resources. A workload assessment procedure, in the context of the MRT, is said to be diagnostic if it shows sensitivity to certain demands of resources and not to others. For instance, measures can be, on one hand, extremely diagnostic and reflect a specific variation at a certain stage or position of demand, or on the other hand, can be very low on diagnostic and express general demands. Formally, diagnosticity has been assessed in various ways. (Tsang and Velazquez, 1996) used *canonical analysis* to study the relationship between a set of predictor workload attributes and a set of criterion variables (the different tasks conditions). (Rubio et al., 2004) proposed to measure it with *stepwise discriminant analysis*, to investigate to what extent workload profiles allowed discrimination of tasks.

2.2.3 Intrusiveness

Intrusiveness refers to the degree of a measurement method to influence the performance of a primary task. In other words, the goal for a good measure is to minimise the disruption in ongoing task performance as a result of its application. For instance, the addition of a secondary task may influence the performance on a primary task, by being intrusive. In some settings, such as laboratory environments, this addition might be more intrusive, with influence on the primary task, but it can still be acceptable. In other settings, such as field test situations or simulations, a higher intrusiveness cannot be accepted. Not only does the property of intrusiveness influence performance on a primary task, but it also introduces problems in interpreting the

data and outcomes produced by a measurement technique. The outcomes of a procedure that has a high degree of intrusiveness cannot accurately represent the level of mental workload required by the primary task (O' Donnel and Eggemeier, 1986). Formally, intrusiveness is usually the difference between the performance on the primary task alone, and the performance of the primary task when the mental workload measure is administered. A significant difference is synonymous of high intrusiveness, thus the assessment method should be avoided.

2.2.4 Requirements

The requirements of a measurement technique, in terms of its implementation, refers to those practical constraints such as the training of operators or the need of specific equipment or software for data collection and analysis. The property mainly deals with the complexity of the measurement procedures and apparatus that have to be considered in the choice of the appropriate workload technique. For example, too much equipment might influence the execution of the primary task, and its intrusiveness can generate degradations in terms of performance. Furthermore, sometimes subjects need to be trained specifically and extensively in order to achieve a reasonable and stable level of performance. However, this might not represent an important issue, but it influences the time necessary before a measure can be taken (O' Donnel and Eggemeier, 1986).

2.2.5 Acceptability

The property of operator acceptance refers to the degree of approval of the measurement procedure by the operator. In particular, those workload measurement methods that consider operational questions must be evaluated to the perception of the subject of the utility and validity of the procedure. This issue is extremely important when the subject population includes an operator with a high degree of experience in the system being evaluated. As a matter of fact, a procedure that is perceived as being artificial or extremely intrusive might be ignored or even worse, performed at substandard levels. In turn, this non-optimality in performing the procedure affects the potential accuracy, effectiveness and correctness of the technique being evaluated. In general, the operator acceptance is higher in the case where the technique is less intrusive, while the face validity may increase acceptance. In other words, acceptance varies among subject populations, and generally the rejection of measurement techniques that lack face validity will increase when experimental situations closely approximate operational environments, familiar to operators. O' Donnell and Eggemeier suggest that if the usefulness of some measure is not fully clear to a subject, explaining its use in details can help the operator to accept it (O' Donnel and Eggemeier, 1986). Acceptability is related to the criteria of *relative simplicity* proposed in (Muckler and Seven, 1992) in which it is suggested that a method should be simple, showing understandability, and directness in order to minimise the interpretation needs and to help designers in achieving accuracy of the definition of the measure. Operator acceptance, as well as face validity are, of course, important factors to consider, but alone they do not guarantee that a technique will reflect the objective level of mental workload. However, the capability to adhere to these criteria can assure that the full potential of an assessment procedure is achieved, above all in operational systems.

2.2.6 Selectivity

The property of selectivity refers to the selective sensitivity to mental workload, rather than to changes of other factors such as physical workload, which might not be related to mental workload, as well as information-processing ability. Selectivity represents the validity of an assessment procedure for mental workload. A measure can be only sensitive to mental workload, or it can be sensitive to other factors as well, such as physical workload. In the case where the assessment procedure is sensitive to other factors, it might be a reason to discard it for mental workload assessment purposes, depending upon the test environment and the task. In other words, if a procedure is sensitive either to mental workload or physical workload, it can be adopted as an indicator for mental workload when physical effort is not required (Wickens and Hollands, 1999, Ch. 11). Formally, selectivity of a measure can be obtained by providing the same task under different physical workload situations which are physical demanding; for instance, an attribute believed not to influence mental workload. If there is a significant difference between the values computed by a workload instrument, then the measure is said to lack selectivity (Zhang and Luximon, 2005).

2.2.7 Bandwidth and reliability

An assessment procedure of mental workload, as any other assessment method of behaviours, should be reliable. This property refers to the estimate of mental workload that needs to be reliable, both across tests and within tests. In particular, measurement procedures deployed in laboratories do not necessarily have to behave as well as operational settings. A workload index should produce the same estimate for a given task and operator, that means it should show repeatability, with small variance compared to the main effects (Cain, 2007). Reliability, as proposed by (Wickens and Hollands, 1999, Ch. 11) is similar to the *sufficient reliability* definition proposed in (Muckler and Seven, 1992) where authors, in addition, indicated that one potential source of unreliability is the changing nature of the process being measured, and not always the measuring technique. If the task, for which subjects are called to interact with, is not stable over time, then the reliability of a measure is compromised. The bandwidth property, instead, refers to the capacity of an assessment procedure to respond quickly to changes in mental workload (Wickens and Hollands, 1999, Ch. 11). The method should be applied sufficiently and timely quick to capture temporary workload changes (Cain, 2007). Strictly, comparisons of workload levels obtained in different contexts with samples obtained from the same population, should provide a good assessment of reliability.

2.2.8 Validity

The criteria of validity is the extent to which a mental workload assessment instrument is measuring the attribute in question, that being mental workload itself. Different typologies of validity have been proposed: construct, content, predictive, face, concurrent and convergent (Zhang and Luximon, 2005). However, two of these variations have mainly been used: convergent and concurrent validity. The former is assessed by studying the correlation between different mental workload scores, while the latter is by investigating the correlation between mental workload scores and objective performance measures. Formally, Pearson Correlation coefficients have been adopted to study validity and the positive they are, the higher the measure is assumed to be valid.

2.2.9 Summary of characteristics

The aforementioned criteria are valuable factors for evaluating a mental workload measurement instrument. Although they refer to specific properties, they are not independent of each other. For instance, diagnosticity might reduce sensitivity and presuppose the property of selectivity. In general, a desirable mental workload assessment method would have high sensitivity, better if in a high bandwidth, low intrusiveness on the primary task and high reliability, as well as showing concurrent and convergent validity. The property of diagnosticity is important, as well, mainly if there is evidence that a specific stage of information processing is affected (De Waard, 1996). The others are more guidelines for implementing measures in different contexts (simulations or laboratories). In addition, a measure should be used across many research and test settings, leading to the development of standards, and probably, to a better understanding and definition of the measure itself. Sensitivity, diagnosticity, concurrent validity, and convergent validity will be subsequently used to evaluate the defeasible framework proposed in this thesis.

2.3 Measures

Mental workload measurement is as vast and heterogeneous topic as its related theoretical counterpart. Several assessment techniques have been proposed in the last 40 years, and researchers in applied settings have tended to prefer the use of ad-hoc measures or pools of measures, rather than any one measure. This tendency is reasonable given the multi-dimensional property that characterises mental workload. Several reviews have attempted to collate the enormous amount of knowledge behind measurement procedures. According to the review of Gopher and Donchin (Gopher and Donchin, 1986), measurements can be divided into subjective measures, performance measures (primary-task), arousal measures, specific measures and psychophysiological measures. Young and Stanton proposed three broader classes of measures: primary and secondary task measures, physiological measures, and subjective measures (Young and Stanton, 2006). This is also supported by O'Donnell and Eggemeier (O' Donnel and Eggemeier, 1986), as well as Wickens and Hollands (Wickens and Hollands, 1999). Vidulich and Tsang proposed four categories: performance, subjective and physiological measures, as well as multiple measures of workload (Tsang and Vidulich, 2006). Xie and Salvendy introduced a further classification based on empirical and analytical methods (Xie and Salvendy, 2000b). In general, measurement techniques are organised into three broad categories, which have emerged in recent scientific articles (Cain, 2007; Tsang, 2006; Wilson and Eggemeier, 2006; Young and Stanton, 2004):

- *self-assessment measures*: these includes self-report measures and subjective rating scales;
- *performance measures*: these consider both primary and secondary task measures;
- *physiological measures*: those derived from the physiology of the operator.

These three categories will be individually described in the following sections, highlighting which of the measurement criteria (previously described in section 2.2) they meet, and their fields of application, outlining the advantages and limitations in relation to their potential use in the field of human-computer interaction, this being the focus of this thesis. More qualitative measures such as open questionnaires and interview techniques, although they are informative, will not be described. The attention is only on those quantitative measures that can be and have been validated empirically.

2.3.1 Self-report measures

The class of self-report measures is often referred to as subjective measures. The former term is preferred to the latter term, for essentially reducing confusion with other categories, such as physiological assessment methods, which can also be subjective. This class is obtained from the direct estimation of task difficulty through subject analysis, and it relies on subjective perceived experience of the interaction operator-system. Subjective measures are appealing to many workload practitioners and researchers because it is strongly believed that no one, but the person concerned with the task, can provide an accurate and precise judgement with respect to the mental workload experienced. In addition, these measures are easily administered, thus they can scale to several subjects. Potentially, in the field of human-computer interaction, these are the most promising techniques to gather behavioural information of users executing tasks. They are relevant to the thesis because they have been adopted in the user studies aimed at evaluating the proposed defeasible framework.

Various dimensions of workload are considered in self-report measures; these include effort and performance, as well as individual differences such as the operatorÕs emotional state, attitude and motivation (De Waard, 1996). The majority of self-report assessment techniques have shown high sensitivity to situations of underload and overload, but do not have high predictive capacity to predict optimal workload. In addition, they can detect changes in region C (loss of performance, figure 2.4, page 17), when severe situations of overload occur, and they are apparent due to extremely low performance or because the operator quit the task (De Waard, 1996; Tsang and Vidulich, 2006) have identified three variables for classifying subjective rating assessment procedures: dimensionality, evaluation style and immediacy. The property of *dimensionality* refers to the dimensions considered in the assessment procedure. Subjects might be asked to rate their experience using single or multiple dimensions (Young and Stanton, 2006). The *evaluation style* indicates the way the rating of the experience is provided. This can either be on an absolute rating or a relative rating in which one experience is compared against another. The property of *immediacy* refers to the time period the subjective rating is provided. This can be done immediately after the execution of a task, a set of tasks or even after an entire experiment. (Tsang and Vidulich, 2006) note that, despite the fact that it is theoretically possible to combine the three properties in different ways, for classifying self-report measures, in practical settings, only two combinations are important . The most adopted techniques combine multi-dimensionality, immediacy and the absolute evaluation style. However, another typical choice is the opposite configuration that is based on a uni-dimensional assessment method, employing a relative comparison evaluation style with ratings collected retrospectively and not immediately after the execution of a task. According to the authors, the assessment of the operator, provided immediately after the completion of a task, should take advantage of the freshest memory related to the experience of executing the trial. In turn, the property of immediacy minimises the potential negative and damaging effect of the guessed answers of the operator, in the case where they are not provided immediately after the trial. In addition, an absolute scale design supports the consideration of the individual workload of each trial, rather than the workload relating to other conditions. Eventually, a multi-dimensional assessment procedure should support and increase diagnosticity, due to the fact that subjects can be more accurate in providing ratings and describing those experimental conditions and factors that have influenced their experience (Tsang and Vidulich, 2006).

Two subjective workload assessment procedures have emerged in the last three decades: the National Aeronautics and Space Administration's Task Load Index (NASA-TLX) (Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). Both the techniques are multi-dimensional, they adopt an absolute scale design, and are based on immediate rating scales. They have been compared several times by various researchers and practitioners (Rubio et al., 2004; Vidulich and Tsang, 1986) and they have been demonstrated to have good concurrent validity with performance as well as high diagnosticity and sensitivity to manipulation of difficulty of tasks. A third, more recent subjective assessment procedure was proposed in (Tsang and Velazquez, 1996): the Workload Profile (WP). This is a relatively new multi-dimensional assessment tool, thus, ithas not been as extensively tested as the NASA-TLX and the SWAT procedures have. It is based on the multiple-resource theory (MRT, page 12) of Wickens (Wickens, 2008; Wickens and Hollands, 1999), and the assumption is that the mental workload dimensions could be represented by the resources dimensions hypothesised in the multiple-resource theory (Tsang and Velazquez, 1996). NASA-TLX, SWAT, and WP are multi-dimensional absolute immediate ratings in contrast with uni-dimensional approaches such as the Cooper-Harper Scale (CH) (Cooper and Harper, 1969) and its modified versions (Modified Cooper-Harper scale (MCH) (Wierwille and Casali, 1983; Wierwille and Eggemeier, 1993)), the Rating Scale Mental Effort (RSME) (Zijlstra, 1993), the Subjective Workload Dominance technique (SWORD) (Vidulich and Ward Frederic G., 1991) and the Bedford Scale (BS) (Roscoe and Ellis, 1990). These measures are usually easily applicable and investigable, but they only provide a general workload score. For this reason, although being sensitive and having low implementation requirements, their diagnostic degree is very poor. In general, as suggested in (De Waard, 1996), diagnosticity is higher in multi-dimensional scales, and sensitivity to task demands is larger in unidimensional mental workload ratings.

In the following paragraphs, the aforementioned subjective assessment procedures are briefly described, highlighting the workload attributes they incorporate. Some of these procedures are re-introduced in section 2.4 for a more detailed description of the computational strategy for aggregating their workload attributes.

Nasa Task Load Index (NASA-TLX)

The NASA-TLX has gained a lot of acknowledgments from different researchers, and it is widely adopted in different contexts, fields and environments (Hart, 2006). It is based on six-sub-scales:

- mental demand;
- physical demand;
- temporal demand;
- performance level;
- effort level;
- frustration level.

Hart et Al. believed that the various dimensions used for assessing mental workload could be clustered together by the six proposed sub-scales, and the assumption is that some combination of these dimensions is likely to represent the mental workload experience by the majority of subjects performing most tasks. These

six dimensions were the result of a multi-year research program aimed at identifying those factors that were mainly responsible for changes in subjective workload between, and within different types of tasks (Hart and Staveland, 1988). The aggregation of the dimensions follows a weighted procedure. In detail, each dimension is weighted according to their relative importance, provided by the subject, and a final overall workload rating, from 0 to 100, is computed. Further details of the computational model behind the NASA-TLX, as well as some computational properties, are described in section 2.4 (page 32). Appendix A.1 shows the questionnaire associated with the procedure.

Subjective Workload Assessment Technique (SWAT)

The SWAT procedure is based on three sub-scales:

- time load;
- mental effort load;
- psychological stress load.

Reid's works suggested that just three components could largely explain mental workload (Reid and Nygren, 1988). Each dimension is weighted according to the subject's ratings of the workload delivered by every combination of the various levels of workload (1-3) in each of the three scales. Subsequently, a conjoint analysis is performed to generate a look-up table aimed at translating the ordinal ratings into ratings with interval-scale properties. The original procedure was recently simplified, and the resulting model was shown to have higher sensitivity for low mental workloads (Luximon and Goonetilleke, 2001). The computational counterpart of this model, as well as some computational aspects, are described in section 2.4 (page 32). Appendix A.2 shows the questionnaire associated with the procedure.

Workload Profile (WP)

The Workload Profile (WP) assessment method is a self-report measure based upon the multiple-resource theory of Wickens (page 12) (Wickens, 2008; Wickens and Hollands, 1999). The workload dimensions considered in this method are those hypothesised in the multiple-resource theory and it is these diverse demands that can be inflicted by a task. They include perceptual/central processing, response selection/execution of both considered stages of processing, spatial and verbal processing, referred to as codes of processing, visual and auditory input processing, and manual and speech output. WP is similar to Workload Index (W/INDEX), another multi-dimensional subjective procedure (North and Riley, 1989). The main difference is that in the former procedure, subjects are asked to estimate the proportion of attentional resources for a task or multiple tasks, on different dimensions, while in the latter experts provide a priori estimates used to predict eventual performance. WP is an assessment technique, while W/INDEX is a projective technique. In WP the diagnosticity of the multi-dimensional post-task ratings is examined, and subjects are not necessarily expert or familiar with the theory behind the model. On the other hand, in W/INDEX, researchers employed for providing estimates are experts in the field of mental workload, and they have previous knowledge about the task to be performed. Experiments done with the WP technique have been shown to have a good degree of diagnosticity in terms of the nature of task demands (Rubio et al., 2004). Its sensitivity to task demands is as good as other measurement techniques, as well as its validity of performance and reliability. The work

of Tsang and Velazquez suggests that mental workload is a multi-dimensional construct and subjects, who performed their WP questionnaire, were able to rate task demands on separate mental workload dimensions (Tsang and Velazquez, 1996). Computational properties of the WP instrument and the aggregation of the dimensions considered in it are examined in section 2.4 (page 32). Appendix A.3 shows the questionnaire associated with the procedure.

Copper-Harper scale

The Copper-Harper Scale (CH) is a unidimensional rating scale that uses a structure in the form of a decision-tree for assessments of mental workload. It is based upon the concept of performance, and subjects are guided along the tree, through questions that lead to an overall workload value in the scale 0 to 10 (Cooper and Harper, 1969). Although applied in aviation, its tree-based structure is simple, with high operator acceptance. However its format lacks diagnostic capacity, similar to other unidimensional measures (Cooper and Harper, 1969). Appendix A.4 shows the questionnaire associated with the procedure along with its functioning.

Rating Scale Mental Effort

The Rating Scale Mental Effort (RSME) is another unidimensional procedure, developed in the Netherlands, which also assesses mental workload. This procedure considers the exerted subject's effort, and subjective ratings are indicated across a continuous line, within the interval 0 to 150 with ticks each 10 units. Labels such as 'absolutely no effort', 'considerable effort' and 'extreme effort' are used along the line. The final mental workload of a subject is related to the exerted effort indicated on the line by the subject, from the origin of the scale (zero). Although the procedure is relatively simple and quick, it shows a good degree of sensitivity. However, on the other hand, it has demonstrated to be a poor diagnostic capacity (Zijlstra, 1993). For details about the scale, its history, and development, we refer the reader to (Zijlstra, 1993).

Subjective Workload Dominance technique

The Subjective Workload Dominance Technique (SWORD) is also a unidimensional technique and is aimed at assessing mental workload of different tasks using relative subjective judgements that are compared against each other (Vidulich and Ward Frederic G., 1991). An operator is required to provide ratings using a structured evaluation form in which multiple tasks are listed, and compared in terms of workload imposed. Specifically, each of the possible paired combinations of the selected tasks to be compared, is presented on a row with the descriptions of the two tasks in the opposite extremities. Along the row, workload descriptors are used to indicate the level of imposed workload with a label 'equal' in the centre, indicating that the two tasks induced the same level of mental workload. On the other hand, other rating expressed by the subject indicates a workload dominance of a task over the other. All the ratings are then organised into a judgement matrix, which is further checked for consistency and aimed at providing the overall mental workload value imposed by the tasks, on an operator. For further information, the reader is referred to (Vidulich and Ward Frederic G., 1991).

Bedford scale

The BS is another unidimensional rating scale conceived for identifying the spare mental capacity of an operator whilst performing a task (Roscoe and Ellis, 1990). One dimension is used to estimate mental workload, and it is based upon a hierarchical decision tree aimed at guiding a subject through the rating scale. This is composed by ten points, each of them labelled by a descriptor and with a numerical associated level of mental workload (from 1 to 10). The assessment procedure is relatively simple, quick to be executed, and easy to apply straight after the completion of a task. It is suitable for assessing mental task load in environments characterised by high mental workload. However, it does not have diagnostic capability, limiting its use. For a detailed description of the Bedford Scale the reader is referred to (Roscoe and Ellis, 1990). Appendix A.5 shows the scale associated with the procedure.

2.3.2 Performance measures

Mental workload practitioners and, more generally, system designers, are typically concerned with the performance of their systems and technologies. The assumption is that the mental workload of an operator, interacting with a system, acquires importance only if it influences system performance. As a consequence, it is believed that performance-based techniques are the most valuable options for designers (Tsang and Vidulich, 2006). According to various reviews (Cain, 2007; O' Donnel and Eggemeier, 1986; Tsang and Vidulich, 2006; Wickens and Hollands, 1999; Wilson and Eggemeier, 2006; Young and Stanton, 2004) performance measures can be summarised using two categories:

- primary task measures;
- secondary task measures.

In primary-task methods, the performance of the operator is monitored and analysed according to changes in primary-task demands. In secondary-task assessments procedures, there are two tasks involved and the secondary task performance might not have a practical use, rather it servers to measure the operator's mental workload during the primary task. A further class of measures, less well-known but classifiable under the secondary-task approaches class, is the reference task measure, in which reference tasks are performed before or after the primary task. In the following paragraphs, these two categories are further described using theoretical explanations and practical applications. Potentially, in the field of human-computer interaction, performance measures are promising techniques to gather objective information of users' behaviour related to their interaction with a computer. Performance measures are relevant to this thesis because they are adopted, along with subjective measures, to evaluate the designed defeasible framework.

Primary-task performance measures

Primary-task performance assessment methodologies are built on a simple assumption: as task demands increase, the performance on the task or the design option of interest are expected to decrease. This hypothetical relation is due because of the limited information-processing capacity of humans to handle and deal with the task demands. Example of commons measures are Response and Reaction Time (RT), accuracy and Errors Rate (ER), speed and signal detection performance, Estimation Time (ET) and Tapping

Regularity (TR) (Tsang and Vidulich, 2006). There is not a prevalent measure and, outside laboratory settings, primary-task performance is intrinsically task specific (De Waard, 1996). However, since these measures directly reflects the outcome of the effort exerted by an operator interacting with a system, they are frequently used as mental workload assessments techniques (O' Donnel and Eggemeier, 1986).

According to O'Donnell and Eggemeier (O' Donnel and Eggemeier, 1986), the performance during a primary-task is an index of the success of the interaction between human and machine. However, there are issues associated with this statement. For example, if one operator is be able to deal with an additional task, while another is not able to, this difference in performance can not be determined. This issue is supported also by Gopher et al (Gopher and Donchin, 1986) who suggested that mental workload is not the unique influencer of performance. Also, measures of direct performance do not reflect changes in the investment of resources. For example, the degree of performance of two individuals during the same task might be the same, however, as they can have different skills, their experienced workload may differ. As a consequence, their spare capacity for executing a further task would also be different. Assessment methodologies that incorporate a person's attitude, personality and skills have been demonstrated to have a higher predictive accuracy (Xie and Salvendy, 2000a,b). Another example is that, in modern technologies, if a system interface is poorly designed, or the data presented is qualitatively weak, performance could be limited. As a consequence, a subject does not need to try hard to understand that the system is poorly designed, limiting its comprehension and interaction. This suggests that, despite the fact that primary-task performance measures are extremely important to system designers and evaluators, individually they do not provide an accurate metric of the mental workload of an operator. This view is also supported in (Gopher and Donchin, 1986). In addition, the performance during a primary task might be not diagnostic of the source of mental workload or simply, it might not be available (Tsang and Vidulich, 2006). We refer the reader to (O' Donnel and Eggemeier, 1986) for a detailed view of primary-task measures.

Secondary-task performance measures

Secondary-task methodologies is another frequently used category of procedures aimed at assessing mental workload. The main characteristic is that they require an operator to concurrently perform two, and sometimes, multiple tasks. However, the focus is on the primary task, and an assessment of the mental workload is derived from the performance of the operator on the secondary task (O' Donnel and Eggemeier, 1986). Applications designed around this category are aimed at measuring the operator's presumed spare capacity exerted to the primary task. This is inferred by analysing the performance on the secondary task, which is an index of the operator's spare capacity, during primary task execution, and that which can be used for the secondary task. The assumption is that since primary and secondary tasks would compete for the finite pool of information-processing resources, an alteration of demands of primary task should result in an alteration of performance of the secondary task, because more or less of those resources can be employed in the secondary task. The most frequently adopted techniques within this category include time estimation and interval production, memory-search, choice reaction-time and mental arithmetic tasks. For a detailed overview of secondary task-based measurement techniques, the reader is referred to (O' Donnel and Eggemeier, 1986) and (Eggemeier and Wilson, 1991).

According to O'Donnell and Eggemeier (O' Donnel and Eggemeier, 1986) this class of measures is more sensitive to discerning differential capacity expenditures than primary-task measures, having high diagnosticity of demands of the primary task. Tsang suggests how procedures based upon these measures are the prototypical mental workload assessments procedures (Tsang and Vidulich, 2006). In one respect, primary task measures can be used to assess performance, even if they are not considered indicators of mental workload. In another respect, the secondary-task paradigm is believed to be an assessment technique highly appropriate to describe the construct of human mental workload (Tsang and Vidulich, 2006). Although secondary-task performance measures are highly diagnostic with a high degree of sensitivity, they lack operator acceptance (Eggemeier and Wilson, 1991). It is argued that adding an extraneous secondary task to the environment under consideration, can actually change levels of workload, and also radically influence the processing of the primary task. This in turn, as mentioned in (Tsang and Vidulich, 2006), would be nothing more than an experimental artefact. For this reason, their application in the heterogeneous field of human-computer interaction seems not to be appropriate. In order to circumvent this problem, the so-called *embedded secondary-task* technique has been adopted, for instance, in (Shingledecker, 1983). This technique is based on the hypothesis that, to maximise operator acceptance and minimise task intrusion, the secondary task can be designed in a way that fully integrates it with the system in use. A natural part that occurs within the primary task can be used as a secondary task, however, their performance can be manipulated and collected individually. Usually the priority of the embedded secondary task is smaller than the priority associated with the primary task. As a consequence, the intrusion on the primary task is expected to be limited with a significant increment of operator acceptance (Eggemeier and Wilson, 1991).

A further technique employed in secondary-task approaches is known as *reference task*. This technique considers standard tasks that need to be executed both before and after the primary task under evaluation. These tasks are mainly employed to investigate whether the instruments have trend effects, and as they represent standard secondary tasks, the changes in performance on them can be employed as indexes of the mental workload on the primary task. If the reference task is built on physiological or subjective measures, then it is possible to infer the costs used by an operator to maintain performance on the primary task. In turn, this is useful for analysing the state of the operator, and whether it has been affected (Gopher, 1984). Although the secondary-task paradigm can be adopted to predict the mental workload on the primary-task, the main drawback is that the method entails experience and background knowledge, both to properly execute the evaluation of the secondary task and to interpret the outcomes. Additionally, the approach might also require further resources for developing the software and the hardware to be used in practical experimentations (Tsang and Vidulich, 2006), therefore limiting its application in real-life human-computer interactive settings.

2.3.3 Physiological measures

Several physiological measures of bodily responses, derived from the physiology of an operator, have been adopted for assessing human mental workload with the assumption that they correlate with it. They are aimed at interpreting psychological processes by analysing their influence on the state of the body and not by

measuring perceptual subjective ratings or task performance. The principal reason for adopting physiological measures is because they are not based upon an overt response by the operator, and the majority of cognitive tasks do not require this type of behaviour[4]. These measures can be collected continuously, within an interval of time, representing an objective way of measuring the operatorÕs state. Generally, physiological measures tend to be systematic indicators of stress, and have extremely high sensitivity to certain aspects of mental workload, above all for situations of underload (O' Donnel and Eggemeier, 1986). However, this high sensitive capacity might represent a misleading indicator of mental workload; in this case, an inappropriate technique is used (Lysaght et al., 1989, p. 137). The intrusiveness of the measures is low because they do not tend to influence the execution of the primary task. Alongside this, a common disadvantage is that they use specialised equipment, as well as trained operators, with technical expertise, to use this equipment. As a result, this has reduced the acceptance of physiological-based assessment procedures. However, the problem of invasive equipment is nowadays mitigated by the miniaturisation of tools and sensors available to researchers and practitioners that might favour their future application in emerging human-computer interactive systems.

From a design perspective, although physiological measures approaches might be sensitive to a number of advantageous applications, they are '... one conceptual step removed from the inference that the system designer would like to make' (Wickens and Hollands, 1999, ch. 6). In other words, the differences of workload provided by physiological-based techniques must be used to infer breakdowns in performance, or how the operator feels about the task performed. Physiological measurements may be extremely appealing in the case where performance measures, or subjective ratings, are not sensitive to covert changes in operator strategies, or when there is dissociation between them. Moreover, it is believed that they should only be applied if they have low intrusiveness, and are actually reliable (Fairclough, 1993), and associated with other measures of mental workload (Young and Stanton, 2006).

In summary, in one respect, most of the drawbacks related to physiological measurement techniques, from an operator's perspective, are technological, and are mainly concerned with improvements in the equipment provided, or in the methodology adopted. In another respect, from a designerÕs and analystÕs perspective, the main issue is the lack of a clear and robust link between physiological measures and performance (Kramer, 1991). According to O'Donnel and Eggemeier (O' Donnel and Eggemeier, 1986), physiological measures can be organised into four classes: brain, eye, cardiac and muscle functions. Although physiological-based measurement techniques have not been adopted for the evaluation of the defeasible framework proposed in this thesis, for the completeness of this mental workload review, the following paragraphs will briefly describe these categories. Here citations might not always be referring to state-of-the-art research; the objective is only to highlight their strengths and limitations.

[4]Overt behaviour refers to any type of behaviour that can be observable by others, as opposed to covert behaviour which cannot be observable by others. For instance, writing is an example of overt behaviour, while thinking is an example of covert behaviour.

Brain functions measures

The most attractive procedure employed in physiological assessments of human mental workload is probably the Electroencephalography (EEG). In this procedure the brain's activity is recorded using surface electrodes placed on the scalp of the operator, while performing a certain task. Various attempts to analyse and extract a potential index of workload from the various bands[5] (alpha, beta, etc.) of the EEG spectrum have shown to generally be inaccurate and imprecise, with a considerable variability and low reliability (O' Donnel and Eggemeier, 1986). In addition, the fact that they require trained operators to use specific equipment, result in them not really being suitable for easily assessing mental workload in modern human-computer interactive environments. For a detailed overview of these brain functions measures, the reader is referred to (O' Donnel and Eggemeier, 1986) and (Kramer, 1991).

Eye function measures

Since the eye is a substantial input channel of information to the human and is directly accessible for observation, it has gained importance in the assessment of mental workload. Several procedures have been developed for analysing eye movements and other parameters, as reviewed earlier in (Young and Sheena, 1975). In particular, the Electrooculography (EOG) [6] has been extensively adopted for studying eye movements with the assumption that it is an index of mental workload. EOG-based assessment techniques have been demonstrated to have low intrusiveness, higher operator acceptance and minimal implementation requirements, compared to other physiological measures (O' Donnel and Eggemeier, 1986). Pupillary response, in particular Pupil Dilation (PD), is another eye based technique. As stated in (O' Donnel and Eggemeier, 1986), in general, eye function-based procedures have the capacity of making fine distinctions between workload levels, thus showing high sensitivity. However, they have severe implementation requirements, and although commercial tools are available, it is very difficult to adopt them in applied environments, and non-laboratory settings. Here, different levels of luminosity or emotional effects can significantly influence the pupil response, thus invalidating their application for workload assessments in more natural, human-computer interactive environments.

Cardiac measures

The Electrocardiogram (EKG), and blood pressure, its volume as well as its oxygen concentration, have all been used as physiological measures of performance, stress and workload (Wilson and Schlegel, 2004). Cardiac rate has frequently been used since it can be obtained by employing non-invasive techniques, and easily administrative procedures. Empirical evidence suggests that Heart Rate Variability (HRV) observed in subjects at rest, may be an index of mental workload, with high sensitivity (Castor, 2003). However, as mentioned in (O' Donnel and Eggemeier, 1986) and (Wilson and Schlegel, 2004, section 4 page 7), the absolute heart rate is influenced by several subtle psychological processes, thus there is scepticism of its value as a workload measure. Blood pressure has been demonstrated to be highly correlated to mental demand, an attribute of workload. Blood Pressure Variability (BPV) is closely related to HRV; decrements in HRV will

[5] A band in Electroencephalogram procedures is a range of frequencies, typically expressed in hertz (HZ), of the EEG signal. For instance, Delta waves are up to 4 Hz, Theta waves are from 4 to 8 Hz, Alpha waves from 8 to 13 Hz and Beta waves are above 13 Hz.

[6] Electrooculography (EOG) is a technique used for the measurement of the resting potential of the retina and its output signal is called the electrooculogram. One of the main applications is eye movements.

be reflected in decrements in BPV. However, its sensitivity is not convincing (Castor, 2003). Respiration Rate Variability (RRV) has been considered as a promising measure which increases under conditions of stress and attention (Wilson and Eggemeier, 1991). In turn, increments in respiration rate correspond to increments in memory load or mental demands. However, the intrusiveness of the equipment for measuring respiration, as well as cardiac rate and blood pressure, is high, thus not suitable for easily assessing mental workload in more natural human-computer interactive systems. For further details on general methodologies and techniques based upon cardiac measures, the reader is referred to (O' Donnel and Eggemeier, 1986) and (Wilson and Schlegel, 2004).

Muscle measures

Human mental workload can also be assessed by monitoring the relative static tension of a muscle not directly involved in the execution of a task. For example, electrodes may be placed on a limb which is not being used in the task as well as on the neck or forehead. This measurement is usually implemented using Electromyography (EMG)[7]. As mentioned in (O' Donnel and Eggemeier, 1986), an increment in mental workload or stress corresponds to an increment in the EMG tension level. However, muscle measures have produced contradictory results, thus there is skepticism in its applicability as a predictor of mental workload (O' Donnel and Eggemeier, 1986). In addition, they require intrusive equipment mitigating their applicability in modern human-computer interactive systems.

2.3.4 Advantages and disadvantages of measurement techniques

Each typology of measurement technique, as just reviewed, has its own advantages and disadvantages for being more or less suitable for modern human-computer interactive environments HCI.

Subjective measures are in general easy to administer and analyse. They provide an index of perceived strain, and multi-dimensional measures can determine the source of mental workload. However, the main drawback is that they can only be administered post-task, meaning after task completion, and as a consequence, they can influence the reliability of long tasks. In addition, meta-cognitive limitations can diminish the accuracy of reporting, and it is difficult to perform comparisons among participants on an absolute scale. Despite this, they appear the most appropriate candidates for assessing mental workload in modern human-computer interactive environments, because they can be easily administered, and they have demonstrated high sensitivity and diagnosticity (Rubio et al., 2004).

Performance measures can be divided into *primary task* and *secondary task* measures. Primary task measures represent a direct index of performance and they are accurate in measuring long periods of mental workload. They are capable of discriminating between individual differences in resource competition. However, the main limitation is that they cannot distinguish between performance of multiple tasks; in this case these are executed in parallel by an operator. If taken in isolation, performance measures do not

[7] Electromyography is a technique for recording the electrical activity generated by skeletal muscles and it is performed using a tool called an electromyograph and produces a record called an electromyogram. The electromyograph is aimed at detecting the electrical potential produced by the cells of a muscle, when these cells are neurologically activated.

represent reliable measures, thus they are not really appropriate for real-world settings and conditions of modern human-computer interactive contexts. Although, if used in conjunction with other measures, such as subjective ratings, they can be useful. Secondary task measures, instead, have the capacity of discriminating between tasks when no differences are detected in primary performance. They are useful for quantifying the individual's spare attentional capacity, as well as short periods of workload. Nonetheless, they are only sensitive to large changes in mental workload and they might be highly intrusive, influencing the behaviours of users when they are carrying out the primary task. This intrusiveness represents the main disadvantages that make them unsuitable procedures to adopt for assessing mental workload, in more natural and modern human-computer interactive contexts.

Physiological measures are extremely good at monitoring data on a continuous interval, thus having high sensitivity in measurement. In addition, they do not interfere with the performance on the primary task. Yet, the main drawback is that they can be easily confounded by external and extraneous interference. Moreover, they require equipment and tools which are, most of the time, physically obtrusive, and the analysis of data is complex, requiring well trained experts. These disadvantages represent the main reasons for excluding physiological measure for assessing mental workload in modern and natural human-computer interactive systems, where the behaviour of operators need to be gathered as naturally as possible.

Having introduced the state-of-the-art measurement techniques for human mental workload assessment, highlighting their advantages and disadvantages, the goal is now to review the strategies employed by state-of-the-art models towards the computation of a numerical representative index of mental workload.

2.4 Aggregation strategies and computational aspects

Several mental workload assessment procedures have been described, which show how non-trivial the assessment problem itself is. For those uni-dimensional procedures, the problem of aggregating the attributes believed to influence mental workload, does not exist: the unique attribute considered here is believed to entirely represent mental workload. As a consequence, the numerical representation of this unique attribute is not important. Still, for multi-dimensional procedures, there is the issue of how to numerically represent multiple attributes, as well as how to aggregate them towards a representative meaningful index of mental workload.

In the NASA-TLX (Hart, 2006), for example, subjective ratings are expressed as natural numbers within the range 0 to 100 ($[0..100] \in \aleph$), while in the SWAT (Reid and Nygren, 1988) model, as natural numbers within the discrete range 1 to 3 ($[1..3] \in \aleph$). These ranges and scales are commonly adopted but they do not represent the only choice for expressing an operator subjective judgement. For example, Moray (Neville et al., 1988) proposed the use of Fuzzy Sets (FS)[8], borrowed from Fuzzy Set Theory (FST) (Zadeh, 1965) as

[8] Fuzzy sets are sets containing elements that have degrees of membership. In classical set theory, the membership of an element in a set is assessed in binary terms, meaning it can either belong, or not belong, to the set. Contrarily, fuzzy set theory allows the assessment of the membership of an element in a set in a gradual way. This gradual membership is described with the use of a membership function, bounded in the real unit interval from 0 to 1 ($[0..1] \in \Re$).

a means for humans to express judgements in a qualitative way but at the same time precisely, formalising the use of verbal judgements. In (Longo and Barrett, 2010a) and (Longo and Barrett, 2010b) the authors attempted a proposal of an ad-hoc formalisation of various attributes believed to influence mental workload. For example, the dimension of 'cognitive ability', as well as the dimension of 'context bias' are modelled as functions requiring three parameters and returning a value in $[0..1] \in \Re$. The concept of 'arousal' is designed as a taxonomy of sub-factors organised as a unidirectional tree where leaf nodes represent subjective judgements, and internal nodes indicate aggregation clusters. Unidirectional weighted edges link child nodes to parent nodes, towards a root node which represents the final value indicating the final degree of arousal. All the values are bounded in the range $[0..1] \in \Re$. The dimension of 'intention' is described as a single value in $[-1..1] \in \Re$ while the dimension of 'perceived difficulty' and 'time pressure' are bounded in $[0..1] \in \Re$. The use of different scales and methods evidently complicates the aggregation of attributes towards a representative index of mental workload. In the last 4 decades, several computational aggregation strategies have emerged. In the following sections these are reviewed, highlighting their advantages and their limitations.

2.4.1 Simple aggregation

In the Workload Profile (WP) assessment procedure, (as described in section 2.3.1, page 24), the accounted workload dimensions are based upon the multiple resource theory proposed by Wickens (Wickens, 2008; Wickens and Hollands, 1999) (as described in section 2.1.2, page 12). Each dimension is quantified through subjective rates (as in appendix A.3, page 175). Subjects, after task completion, are required to rate the proportion of attentional resources used for performing a given task, answering each question with a value within the discrete range 0 to 1 ($[0..1] \in \Re$). A rating of 0 means that the task placed no demand on the dimension being rated, while a rating of 1 indicates that the task required maximum attention on that dimension (Tsang and Velazquez, 1996).

The aggregation strategy employed in the WP method is relatively simple as it only sums each of the 8 rates provided by a subject for each workload dimension d:

$$MWL_{WP} : [0..8] \in \Re$$

$$MWL_{WP} = \sum_{i=1}^{8} d_i$$

Although, this aggregation method is extremely simple, it implies that each dimension has the same strength in affecting overall mental workload. Additionally, it does not consider external factors affecting the execution of the task, nor the state of the operator and his/her previous knowledge of the task being executed.

2.4.2 Weighted aggregation and preferences

In the NASA Task Load Index (NASA-TLX) instrument (section 2.3.1, page 23), the combination of the factors believed to influence mental workload is not based on a simple sum, rather, on a weighted average. Each factor is quantified through a subjective judgement formally bounded in $[0..1] \in \aleph$ whose weight is computed via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial

coefficient) of the 6 attributes (see appendix A.1, 173), *'which of the two contributed more to their workload during the task'*, such as 'Mental or Physical Demand?', 'Physical Demand or Performance?', and so forth, giving a total of 15 preferences.

$$\binom{6}{2} = \frac{6!}{2!(6-2)!} = 15$$

The weights w are the number of preferences, for each dimension, in the 15 answer set. In other words, the number of times that each dimension was selected. In this case, the range is from 0 (not relevant) to 5 (more important than any other attribute). Eventually, the final human mental workload score (MWL) is computed as a weighted average, considering the subjective rating of each attribute d_i (for the 6 dimensions) and the correspondingcorrespondent weights w_i.

$$MWL_{NasaTLX} : [0..1] \in \Re$$

$$MWL_{NasaTLX} = \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15}$$

The main issue associated with this aggregation approach is that, in the case where a new dimension has to be added, the paired comparison procedure will be more tedious, as it requires more judgements by the subjects. With only 9 or 10 dimensions, the comparisons required are respectively 36 and 45 which can be too burdensome for an operator to perform. This issue has been acknowledged by various authors who have proposed a modified version of the NASA-TLX. These do not require the paired comparison procedure, thus reducing the aggregation strategy to a simple average of the 6 dimensions (Thomas E., 1991). For further and detailed explanations about the Nasa Task Load Index (NASA-TLX) and its modified versions, the reader is referred to (Thomas E., 1991) and (Hart, 2006).

2.4.3 Ranking-based and correlation-based aggregation

In the Subjective Workload Assessment Technique (SWAT), as described in section 2.3.1, page 24) three workload attributes: (time, effort and stress), are modelled using discrete numbers from 1 to 3 ($[1..3] \in \aleph$). Each number has an associated description (as in table A.2, appendix A, 174). A pre-task procedure requires subjects to rank 27 cards, yielded from the combinations of the three dimensions at the three discrete levels, beginning with the card representing the lowest workload, and ending with the card representing the highest workload. The main reason for executing the card sort procedure is to build data useful to producing a scaling solution which is tailored to the perception of workload by the group of subjects, or an individual. This step is very important as it differentiates SWAT from other subjective assessment techniques. The subsequent step, called prototyping, analyses the sorted card data in order to determine the degree of agreement among the participants (raters), for a certain experiment on a given task. In this step, the Kendall's coefficient of concordance (W) is employed: a non-parametric statistic[9] used for assessing agreement among raters. Kendall's W ranges from 0 (no agreement) to 1 (complete agreement). Assuming that card i is given the rank $R_{i,j}$ by the subject number j, where there are in total n cards (27 in the SWAT model) and m subjects, then the

[9] A non-parametric statistic is a statistical method in which the data is not required to follow a normal distribution, or any other particular probability distribution.

total rank given to card i is:

$$R_i = \sum_{j=1}^{m} r_{i,j}$$

and the mean value of these total ranks is

$$\overline{R} = \frac{1}{2} m(n+1)$$

The sum of squared deviations, S, is defined as:

$$S = \sum_{i=1}^{m} (R_i - \overline{R})^2$$

and then Kendall's W is defined as:

$$W = \frac{12S}{m^2(n^3 - n)}$$

If the statistic W is 1, then all the subjects (raters) have been unanimous, meaning each respondent has assigned the same order to the list of cards. If W is 0, then there is no overall trend of agreement among the subjects. Intermediate values of W indicate a greater or lesser degree of unanimity. In the SWAT procedure, a single scale is developed by averaging data if $W > 0.75$. However, depending on the typology of the study being conducted, scales for individual subjects can be developed, in the case where individual differences have to be considered her, for example when $W < 0.75$, homogeneous subgroup scales can be developed. In the original SWAT (Reid and Nygren, 1988), the authors developed six hypothetical orderings, based on the relative importance of each attribute, as depicted in table A.3 (appendix A, page 174). For instance, TES is the ordering in which the greatest emphasis is on T (time), the second greatest is on E (effort) and the third on S (stress). The same principle is applied for TSE, ETS, EST, STE, as well as SET weighting schemes. The subsequent step is devoted to the application of the Spearman correlation coefficient [10] between the sorting provided by the subject and the hypothetical ordering. This is aimed at deciding which of the six subgroups is more suitable, that means which group a subjects belongs to. For instance, a subject that correlates to the SET group, can be considered a stress subject.

In the SWAT assessment technique, once the number of groups has been determined, a conjoint analysis is performed in order to generate a final workload scale bounded between 0 and 100 ($[0..100] \in \aleph$). Specifically, after task completion, a subject is required to rate the three dimensions, providing for each one a natural value from 1 to 3, thus generating a 3-tuple. Each combination in the rank associated with the subject in the previous phase, had to previously be rescaled in the range 0 to 100 ($[0..100] \in \aleph$), indicating the corresponding mental workload level for each combination. Finally, the last step consists of extracting the workload value associated with the 3-tuple in the rank associated with the subject that matches the 3-tuple provided by the subject, after task completion. Formally the mental workload provided by SWAT is:

$$MWL_{SWAT} : [0..100] \in \aleph$$

[10] In statistics, the Spearman's rank correlation coefficient (ρ), is a non-parametric measure of statistical dependence between two variables. It is aimed at assessing how good the relationship between two variables can be described, using a monotonic function. The Spearman correlation coefficient is similar to the Pearson correlation coefficient but it is applied between ranked variables. A correlation of $+1$ or -1 indicates that each variable is described by a perfect monotone function of the other, positive or negative. Values tending to 0 indicate a non-correlation between the two variables being measured.

It is worth noting that although it has been demonstrated that the SWAT procedure has high diagnosticity and content validity (Rubio et al., 2004) (Vidulich and Tsang, 1986), it relies on a very burdensome and tedious procedure for subjects to obtain the workload ratings. Similarly, it might not be straightforward to understand even by different mental workload designers. For further and more detailed explanations, and for the rationale behind the design choices of the SWAT, the reader is referred to (Reid and Nygren, 1988).

2.4.4 Ad-hoc aggregations and frameworks

The growing use of human-computer interactive systems has caused an increasing interest in the development of interfaces, these being the direct contact between a computer and an operator. In order to minimise the complexity of these systems, adaptive strategies have been proposed to improve the efficiency of the human-computer interaction. Hancock and Chignell employed the construct of mental workload as a means for investigating the capability of operators interacting with machine through interfaces (Hancock and Chignell, 1988). Their theoretical formulation of mental workload includes the ideas of the skill of the operators, and the time pressure they are exposed to, as well as the effort exerted for the execution of the task. Psychology being their main research field, the authors were inspired by the proposal of a computational model that included a power function to represent and assess mental workload, formalism widely applied for fitting psychological data. Their approximation of overall workload may be described by the following formula:

$$MWL_{HC} = \frac{1}{et^{s-1}}$$

where MWL_{HC} (Hancock and Chignell) is the overall workload level, e is the effort exerted by an individual operator, t is the actual time available for action and s indicates the operator's degree of skill. The issues associated with this formulation of workload, are various. Firstly, as also agreed with by the authors (Hancock and Chignell, 1988), the use of the function does not solve the problem of workload assessment since the degree of effort (e), skill (s) and temporal constraint (t) should be quantified and scaled using the same data range. Secondly, the formalism is not extensible: it is difficult to be expanded if further factors are considered. Finally, it does not account for the potential interactions that might occur between workload factors and their theoretical relationships. In a more recent work, it has been argued that human mental workload could be better defined in a framework consisting of multiple-indexes of workload, instead of a single-index such as overall mental workload (Xie and Salvendy, 2000b). In their work, Xie and Salvendy have introduced theoretical indexes:

- instantaneous workload;
- peak workload;
- average workload;
- accumulated workload;
- overall workload.

Figure 2.5 describes the meaning of each index. *Instantaneous workload* is aimed at measuring the dynamics of workload which have been defined as a dynamic and not a static process (Rouse et al., 1993). In the figure, the curve is an indication of the instantaneous workload versus time. The majority of physiological measures are examples of instantaneous indexes of mental workload. Subjective measurement and

performance measurement techniques do not usually consider instantaneous workload, since an overall index is usually computed after the completion of the task under examination. Nonetheless, some studies have shown that it is possible to measure instantaneous workload even when using these measurement techniques, in particular short-period workload (Verwey and Veltman, 1996). By employing the idea of instantaneous workload, it is possible to measure the mental workload at any given time during the task execution. This is the basic and most important index from which the other indexes can be derived.

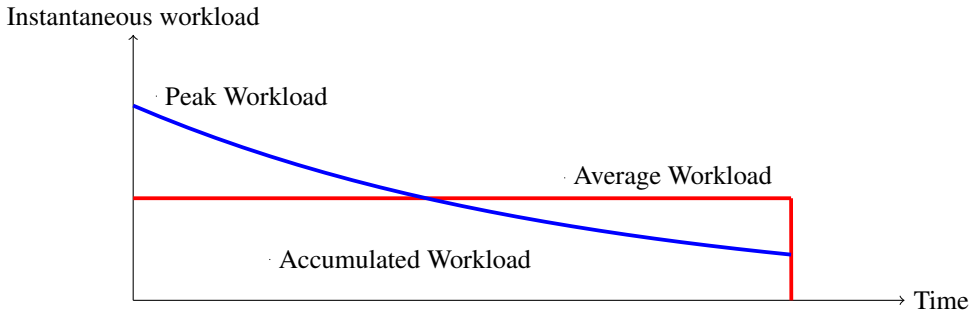


Fig. 2.5: Attributes of mental workload in the framework of (Xie and Salvendy, 2000b)

The *peak workload* is the maximum degree of instantaneous workload identified during the execution of the task. This value can be obtained by comparisons of all the instantaneous workload values. If this value exceeds the assumed maximum mental workload limit, as defined by the *redline* threshold (section 2.2.1, page 17), the operator may be affected by a consequence on performance which starts degrading. The *accumulated workload* is a measure of the overall amount of mental workload experienced by the operator during task execution. In figure 2.5 this is represented by the area below the instantaneous-workload curve. The *average workload* is a measure of the intensity of workload, and it is the average of the instantaneous workload values that coincide with the accumulated workload per unit time. Intuitively, a limit is assumed for the average workload, and if the latter exceeds the former, performance suffers. Since mental workload is related to the duration of a task, both the accumulated workload and the average workload are needed. Their combination allows to accurately measure the workload of both long-term and short-term tasks. Eventually the *overall mental workload* can be derived from the previous indexes and it describes the individual's experience of mental workload. Xie and Salvendy suggest that the overall workload coincides with the instantaneous workload or the accumulated and average workload in the brain of the operator (Xie and Salvendy, 2000b). The relationship between instantaneous and overall workload can be described by a mapping function f_1. Similarly, the relationship between the average, accumulated workload and the overall workload can be represented by a further mapping function f_2. In the case where the time interval of the task is fixed, accumulated and average workload should be proportional to overall workload. These relationships among the indexes of mental workload, are depicted in figure 2.6, and are formally defined as:

$$W_{peak} = Max\{W_{inst}(t)\}$$

$$W_{acc}(t) = \int_0^t W_{inst}(u)\, du$$

$$W_{avg}(t) = \frac{1}{t}\, W_{acc}(t)$$

$$MWL_{tot}^{XS} = f_1[W_{inst}(t)] = f_2[W_{acc}(t), W_{avg}(t)]$$

where t indicates time, W_{inst} is the instantaneous workload, W_{peak} indicates the peak workload, W_{acc} is the accumulated workload, W_{avg} indicates the average workload and finally MWL_{tot}^{XS} (Xie and Salvendy) is the overall workload. f_1 and f_2 are the mapping functions and they depend both upon the given task, and the particular individual.

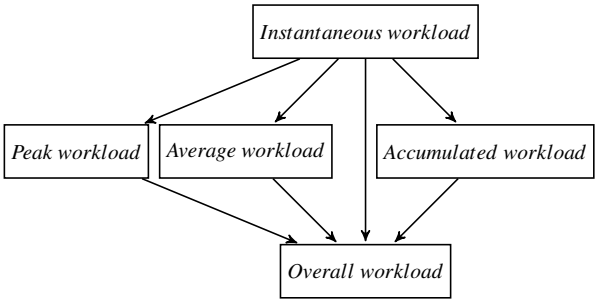


Fig. 2.6: Relationships among the indexes of workload within the framework of (Xie and Salvendy, 2000b)

Xie and Salvendy agree that the incorporation and the consideration of individual factors in the assessment of mental workload plays an important role. As a consequence they proposed an extension of the aforementioned model in order to make its predictive capacity consistent with an individual's subjective experience of mental workload. People are not able to accomplish a given task with ful efficiency and they often make errors, or are distracted. This suggests that all the effort a subject exerts to accomplish the task, does not fully contribute to its fulfilment. For this reason mental workload has been divided into ineffective and effective parts. *Effective workload* contributes to the fulfilment of the task directly, and it coincides with the amount of workload people deliver when they execute the task correctly and most efficiently. *Ineffective workload* does not contribute positively to the fulfilment of the task, and it is part of the workload that people can significantly reduce through training and learning. As a consequence, this factor is individual-dependent, meaning two different subjects may generate different degrees of ineffective workload. According to these two new typologies of workload, accumulated workload and average workload could be updated as:

$$W_{acc} = W_{eff} + W_{ineff}$$
$$W_{avg} = \frac{W_{acc}}{T}$$

where T is the time available to perform the task, W_{eff} is the effective workload, and W_{ineff} is the ineffective workload. The approach assumes that people fully concentrate on the given task, but in reality this is often not the case. So even if a task is very difficult to be executed, a subject can still decide not to do anything to accomplish it, thus the workload might be null. In turn, this implies that there must be another reason that influences the overall workload. In their model, Xie and Salvendy related to this issue as a *degrading factor* (DF) which is a number bounded in the range 0 to 1 ($[0..1] \in \Re$) where 0 indicates the total lack of willingness to perform the task and 1 the full concentration devoted to it. Intermediate values represent partial degrees of concentration or full attention only applied to a certain part of the given task. According to this, the accumulated workload can be redefined as:

$$W_{acc} = DF \times (W_{eff} + W_{ineff})$$

where DF is the degrading factor. The implication behind the effective/ineffective classification, is the possibility to reduce mental workload and, as a consequence, the efficiency (EFC) can be enhanced by controlling those factors that yield ineffective workload.

$$EFC = \frac{W_{eff}}{W_{eff} + W_{ineff}}$$

Effective workload, as well as ineffective workload and the degrading factor can be influenced by other factors such as stress, fatigue, knowledge, task importance, motivation, attitude and task complexity, task uncertainty and task duration. These factors are domain-specific as well as user-specific and they are important for the implementation of a model for each particular case. Finally, in order to define a task in a precise way, the typology of the environment in which the task is executed should be taken into account. So far the model was suitable for single-task environments, but it could be easily extended for multi-task environments. Here, each task might not only influence the other tasks, but the operator has a further mental effort which is devoted to the management of the simultaneous tasks to be executed. This *management load* (*ML*) is needed to control the concurrent tasks, their scheduling and switching. The previous model applied in single-task environments can be defined, for multi-task environments, as:

$$W_{eff} = \sum_{i=1}^{n} W_{eff} \; for\, task\, i$$

$$W_{ineff} = \sum_{i=1}^{n} W_{ineff} \; for\, task\, i + ML$$

where n is the number of tasks simultaneously performed and MN is the management load. The final workload level is redefined as:

$$MWL^{XS}_{tot-multi-t} = \sum_{i=1}^{n} HMW^{XS}_{tot} \; for\, task\, i \; + EFC + ML$$

where $HMW^{XS}_{tot-multi-t}$ is the overall multi-task workload, EFC is the efficiency and ML is the management load. The implication behind this model is that the mental workload exerted during execution of simultaneous tasks always generates a higher degree of workload than the sum of the workloads of the same tasks performed individually. A practical experimentation of this model can be found in (Xie and Salvendy, 2000a).

Although Xie and Salvendy's proposal is an important step towards a better definition of mental workload (Xie and Salvendy, 2000b), it is a theoretical implementation for modelling it that needs to be validated empirically. In addition, although it is a more complete model for better describing mental workload, allowing a designer to embed in it user-specific, task-specific and context-aware factors, it does not take into consideration how to formally model each of these factors. As mentioned in (Xie and Salvendy, 2000b), another issue is the representation of the two mapping functions that are unknown in the literature, and they represent only a theoretical proposal. Additionally, their framework does not consider theoretical relationships and potential interactions between these factors believed to influence workload, and the inconsistencies that might emerge from their interactions.

2.5 Fields of application

In the last 4 decades, the concept of human mental workload has been applied in many different fields. Earlier applications were in transportation, in particular in the aviation and automobile industry. Use of MWL was subsequently extended to adaptive automation and manufacturing systems. In recent years, it has been increasingly employed in medicine and health-care. Eventually, with the proliferation of computer-based systems and web-based applications, that drove the work of humans more cognitive, the construct was also adopted in the field of human-computer interaction (HCI). In this section, several applications, based on MWL measurement, are briefly described. The main objective is to complete the review of mental workload, showing how the construct has been practically employed so far and used in different environments.

2.5.1 Transportation

One of the first fields of application of mental workload was transportation. In particular, in aviation, several studies have been carried out for testing the mental workload of pilots interacting with cockpit interfaces. One of the reasons for analysis in this area, is to avoid pilot overload, by performing interface structural changes or introducing automation. Similarly, as aircraft and other systems are becoming more automated, the problem is also to avoid underload situations in which a pilot's attention can be at minimal levels. Wierwille et al have evaluated 16 different measures of mental workload, in terms of sensitivity and intrusiveness, using a simulated flight task for meditational activity[11]. These measures included performance and physiological assessment techniques, as well as subjective ratings and opinions (Wierwille et al., 1985). Similar studies were aimed at evaluating HRV as a physiological index of mental workload in complex flight scenarios (Veltman and Gaillard, 1993) or a terminal radar approach control simulator through a simulated task with increasing difficulty employing the NASA-TLX, the RRV and the HRV (Brookings et al., 1996). A study, commissioned by the Swedish Air Force, was aimed at analysing the performance and effects of the complexity of the information provided to pilots through a head down display, on their mental workload, according to different tactical situations (Svensson et al., 1997). Applications of HRV, the NASA-TLX, the SWAT and the Bedford Scale (BS) showed how even a small increment in information complexity corresponded to a higher increment in mental workload, with negative effect on performance. Besides aviation, the automobile sector gained benefit from the application of the concept of MWL. Several studies were devoted to analysing the performance of drivers (De Waard, 1996) under various psychological states in order to enhance automotive safety. These employed physiological measures (Reimer and Mehler, 2011) and primary/secondary task measures jointly with gaze tracking (Zhang et al., 2004), or performance measures jointly with the SWAT subjective procedure (Baldauf et al., 2009). MWL has also been applied in the rail industry (Pickup et al., 2005), in this caseto improve the efficiency and safety of system performance (Macdonald, 1999), as well as the workload imposed on train control officers (Pretorius and Cilliers, 2012).

[11] Mediational activity include activities such as reasoning, logic, judgement as well as decision-making.

2.5.2 Critical environments

The construct of mental workload plays an important role in critical environments, such as Nuclear Power Plants (NPP). In this context, it was adopted to evaluate interfaces of simulators of shutdown tasks and alarm reset tasks, as well as human performance employing primary/secondary task measures, subjective rating procedures and physiological assessment techniques (Hwang et al., 2007; Jou et al., 2009a,b; Vitorio et al., 2012). Similarly, in military operations and tasks which are performed in extreme settings and environments such as oceanic and space exploration, and disaster search-and-rescue, measures of mental workload have been shown to be extremely useful. In these contexts, the use of robots allows personnel to perform tasks that were previously thought impossible or life-threatening. However, they might impose high levels of mental workload on the human operators who control them. Indexes of MWL can allow the detection of these situations of overload and can be used to optimise operator performance (Prewett et al., 2010).

2.5.3 Automation, adaptive and manufacturing systems

Complex systems, in some cases, incorporate automative and adaptive procedures aimed at moderating the balance of work between the machine and the human. In these systems, measures of workload can be adopted to trigger adaptive automation in order to decrease the cognitive burden on the operator and, as a consequence, to improve the performance both of the operator and the system itself (De Greef et al., 2009). For instance, in the work of Her and Hwang (Her and Hwang, 1980), the objective was to individuate the limits of a supervisory manufacturing control system by analysing the workload levels of operators, by employing a model based upon Queueing Theory (QT)[12] Measures of mental workload were adopted to inform designers about when to introduce automation during the use of agricultural sprayers (Dey and Mann, 2010).

2.5.4 Medicine and health-care

In the last few years, the evolution of paper-based to digital-based systems has negatively impacted the workload of health-care clinicians and practitioners that are now required to interact with new technologies, such as Electronic Health Records (EHR), during their daily activities. Unfortunately,they are mostly not computer experts, and the burden caused by this interaction might have negative consequences on their performance, leading to errors, and reducing patient care, as well as decreasing safety and human satisfaction (Byrne, 2011). For these reasons, the construct of mental workload has been employed to design better systems and interfaces in health-care and medicine. In (France et al., 2005; Gaba and Lee, 1990) and (Leedal and Smith, 2005), subjective ratings and primary-task measures were employed to assess workload of physicians whilst interacting with an integrated electronic whiteboard or anaesthetists working in operating theatres. The purpose was to test their spare mental-capacity in order to handle additional tasks. Measures of mental workload associated with clinicians respectively during simulated and clinical practice were employed with the aim of improving safety (Byrne et al., 2010; Davis et al., 2009), or improving the quality of care provided to patients (Bertram et al., 1990, 1992). Similarly, MWL was used for the evaluation and initial development of novel technologies, applied to training and performance of laparoscopic surgery (Carswell et al., 2005; Stefanidis et al.,

[12] Queueing theory is a mathematical study concerned with waiting queues. This theory assumes that a model is constructed for the prediction of queue lengths and waiting times.

2006), for the evaluation of the usability of an Electronic Health Record interface (EHR) (Longo and Kane, 2011), in enhancing medication administration processes (Kataoka et al., 2011) and perianaesthesia nursing (Young et al., 2008). Over time, as new technologies and computer-based systems have driven the work of humans to be more cognitive and less physical, situations of high mental workload accumulations have become more frequent, and if recovery from these do not occur, health problems such as burnout, chronic stress or even depression can occur. Regular assessment of mental workload might offer new ways of supporting and preventing such mental disorders and maintaining mental health (Cinaz et al., 2011).

2.5.5 Human-Computer interactive and web-based environments

Human mental workload, as reviewed so far, appears to be a key concept for analysing people's interaction with machines and new technological devices. Cook and Salvendy's approaches provide a methodology for designing computer-based jobs in industry, accommodating individual's preferences, such as enhancing the degree of job enrichment and mental workload, in order to increase job satisfaction (Cook and Salvendy, 1997). Chaouachi et al proposed a methodology based on EEG features for the assessment of mental workload in intelligent systems, with focused applications in educational contexts and user modelling (Chaouachi et al., 2011). Mental workload has been employed in the design of Brain-Computer Interface (BCI) systems [13] and to prevent fatigue and increase performance of both able-bodied and motor-disabled people during brain-computer interface training (Felton et al., 2012). With the advent of human-web interactive systems, information can be presented to end-users in the form of text, video, audio and other forms of multimedia, eliciting different cognitive modalities and resources, sometime overloading the limited capacity of the human information-processing system. If web-sites and web-interfaces are not optimally designed, the risk is that tasks performed over them can lead to situations of overload. In turn, this may affect objective and perceived usability as well as user satisfaction and engagement.

Mental workload and perception of usability were jointly employed to measure the cognitive obstacles imposed by a particular design (Albers, 2011; Kokini et al., 2012; Tracy and Albers, 2006; Tremoulet et al., 2009) and to enhance current usability design practices (Longo et al., 2012b). Gwizdka's studies were aimed at identifying those factors that influence mental workload of users during search tasks, for the identification of search system features and search task types that imposed increased levels of load on end-users (Gwizdka, 2009, 2010). Other studies in the World Wide Web (WWW) include the application of mental workload for investigating its effect on user satisfaction in eCommerce systems (Schmutz et al., 2009), in adaptive hypermedia systems (Schultheis and Jameson, 2004), learning (Berka et al., 2007) and multimedia environments (Wiebe et al., 2010).

[13] A BCI is a direct communication system between the brain and an external device, and it is designed for assisting or repairing human cognitive or sensory-motor functions. For example, a speller, based upon a BCI, enables the system to spell letters by using the only brain activity. This typology of interface offers the promise to provide functionality as well as independence to people affected by severe motor disabilities, because it allows them to directly interact with computers or assistive devices.

2.6 Discussion in modelling human mental workload

The construct of Human Mental Workload (MWL), as reviewed in the previous sections, is certainly complex and multi-dimensional, and its assessment is not a trivial issue. Several definitions have been proposed by various researchers, with different backgrounds and influences, and they appear to be intuitively appealing. However, most of the current state-of-the art definitions all lack to demonstrate quantitative validation, as well as widespread acceptance. Gopher and Donchin argued that mental workload can be regarded as a hypothetical construct[14] and not an intervening variable [15] (Gopher and Donchin, 1986). Not only is it difficult to define workload precisely, but also to measure and assess it. Several measurement techniques exist, which are mainly classified into performance, subjective or physiological measures. However, each of them considers a different pool of workload factors, sometimes influenced by the context of application, other times affected by the designer's background, knowledge and choices, or simply driven by intuition. In some contexts not all the factors believed to influence mental workload, can be quantified, in other contexts they can only be gathered partially and incompletely, thus introducing uncertainty. Yet, to further complicate this, each assessment technique aggregates attributes in a different way, using different scales, weights and ad-hoc computational methods. Some studies have also attempted to propose hybrid models combining subjective and objective metrics together. Regardless of the chosen mental workload methodology, designing a general theory aimed at putting measures into context, requires several and heterogeneous experiments (Wierwille, 1988). Workload attributes can be static or dynamic, reflecting the mental workload within an interval of time or at a single moment. In addition, these attributes might be related, and sometimes not independent of each other. These relationships can be theoretical, such as the one between demand and performance (2.4, page 17), or empirically demonstrated through experiments, such as the U-shaped relationship between arousal and performance, as depicted in figure 2.7 (page 44), derived by the Yerkes-Dodson's law and widely accepted (Yerkes and Dodson, 1908).

However, none of the present state-of-the-art assessment techniques include a way of handling these theoretical relationships among workload attributes and the inconsistencies that might emerge from their interaction. According to Annett, the validity of individual attributes and, more generally complex constructs such as mental workload, lies especially in their relationships with the other attributes of interest, in the context of a specific situation (Annett, 2002a). In other words, he suggested that the validity of measures, especially subjective ratings, in a given context, is essentially the determination of the relationships with other measures of interest. These may be either behavioural or physiological, subjective or objective, as well as the expression of intentions or opinions. A measure is rarely valid in isolation, but as a predictor of some other measures or observations. Intuitively, more workload attributes and their interaction should provide more insights than one single non-interactive attribute. On the other hand, if interaction of attributes is acknowledged by a given assessment technique, a method for resolving inconsistencies that might arise from their interaction is needed.

[14] A hypothetical construct is an explanatory variable that is not directly observable, and it differs from an intervening variable because it has properties and implications that have not been demonstrated in empirical terms.

[15] An intervening variable can be summarised by findings observed empirically.

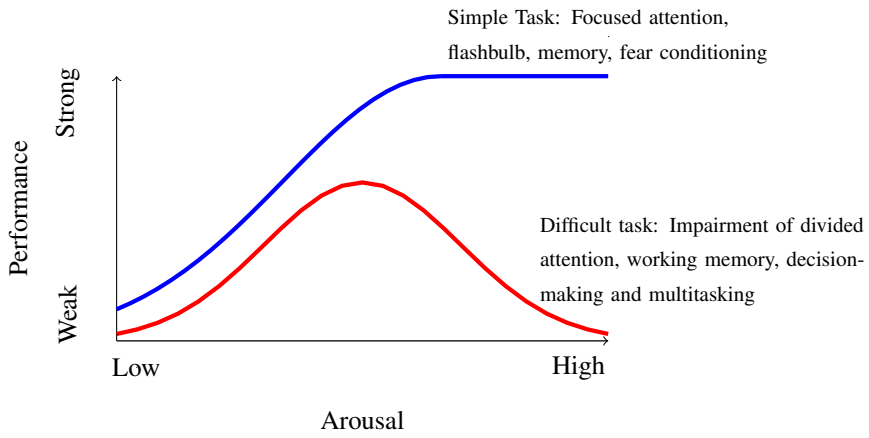


Fig. 2.7: Relationships between task difficulty, arousal and performance

2.6.1 Mental workload core tenets

To facilitate the understanding of the construct of mental workload and the issues associated with representing and modelling it as a computational concept, the core tenets found are summarised as follows.

1. *Multi-dimensionality*: empirical evidence and several researchers suggest that mental workload is believed to be a multi-dimensional construct influenced by many factors (Eggemeier and Wilson, 1991; Hart and Staveland, 1988; O' Donnel and Eggemeier, 1986; Reid and Nygren, 1988; Tsang and Velazquez, 1996; Wierwille and Eggemeier, 1993; Xie and Salvendy, 2000b; Young and Stanton, 2004). These factors have both static and dynamic properties. The former reflect mental workload within an interval of time while the latter within a single moment. Some factors can be related to a particular cognitive resource of the limited pool of expendable resources, of the human processing capacity. The requirement for individual resources can be unbalanced while performing a task. This means that some resources may remain unaffected, some overloaded and some underloaded (Xie and Salvendy, 2000b).
2. *Context-awareness*: mental workload is a context-aware construct, meaning that it is influenced by the context of application. For instance, it can be considered in a single or in a multi-task environment and, as a consequence, it might be affected by external factors, or by other concurrent tasks (Addie and Widyanti, 2011; Cain, 2007; Eggemeier and Wilson, 1991; Tsang and Vidulich, 2006; Xie and Salvendy, 2000a). The same assessment technique applied in two different contexts, can generate different workload assessments (Noyes and Bruneau, 2007).
3. *User-specificity*: workload is a user-specific construct, influenced by individual factors. These include, for instance, previous knowledge, skills and experience (Damos, 1988), but also the state of the individual (Xie and Salvendy, 2000b), as well as intention, motivation, effort manifested (Hancock, 1988), and subjective perception (Hancock, 1989; Meshkati and Loewenthal, 1988).
4. *Task-specificity*: mental workload is task-specific, influenced by task-related factors. These include, for example, task demands in terms of resources (Hart and Staveland, 1988; Tsang and Velazquez, 1996), objective task difficulty (Hancock, 1989), and its perception (Hancock, 1989; Hart and Staveland, 1988; Reid and Nygren, 1988).

5. *Relationality*: the factors considered within a workload assessment technique might be related monotonically or not, influencing each other and mitigating or enhancing each otherÕs strengths (Annett, 2002a,b) (examples of relationships in (Yerkes and Dodson, 1908) and (O' Donnel and Eggemeier, 1986)).
6. *Preferentiality*: a dimension considered within a workload assessment technique might be preferred than another dimension, thus having a greater influence on overall mental workload. Preferences can be modelled, for instance, by a ranked order between dimensions (Reid and Nygren, 1988) or with numerical weights (Hart and Staveland, 1988), either computed objectively or provided subjectively by raters.
7. *Subjectivity*: assessments of mental workload are characterised by a degree of subjectivity. This lies in the gathered measures, regardless if they are subjective ratings such as in (Hart and Staveland, 1988; Reid and Nygren, 1988; Tsang and Velazquez, 1996) or physiological such as in (Fairclough, 1993; Kramer et al., 1987; Wilson and Eggemeier, 1991). In addition, subjectivity refers to the design choices adopted for the development of an assessment technique, by a certain designer, thus outlining which factors to include and how to aggregate them (Annett, 2002b; Young and Stanton, 2002a).
8. *Uncertainty*: mental workload is a construct characterised by uncertainty. This is intrinsic in the concept and in its definition, but also in the values carried by each workload attribute. The former is due to the disagreement amongst researchers about how to define mental workload and how to measure it (Annett, 2002b; Cain, 2007). The latter refers to the accuracy of the measurement of each attribute. In particular, in the case of subjective measures, judgments are often made under uncertainty: often raters, not familiar with the concept of workload and associated factors, might have difficulty in even understanding the questions aimed at quantifying each dimension.
9. *Partiality*: the quantification of the factors employed within a workload assessment technique may be partial and incomplete. This mainly refers to objective measures (example is physiological), that can be gathered incompletely by employed devices and sensors. Partiality may also refer to the environment and context in which the assessment procedure is applied. The factors considered by a model might be correctly measured in laboratory settings, yet partially measured in practical settings, thus invalidating the theoretical model (Wierwille and Eggemeier, 1993) in the case it strictly requires them.
10. *Hypotheticality*: mental workload is believed to be a hypothetical construct. In other words, it cannot be detected directly but through the measurement and aggregation of other factors believed to have a high correlation with it (Gopher and Donchin, 1986; Xie and Salvendy, 2000b). In addition the relationship between these dimensions might be hypothetical/theoretical and not empirically demonstrated.

2.6.2 Mental workload as a defeasible phenomenon

Human mental workload is a complex and multi-dimensional construct built upon a network of pieces of evidence, as have emerged so far. This network can vary according to the knowledge-base of a workload designer considered in a given context. It is composed of the workload factors and their hypothetical or empirically-demonstrated relationships, assumed to be useful for predicting the mental workload of a user performing a given task, in a given context. Different workload factors, as suggested in the example in

section 1.3 (page 4), might support different and sometimes contradictory assessments of workload, creating inconsistent scenarios. In summary, it is reasonable to assume that:

- *Assumption 1*: human mental workload is a complex construct built over a network of pieces of evidence;
- *Assumption 2*: accounting and understanding the relationships of pieces of evidence as well as resolving the inconsistencies arising from their interaction is essential in modelling human mental workload.

In formal logics, these assumptions are the key components of a *defeasible concept*: a concept built upon a set of interactive pieces of evidence, called arguments, which can be defeated by additional arguments. The term 'defeasible' comes from the multi-disciplinary fields of defeasible reasoning (DR) and argumentation theory (AT), aimed at studying the way humans reason under uncertainty and with contradictory and incomplete knowledge. A reasoning process is defeasible when accounted arguments are rationally compelling but not deductively valid. In other words, DR is a form of reasoning built upon reasons that are defeasible, not infallible and a conclusion or claim, derived from the application of previous knowledge, can be retracted in the light of new evidence. DR is also known as non-monotonic reasoning (NMR), because of the technical property (non-monotonicity) of the logical formalisms that are aimed at modelling defeasible reasoning activity (Baroni et al., 1997). A formal implementation of DR is provided by AT, a recent multi-disciplinary topic in AI based on elements borrowed from psychology, philosophy and sociology. This topic investigates how people reason, and how arguments can be constructed, supported or neglected in a defeasible reasoning process. Additionally, AT studies the validity of the conclusions of a reasoning process via resolution of potential inconsistencies that might emerge from the interaction of arguments. Argumentation theory has been proved appealing for knowledge representation in various fields, thanks to its simplicity and modularity as compared to other approaches of reasoning, and has delivered an interesting explanatory capacity for tackling and describing complex constructs (Toni, 2010). These features seem to be appealing for creating a framework for mental workload representation and an assessment that matches the ideal requirements proposed in the previous section: flexibility, falsifiability, replicability, simplicity, inconsistency-awareness. As a consequence, these features have lead to the definition of the research question behind this thesis:

> *Can defeasible argumentation theory enhance the representation of the construct of mental workload, and improve the quality of its assessment in the field of human-computer interaction?*

As anticipated in the introductory chapter, a framework based on defeasible reasoning, and implemented with argumentation theory will be designed in order to represent and assess human mental workload, and tested in the field of human-computer interaction. However, before formally designing this framework, the next chapter is aimed at providing the reader with the basic principles of defeasible reasoning and non-monotonic logics, whilst describing state-of-the-art works in the field of argumentation theory, relevant to this thesis.

Chapter 3

Defeasible reasoning and argumentation theory

The aim of this chapter is to introduce the basic building blocks of defeasible reasoning and non-monotonic logics, notions that stand at the core of this thesis, through a review of state-of-the-art works in the field. Subsequently argumentation theory, based upon these notions, is introduced with particular emphasis on its role for knowledge representation. The starting assumptions behind this thesis is that the multi-dimensional construct of mental workload can be reasonably seen as a defeasible phenomenon, built upon a set of pieces of evidence, the arguments, that might interact and defeat each other creating inconsistent scenarios. State-of-the-art argument-based models are described highlighting how a single argument can be represented and the ways it can interact in terms of conflicts and defeats relations with other arguments. A meaning of the dialectical status of arguments is described, aimed at defining a non-monotonic notion of logical consequence for resolving potential inconsistent scenarios of conflicting arguments. The notions provided throughout this chapter are mainly formal definitions articulated with illustrative examples in relation to workload modelling.

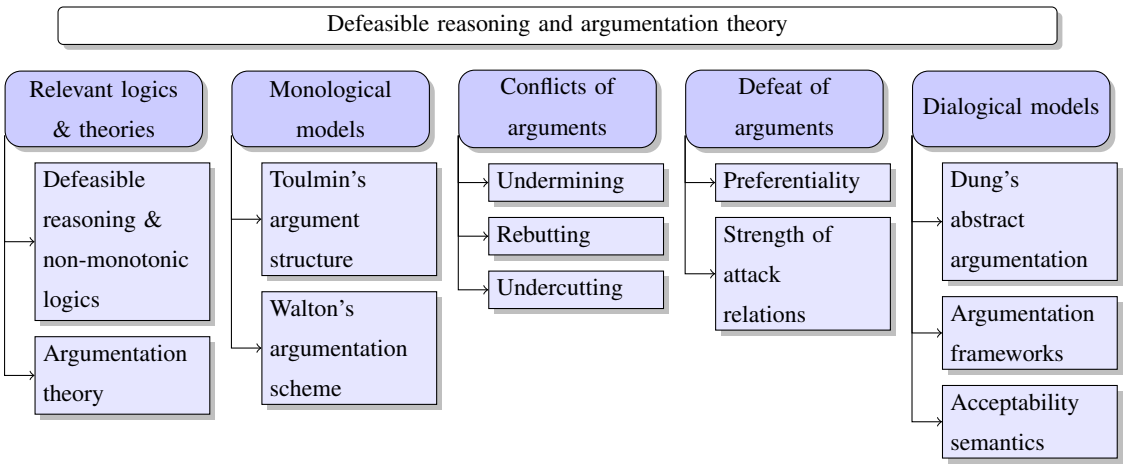


Fig. 3.1: Structure of the literature review of defeasible reasoning and argumentation theory

3.1 Relevant logics and theories

3.1.1 Defeasible reasoning and non-monotonic logics

The capability of delineating defeasible conclusions with partial information is an important aspect of any intelligent behaviour. In order to achieve such a capability, humans make use of a particular typology of knowledge, named *default knowledge*. The main property of this kind of knowledge, in a reasoning process, is that it can be exploited even if the preconditions used to its application are partial. In the case new information becomes available and the falsity of such preconditions can be deduced, then the conclusion derived from the application of the default knowledge can be retracted. This form of reasoning that employs default knowledge is referred to as *defeasible reasoning* (Baroni et al., 1997). In *default logic*, default knowledge is represented by using *defaults* that are specific inference rules. These are expressions of the form:

$$\frac{p(x) : j_1(x), ..., j_n(x)}{c(x)}$$

with $p(x)$ is the pre-requisite of the default, $j(x)$ is the justification and $c(x)$ is the consequent. If $p(x)$ is known and if $j(x)$ is consistent with what is known, then $c(x)$ can be possibly deduced. In other words, if it is believed that the pre-requisite is true, and each of the n consistency conditions (justifications) are consistent with current beliefs, this leads to believe that the conclusion is true.

Example 1

A classical example of a default is:

$$\frac{bird(Tweety) : fly(Tweety)}{fly(Tweety)}$$

that in natural language is: if *Tweety* is a bird and it is consistent with other available information to assume that *Tweety* flies (for example all birds fly), then it is assumed (inferred) that *Tweety* flies.

Default logic is a form of non-monotonic logic conceived to formalise reasoning with default assumptions. *Non-monotonic reasoning* is different from standard deductive reasoning: in the latter a conclusion follows from a set of true premises while in the former this is not always the case. To clarify this important property, consider again example 1: *Tweety is a bird* and *all birds fly*, so following a syllogistic reasoning, *Tweety flies*. If A is the set of premises and p is the conclusion, the deductive reasoning is:

$$if \;\; A \vdash p, \;\; then \;\; A, B \vdash p$$

If any additional set of information B is added to the set of evidence A, the conclusion p is still valid. This property is called *monotonicity* and conclusions do not change if new evidence is added to the existing set of premises, since the validity of the conclusions is all embedded in the premises. On the other hand, non-monotonic reasoning is not based on the monotonicity property and conclusions can be retracted when new evidence is available. Consider again example 1. If, in addition to the fact that *Tweety is a bird*, it is known that *Tweety is a penguin*, then the conclusion that *Tweety flies* can be retracted, as a special exception

raised. Non-monotonic logic relies on the idea that the pieces of knowledge employed in a reasoning activity such as *birds fly* may admit exceptions and that is impossible to include a detailed list of exceptions within the reasoning rules (Baroni et al., 1997). In these cases, the premise of a certain rule is only partially specified and a conclusion can be derived from the premises. In the case an exception subsequently arises then the derived conclusion has to be retracted. In other terms, the basic idea of non-monotonic inferences is that, when more information is obtained, some inference that were earlier reasonable may be no longer so.

The reasonable assumption behind this thesis, as emerged from the review of chapter 2, is that the representation of the construct of mental workload is a reasoning activity with the property of non-monotonicity. As a consequence, defeasible reasoning and non-monotonic logics seem to be plausible candidates for representing and modelling it. However, mental workload is a complex and ill-defined construct, subjectively considered by different researchers and practitioners of different fields. Its multi-dimensional and abstract nature make its definition, representation and assessment a non-trivial problem. For these reasons, mental workload designers call for a methodology able to improve the representation of such a multi-dimensional construct in a more intuitive way and at the same time having a precise framework for investigating its complexity. In the last two decades, a new field within artificial intelligence (AI), named argumentation theory (AT) has emerged as a useful paradigm for capturing and expressing the way humans reason in the form of formal arguments. The following sections are entirely devoted to the introduction of this new emerging field with a particular emphasis on its important role for Knowledge Representation (KR) and with a precise review of the state-of-the-art argument-based models proposed so far.

3.1.2 Argumentation theory

Argumentation theory has acquired importance in artificial intelligence and computer science, emerging as a multi-disciplinary approach that intersects the fields of law and philosophy, with aspects borrowed from psychology and sociology. It studies how pieces of evidence, seen as arguments, can be represented, supported or discarded in a defeasible reasoning process, and it investigates the validity of the conclusions achieved (Toni, 2010). Argumentation theory has gained interest with the introduction of computable models inspired by the way humans reason. These models extended classical reasoning approaches, based on deductive logic, that were increasingly inadequate for problems requiring non-monotonic reasoning, commonly adopted by humans, and explanatory reasoning, not available in standard non-monotonic logics such as default logic (Dung, 1995). Argumentation theory is different from standard deductive reasoning because it implements non-monotonic reasoning. In non-monotonic reasoning a conclusion can be retracted in the light of new evidence, whereas in standard deductive reasoning the set of conclusions always grows. Additionally, argumentation is a form of explanatory reasoning because it is built upon modular and intuitive steps that differ from the monolithic approach followed by many traditional logics for non-monotonic reasoning. Argumentation provides a useful paradigm for dealing with incomplete and possibly inconsistent information, and thus being fundamental for resolving conflicts and difference opinions of different parties or contradicting pieces of evidence or arguments. Furthermore, it is a powerful mechanism for explaining the outcomes generated automatically of an inference process. These features have enabled the introduction and the application of

argumentation theory in many fields, especially for inference, decision support, decision-making as well as dialogue, negotiation and practical reasoning (Bench-Capon and Dunne, 2007; Rahwan and McBurney, 2007; Toni, 2010). For example, thanks to the increasingly prominence of medical applications, early expert systems evolved towards more complex systems incorporating argumentative capabilities. As a consequence several works in health-care were proposed, such as applications devoted to support decision-making and to advise doctors on best specific medications or best treatment for breast cancers (Fox et al., 1993; Longo et al., 2012a). Argumentation has been used for conflict resolution in multi-agent systems with different and sometimes incomplete beliefs (Amgoud and Kaci, 2005) and also for trust computing (Matt et al., 2010; Parsons et al., 2010; Prade, 2007).

A general view of argumentation logics

In a simplistic view, argumentation focuses on interactions where different parties or different pieces of evidence argue for and against some conclusions (Matt et al., 2010). In argumentation theory arguments can be seen as 'tentative proofs for propositions' (Fox et al., 1993; Krause et al., 1995). Here knowledge is usually expressed in a logical language and its axioms correspond to premises according to the domain under consideration. Theorems in the chosen language are identical to claims, in the underlying domain, derivable from the premises by applying some rules of inference (Prakken and Vreeswijk, 2002). In general, the premises are not consistent because they might lead to contrary propositions. Arguments for propositions (claims) coincide with proofs, in a deductive logic but with the difference that the premises, on which these proofs are built upon, are not all known to be true. Viewing an argument as a tentative proof is related to the understanding of its internal structure. Various formalisms aimed at addressing the internal structure of arguments have been proposed in the literature, originated by the philosophical works of Toulmin (Toulmin, 1958). These models are mainly focused on the logical connection between the different elements of an argument and how a set of premises is linked to a conclusion in a monological structure. They are often referred to as *monological models* (Bentahar et al., 2010). A second branch of artificial intelligence, on the other hand, has investigated the relationships among arguments, sometimes not considering their internal structure, and treating them as abstract entities. These models emphasise the structure of arguments within a dialogical framework and thus they are often classified as *dialogical models*. Example of these models are proposed in (Dung, 1995) and in (Atkinson et al., 2006). In general, dialogical models are related to the process of arguing whereas monological models concerns the production and construction of arguments. The former consider the external, macro structure of arguments whereas the latter consider the internal, micro structure of arguments. In addition, dialogical models have driven argument-based approaches to be referred to as *defeasible reasoning systems* incorporating defeasible arguments: an argument is not a final absolute reason for the conclusion it supports, instead it is open to attacks by other arguments. In other words, a reasoning, in which a rule that supports a certain conclusion, might be defeated by new evidence, is called *defeasible* (Pollock, 1974, 1987). In the case defeasible reasons are connected and chained to reach a certain conclusion, arguments have place instead of proofs. As mentioned before, defeasible reasoning stands on the non-monotonicity property: conclusions previously drawn may be later withdrawn in the light of new information. Example of systems, based on non-monotonic reasoning, are the works of (Dung, 1995; Pollock, 1994; Prakken and Vreeswijk, 2002; Vreeswijk, 1993) and (Toni, 2008). A third classification of argument-based models have been proposed in which neither

the monological nor the dialogical structure is considered. These models are called *rhetorical models* and the rhetorical structure of arguments is stressed (Bentahar et al., 2010). The main characteristics of these models is the consideration of the audience's perception of arguments and they are aimed at investigating how arguments can be employed as a means of persuasion and not for establishing the truth of a conclusion. Example of these models are in (Grasso, 2002) and in (Pasquier et al., 2006). Table 3.1 summaries the three main categories of argumentation models.

Monological models	
Structure	Micro structure of arguments
Foundation	Arguments as tentative proofs
Linkage	Connecting a set of premises to a claim at the level of each argument

Dialogical models	
Structure	Macro structure of arguments
Foundation	Defeasible reasoning
Linkage	Connecting a set of arguments in a dialogical structure

Rhetorical models	
Structure	Rhetorical structure of arguments
Foundation	Audience's perception of arguments
Linkage	Connecting arguments in a persuasion structure

Table 3.1: Classification of argumentation models

In the literature of argumentation theory, models belonging to one category difficultly belong to the other categories. For instance, dialogical models do not address the internal representation of an argument and do not consider their perception by an audience. However, according to (Bentahar et al., 2010), in order to design and create intelligent systems that incorporate powerful argumentative capabilities, the micro-structure of an argument, its relation with other arguments as well as the rhetorical structure should be addressed. In other words, the internal representation of an argument should clearly relate premises to conclusions, and at an external levels, the argument should be considered within the set of other arguments it interacts with. Eventually, the perception by an audience is important because in real life implementations, arguments are built to achieve predefined objectives, according to the participating agents' believes. In the specific case of mental workload modelling, single arguments should be built around each individual factor believed to influence workload. The internal structure should clearly link one or more factors to a certain level of mental workload. Once each argument is internally built, it might be linked and related to some of the other arguments a designer is wiling to implement for shaping the construct of mental workload, creating a macro-structure. Eventually, when the macro-picture is conceived, the rhetorical structure should be addressed for producing convincing arguments for the audience. Indeed, if the micro-structure of an argument is not considered, it is difficult to generate meaningful and convincible arguments for an audience. Similarly, if the rhetorical structure is not addressed, the macro-structure cannot be efficient enough. Micro, macro and rhetorical structures are strongly related (Bentahar et al., 2010).

In summary, the general idea of argumentation systems is that they formalise non-monotonic reasoning as the internal construction of arguments (micro-structure) as well as their comparisons for and against certain conclusions (macro-structure). The construction of arguments, based on a theory, is monotonic that means an argument remains the same even if the theory is expanded with new information. On the other hand, non-monotonicity is expressed in terms of interaction between conflicting arguments. This is because the additional information may generate stronger arguments that in turn defeat previous arguments. Argumentation systems and the notion of an argument are typically constructed upon an *underlying logical* language and around an associated notion of *logical consequence*. As mentioned before, this notion of consequence is monotonic. New information can not invalidate existing arguments as constructed, but can only be responsible for the generation of new counterarguments. Some argument-based applications assume a particular and well-defined logic whereas other leave the underlying logic part of the context of application or even totally undefined. In the case the logic is left unspecified, the system can be instantiated with different alternative logics, thus they are often referred to as frameworks rather then systems. Beside the chosen underlying language, argumentation systems generally incorporate four elements:

- definition of an argument
- definition of conflicts among arguments
- definition of defeat relations among arguments
- definition of the dialectical status of arguments.

The literature of argumentation logics is really vast and it would be impossible to review all the research studies conducted for each element as well as all the applications appeared in the field of law, philosophy and computer science. For these reasons, the remaining of this chapter is devoted to the introduction of those essential concepts and formalisms necessary for orientating the reader towards the understanding of the rest of this thesis.

3.2 Internal structure of arguments and monological models

The internal representation of arguments is addressed by monological models which do not take into consideration the relationships that might exist between other arguments. In logical proofs, a conclusion follows from a set of premises and many argumentation systems do not make any distinction between them. However, when arguments are expressed in natural language, premises might play different roles, and their understanding can make the argument itself more understandable. In addition, the identification of the role of each premise can support the investigation of the different ways an argument can be accepted or defeated. In the literature of argumentation theory, this way of structuring and representing an argument is defined as *argument scheme*.

3.2.1 Toulmin's argument structure

Probably, one of the most widely argument scheme adopted in artificial intelligence is represented by the model proposed by Toulmin (Toulmin, 1958). Conceived in the context of law and philosophy, Toulmin introduced a diagrammatic representation of legal arguments as a theoretical model of argumentation composed by six distinct elements, as depicted in figure 3.2:

1. *Claim (C)* - an assertion or a claim (conclusion) that has a potentially controversial nature and that might not meet the initial beliefs of the audience;
2. *Data (D)* - statements specifying facts or beliefs previously established related to a certain situation in which the claim is made;
3. *Warrant (W)* - statement that justifies the inference of the conclusion from the data;
4. *Backing (B)* - a set of information that assure the trustworthiness of a warrant. It is the ground underlying the reason. A backing is invoked as soon as the warrant is challenged;
5. *Qualifier (Q)* - a statement that expresses the degree of certainty associated to the claim;
6. *Rebuttal (R)* - a statement introducing a situation in which the conclusion might be defeated.

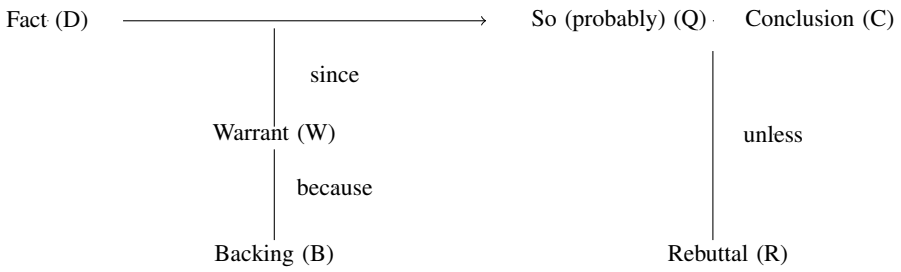


Fig. 3.2: An illustration of Toulmin's argument representation

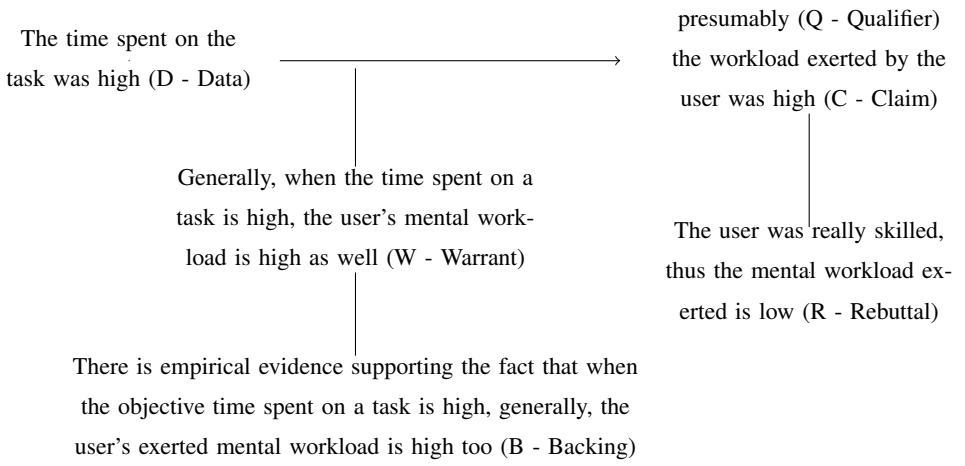


Fig. 3.3: An illustrative argument for mental workload using the Toulmin's structure

Figure 3.3 illustrates an example using the Toulmin's way of representing an argument in the context of mental workload. This argument claims that the mental workload exerted by a user during the execution of a task was high (conclusion) because execution time was high as well. This conclusion (claim) is presumably true since, in general, when execution time is high the user's mental workload is high as well (warrant). In turn, this is due because there is empirical evidence supporting the fact that when the objective time spent on a task is high, generally the user's exerted mental workload is high too (backing). However, exceptions to this rule might exist such as when the user is really skilled thus likely to exert low mental workload (rebuttal).

The Tulmin's model plays a significant role in highlighting the elements that might form a natural language argument. However, in real world scenario, arguments might not consider all these elements, making the argument itself weaker and less expressive. Toulmin's structure has been extended in several works and applied in various ways such as for facilitating the construction of text-based arguments resulting from a dialogue or for dialectical and legal reasoning (Prakken, 2005). Here, counterarguments can occur: these are also arguments that may defeat (attack) any of the first four elements of another argument (claim, data, warrant, backing). For instance, a debate can be visualised by chaining diagrams of arguments (Bentahar et al., 2010). Once the single representation of an argument is extended with the consideration of its relations with other external arguments, its monological structure is expanded towards a dialogical structure. The Toulmin's contribution has several advantages but also some limitations. Firstly, as mentioned before, it explicitly considers the various components of an argument and how they are linked, providing a very useful means for knowledge representation. Moreover, arguments represent the inference rules used to infer a conclusion from a set of premises. The main disadvantage of the Toulmin's representation is that it does not specify the way different argument structures can be combined for illustrating the dynamics of an argumentation process. It is uniquely aimed at emphasising the internal representation without accounting the participants and their knowledge-bases as well as not specifying any criteria for accepting an argument.

3.2.2 Walton's argument scheme

Another well-known monological paradigm has been proposed by Reed and Walton to model the notion of arguments as product (Reed and Walton, 2003; Walton, 1996). It is based upon the notion of *argumentation scheme* and it is useful for identifying and evaluating common and different types of argumentation in every-day discourses, having its strength in representing knowledge used for arguing and explaining (Bentahar et al., 2010). This type of argumentation scheme is aimed at capturing common stereotypical patterns of reasoning that are non-monotonic and defeasible in nature. In order to understand their way of structuring an argument, consider example 2, taken from (Reed and Walton, 2003).

Example 2

Suppose that Bob and Helen are discussing about tipping, and that Helen is not in favour of tipping because she thinks that it is a bad practice and it should be discontinued. Also suppose Dr. Phil is an expert in psychology. Her argument is: *Dr. Phil says that tipping lowers self-esteem.*

From example 2 it appears that Helen's argument is implicitly *an appeal to expert opinion*. In addition, it is evidently an instance of *argument from consequences*. Helen is sustaining that lowering self-esteem is a bad consequence of an action. The argument is based upon the assumption that, since the bad outcome is a consequence of tipping, therefore tipping itself is a bad thing. This way of reasoning is a chain of argumentation reconstructable as follows:

- Dr. Phil, an expert psychologist, says that tipping lowers self-esteem, because he has knowledge about self-esteem.
- tipping lowers self-esteem
- lowering self-esteem is a bad thing
- anything that leads to bad consequences is a bad practice
- tipping is a bad practice

Argumentation schemes can be used to link premises each other and to conclusions in a chain that represents the argument provided by Helen. Walton identified and proposed 25 different argumentation schemes that can be used to construct valid and robust arguments. For instance, as mentioned before, in example 2 two argumentation schemes are taken into account and they can be generalised, in line with (Walton, 1996) as follows:

Argument from expert opinion:

- major premise: source E is an expert in subject S containing proposition A
- minor premise: E asserts that proposition A (in the domain S) is true (or false)
- conclusion: A may plausibly be considered true (or false)

Argument from consequences:

- major premise: if an argument leads to good (or bad) consequences, it should (should not) be brought about
- minor premise: if action A is brought about, a good (or bad) consequence will occur
- conclusion: therefore A should (should not) be brought about.

These two schemes can be used by Helen to build her point of view. The argument can be built by firstly populating the implicit premises that are necessary to support the requirement of the appeal to expert opinion, and subsequently to give a reason to backup the conclusion that an action should not be taken. Each scheme proposed by (Walton, 1996) comes with a set of *critical questions* such as 'is the expert E in the position to know about the proposition A?'. Such questions have to be answered to assess whether their application is warranted or not in a specific context and case. Intuitively, the possibility to use critical questions makes argument schemes defeasible as open to counterarguments. The main advantage of the paradigm proposed by Reed and Walton is the capacity to illustrate an argument's structure using real cases as examples. They represent the inference process and the defeasible rules by using critical questions and they consider various criteria for accepting an argument related to the nature of the schema. However, as the Toulmin's model,

it only emphasises the individual structure of arguments, without considering the participant's knowledge-bases. The interaction with other argumentation schemes as well as with other arguments is not specified. In summary, in argumentation logics, the notion of argument coincides with a *tentative proof* in the underlying logic. Sometimes arguments are defined as inference trees that are grounded in the premises, and some other times as deduction that means as consequence of such inferences (Prakken, 2011). Simpler systems might be constructed with arguments being premises-conclusion pairs, and being the underlying logic left to implicitly validate a proof of the conclusion from the premises. Other monological paradigms have been proposed in the literature of argumentation theory: the reader can refer to (Bentahar et al., 2010) for further information about monological models of arguments. The notions provided so far were aimed at identifying the possible ways arguments can be internally constructed. However, in the reminder of this thesis, the Toulmin's and the Walton's approaches are left in favour of a simpler representation of an argument in the form of premises-conclusion using the notion of logical consequence.

3.3 Conflicts between arguments

Monological models, aimed at internally represent an argument can be complemented by dialogical models, focused on the relationships among arguments. The latter investigates the issue of invalid arguments that appear to be valid (fallacious arguments). *Conflict* is an important notion in argumentation theory, often replaced by the terms *attack* or *counterargument*. In the literature of AT three types of conflict have emerged (Prakken, 2011): undermining, rebutting and undercutting attacks. In general, any typology of argument can be attacked on their premises, however just defeasible arguments can be attacked on their inference link or on their conclusion. The reason that does not allow deductive arguments to be rebutted or undercut is that they, by definition, have a deductive inference that is truth-preserving, thus the truth of their conclusion is guaranteed by the truth of their premises. The only way to disagree with them is to deny one of its premises. In contrast, the conclusions of defeasible arguments might be rejected even if all its premises are accepted.

3.3.1 Undermining attack

The first typology of attack is referred to as *undermining attack* (figure 3.4): an argument can be attacked on one of its premises by another argument having a conclusion that negates that premise. This can be observed in the extended version of example 1 where an argument 'Tweety flies because it is a bird' can be attacked by another argument 'Tweety is not a bird'.

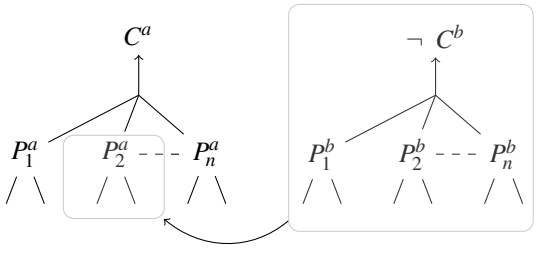


Fig. 3.4: Undermining attack

3.3.2 Rebutting attack

The second type of attack is called *rebutting attack* (figure 3.5) and it happens when an argument negates the conclusions of another arguments. For instance, 'Tweety flies because it is a bird' can be negated by 'Tweety does not fly because it is a penguin'.

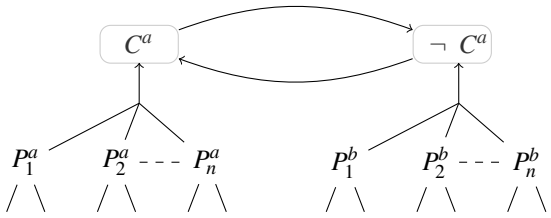


Fig. 3.5: Rebutting attack

3.3.3 Undercutting attack

The third type of attack (figure 3.6) occurs when an argument uses a defeasible inference rule and it can be therefore attacked on its inference by arguing that there is a special case that does not allow the application of the defeasible inference rule. After the important contribution of Pollock (Pollock, 1974, 1987) this type of attack is referred to as *undercutting attack*. In contrast to a rebutting attack, an undercutting attack does not negate the conclusion of its target argument, rather it argues that the target's conclusions is not supported by its premises and, as a consequence, cannot be drawn.

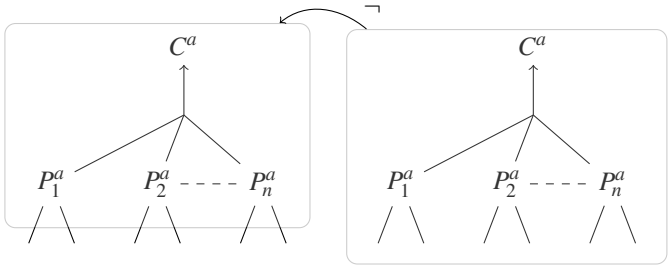


Fig. 3.6: Undercutting attack

3.4 Defeat between arguments

Conflicting arguments is an important characteristics of argumentation systems, however it does not embody any approach for evaluating an attack. The determination of the success of an attack from one argument to its target is another important aspect of argument-based systems. In the literature of argumentation theory, an attack has generally a form of a binary relation between two arguments and usually terms such as 'defeat' or 'interference' are interchangeably used to indicate a proper successful attack. Some work distinguish a defeat relation in a weak form (attacking another argument and not weaker) or in a strong form (attacking another argument and stronger). The former is generally referred to as 'defeat' whereas the latter as 'strict

defeat' (Prakken, 2011). Defeat relations are determined in various ways, depending on the argumentation system. Often, they are influenced by the domain of application and are usually defeasible. For instance, in those domains where observations are important, defeat relations might depend on the reliability of tests as well as on observers. In consultancy, defeats might be influenced by the level of expertise of consultants, whereas in legal applications, legal hierarchies among statutes, level of authorities or moral values all might have a different role in determining defeat relations. Evaluating an attack can occur through the notion of preferentiality or introducing the concept of strength of an attack relation. These two types of evaluation have been emerged in the literature of argumentation theory and as they are relevant to the rest of this thesis, they are described in the following sections.

3.4.1 Preferentiality between arguments

To establish whether an attack can be considered a successful attack (defeat relation between two arguments), a trend in argumentation theory is devoted to the consideration of the *strength* or arguments. In this respect a key concept is represented by the inequality of the strength of arguments that has to be accounted in the computation of extensions of arguments and counter arguments (Dunne et al., 2011). Several works have adopted the notion of *preferentiality* of arguments (Modgil, 2009). For example, in (Pollock, 1987) and (Prakken and Sartor, 1997), the authors formalised the role of preferences and if an arguments X undercuts another argument Y, then X is a successful attack (defeat) if Y is not stronger than X. Other approaches adopt preferentiality at a more abstract level. For instance, in the *Preference-based Argumentation Framework (PAF)* proposed by (Amgoud and Cayrol, 2002), an attack from X to Y is successful only if Y is not preferred to X. (Bench-Capon, 2003) proposed a *Value-based Argumentation Framework (VAF)* in which an attack from X to Y is successful only if the value promoted by X is ranked higher or equal than the valued promoted by Y, in accordance to a given ordering on values. Example 3 and figure 3.7 illustrates these various scenarios of preferentiality, given an attack set and the resulting defeat (successful attack) set.

The information necessary to decide whether an attack between arguments is successful is often assumed to be pre-specified and implemented as an ordering of values or a given partial preference. However, according to (Modgil, 2007) and (Modgil, 2009), the information related to preferentiality might be contradictory, as the preferences may vary depending on the context and on different subjects who can assign different strengths, to different arguments, employing different criteria. As a consequence, a subject has to argue and reason about defeasible and likely conflicting information about preferences. Motivated by this, Modgil has proposed the concept of *meta-level argument*, a special argument about preferences. An argument expressing preferences is a simple node in a graph of nodes, and preferentiality is abstractly defined, by creating a new attack relation that comes from a preference argument. This new attack relation defeats another attack relation between those arguments subject of the preference claim. A clarification is provided in example 4 and figure 3.8. Meta-level arguments allow no commitment regarding the definition of the preferences within the argumentation framework, rendering it simple as no preference list or strength need to be specified. The simple scenario of example 4 can be extended considering preference arguments underlying preferences that are contradictory, thus attacking each other, and also being open to be attacked by other preference arguments.

Example 3

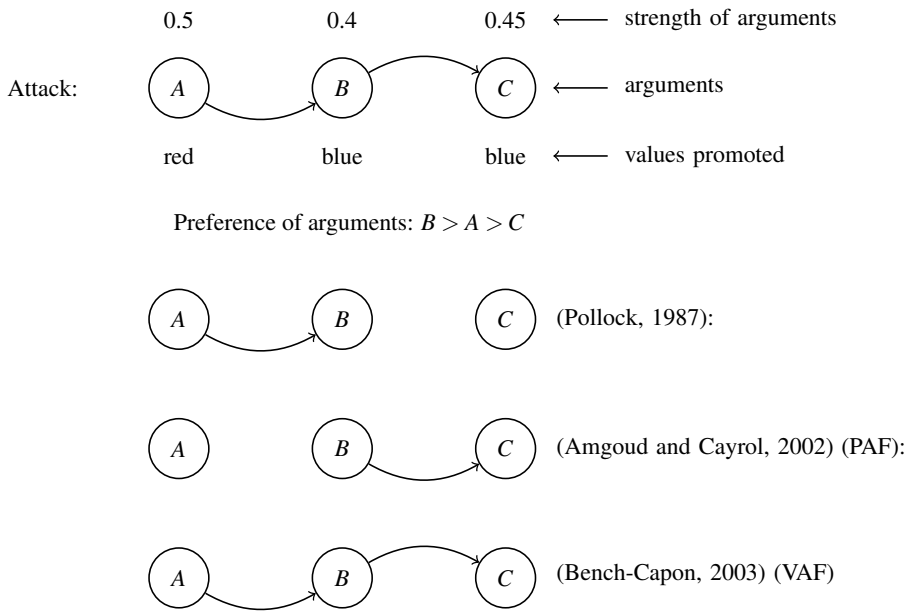


Fig. 3.7: Implementations of preferentiality between arguments

Example 4

Let us consider two arguments A, B, claiming two different contradicting conclusions, being subject of a rebutting symmetrical attack. Suppose the existence of a pre-defined preference list in which argument A is preferred to argument B (figure 3.8 - a). According to Modgil, this situation can be expressed as in figure 3.8 - (b) where another argument C is added, undercutting argument B.

(a) Standard preferentiality (b) Meta-level argument

Fig. 3.8: Standard preferentiality and meta-level arguments for expressing preferentiality

3.4.2 Strength of attack relations

Preferentiality, as reviewed so far, is implemented by assigning to arguments an importance value. This is usually pre-defined, in form of a full or partial priority list of available arguments, or in form of a numerical value attached to each of them, explicitly provided or implicitly derived from the strength of the rules used within the argument. In turn preferentiality allows to establish whether an attack can be considered successful, thus formalising a proper defeat relation, or considered a weak/false attack, thus being disregarded. As opposite to this approach, another branch of argumentation theory is devoted to associate weights to attack relations instead to arguments. In (Dunne et al., 2011) the authors investigated the role of adding weights on the attack links between arguments, introducing the notion of *inconsistency budget*. This quantifies the amount of inconsistency a designer of an argumentation system is willing to tolerate. With an inconsistency budget α, the designer is open to disregard attacks up to a total weight α. It turns out that, increasing this threshold, more solutions can be achieved progressively as less attack would be disregarded. As a consequence, this gives a preference order over solutions, and the solutions having a lower inconsistency budget are preferred. A similar recent approach that considers the strength of attacks is incorporated in (Martínez et al., 2008). In this proposal, referred to as Varied-Strength Attack Argumentation Framework (VSAAF), each attack relation is assigned a type, and the argumentation framework is equipped with a partial ordering over the types. Let us consider figure 3.9 where the type of attack from argument A to argument B is i and from argument B to argument C is j. Intuitively, depending on whether the type j is higher, lower or equally ranked than the type i, different ranges of solutions are possible.

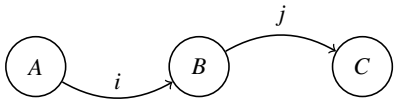


Fig. 3.9: Varied-strength attacks

The classical binary relation of attack has been extended in (Janssen et al., 2008) with the notion of *fuzzy relation* borrowed from Fuzzy Logic (FL). This approach allows the representation of the degree to which an argument attacks another one, creating a Fuzzy Argumentation Framework (FAF). Similarly, in (Kaci and Labreuche, 2010), a Fuzzy Preference-based Argumentation Framework (FAF) has been proposed in which the notion of fuzzy has been used to model the preference relation among arguments. In other terms, a value X attached to a preference relation between two arguments A, B corresponds to the degree of credibility by which A is strictly preferred to B. Strength of arguments and defeat relation has been considered also in (Li et al., 2011). Here, the authors assigned probabilities both to arguments and defeat, introducing the notion of Probabilistic Argumentation Framework (PRAF). Here probabilities refer to the likelihood of the existence of a specific argument or defeat relation, thus capturing the inherent uncertainties in the argumentation system. In PRAF all possible arguments neither definitely are disregarded nor they definitely exist: they have different chances of existing. Another interesting approach to assess the strength of a given argument has been investigated delivering appealing properties. Here, two fictitious people have to be confronted, endorsing respectively the roles of proponent and opponent of the argument. Such a situation of conflict between them can subsequently be analysed employing the paradigm of *game theory* (Matt and Toni, 2008).

3.5 The dialectical status of arguments and dialogical models

Defeat relations, as previously reviewed, are usually modelled as binary relations on the set of arguments. However, this form of relationship does not tell yet what arguments, within this set, can be seen as justifiable. Rather, it focuses on the relative strength of two individual arguments that are in conflict. The final ultimate status of each individual argument depends on the interaction with the other arguments in the set. As a consequence, a definition of the *dialectical status* of arguments, depending on their interaction, is needed. This last step of the argumentative schema is usually aimed at determining the outcome of an argumentation system and typically it splits the set of arguments in two classes: those arguments that allow a dispute to be won or be lost. Sometimes a third class contains those arguments that leave the dispute in an undecided status. The terminology used for the dialectical status of arguments varies and terms such as 'justified', 'defensible', 'defeated' or 'overruled' are usually adopted. The dialectical status of argument is investigated by dialogical models. Modern and current implementations of dialogical arguments-based systems are built upon the theory of (Dung, 1995). His work, historically speaking, derives from other more practical and concrete works on argumentation and defeasible reasoning, such as (Pollock, 1987, 1994; Vreeswijk, 1993). In the following sections, the Abstract Argumentation Theory (AAT) proposed by Dung is introduced and as it represents an important element for the reminder of this thesis, related notions are formalised and clarified with illustrations and examples.

3.5.1 Abstract Argumentation Theory

Dung's implementation of abstract argumentation was and is still a success due to the fact that it provides a way, applicable to all types of system that instantiate his framework, for assigning justification statuses to arguments. It is useful to mention that Dung-style argumentation approaches, contrary to, for instance, standard logic approaches, are not based on the notion of *truth*. These approaches formalise reasoning processes that are defeasible in nature and are not concerned with truth of propositions, rather they focus on accepting a proposition as true. Dung's frameworks allow comparisons among different systems by translating them into his abstract format (Vreeswijk, 1993). This property was a breakthrough because it showed how several logics for non-monotonic defeasible reasoning could be translated into his abstract framework.

The underlying idea that characterises abstract argumentation is that given a set of abstract arguments and a set of defeat (attack) relations between them, a decision to determine which arguments can ultimately be accepted has to be taken. Solely looking at an argument's defeaters to decide the acceptability status of an argument is not enough: it is also important to investigate whether the defeaters are defeated themselves. Generally, an argument B *defeats* an argument A if and only if B is a reason against A. If the internal structure of arguments and the reasons that lead to the definition of the defeat relation between them are not considered, an *Argumentation Framework (AF)* takes place (Dung, 1995).

3.5.2 Argumentation framework

An *argumentation framework* is a set of (abstract) arguments and binary attack (defeat) relations among these arguments. It is a directed graph in which arguments are presented as nodes and the attacks as arrows (figure 3.10).

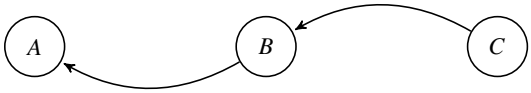


Fig. 3.10: Argument and reinstatement

Definition 1 (Argumentation Framework)
An argumentation framework is a pair

$$AF =< Ar, Attacks >$$

where Ar is a set of arguments and $Attacks \subseteq Ar \times Ar$. *A Attacks B iff (A, B)* $\in Attacks$. ■

Given an abstract argumentation framework, the issue is to decide which arguments should ultimately be accepted. The notion of attack of Dung is equivalent to the notion of defeat (section 3.4, page 57) because all the attacks in an argumentation framework are implicitly considered proper defeats, and they do not need to be evaluated. Therefore, for abstract argumentation frameworks, $AF =< Ar, Attacks >$ and $AF =< Ar, Defeats >$ are equivalent definitions. In other words, in abstract argumentation frameworks, there is no need to evaluate whether an attack is valid or not. In figure 3.10, A is attacked by B, and apparently A should not be accepted since it has a counterargument. However, B is itself attacked by C that is not attacked by anything, thus C should be accepted. But if C is accepted, then B is ultimately rejected and does not form a reason against A anymore. Therefore A should be accepted as well. In this scenario it is said that C *reinstates* A. This issue is referred to as *reinstatement* and in order to determine which arguments of an AF can be accepted, a formal criterion is necessary. In argumentation theory, this criterion is known as *semantics* (acceptability semantics), and given an AF, it specifies zero or more *extensions* (sets of acceptable arguments) (Baroni et al., 2011).

3.5.3 Acceptability semantics

Various argument-based semantics have been proposed (Baroni et al., 2011; Baroni and Giacomin, 2009; Caminada et al., 2012; Dung et al., 2007), but for the reminder of this thesis, the focus is on complete, grounded and preferred semantics as proposed in (Dung, 1995). The issue of argument semantics is clarified using the labelling approach by (Wu et al., 2010), as it follows.

Each Argument is either *in*, *out* or *undec* according to the following conditions:

- an argument is labelled *in* if and only if all its defeaters are labelled *out*, and
- an argument is labelled *out* if and only if it has at least one defeater labelled *in*.

Informally speaking, an argument labelled *in* means that is has been accepted, *out* means rejected and *undec* if it cannot be neither accepted nor rejected.

Definition 2 (Complete labelling)
Let $< Ar, Defeats >$ *be an argumentation framework and* $Lab : Ar \rightarrow \{in, out, undec\}$ *be a total function. Lab is a complete labelling iff it holds:*

- *if* $Lab(A) = in$, $\forall B \in Ar : (B\ Defeats\ A \supset Lab(B) = out)$
- *if* $Lab(A) = out$, $\exists B \in Ar : (B\ Defeats\ A \wedge Lab(B) = in)$
- *if* $Lab(A) = undec$, $\neg\forall B \in Ar : (B\ Defeats\ A \supset Lab(B) = out)$ *and* $\neg\exists B \in Ar : (B\ DefeatsA \wedge lab(B) = in)$.

■

Example 5

A concrete version of the argumentation framework of figure 3.10 could concern a reasoning process to predict mental workload using three arguments:

- A: the user was really skilled in the given task, so there is a reason to believe the workload was low.
- B: the user took long time to complete the task therefore there is a reason to believe the workload was high;
- C: the user was interrupted several times during the execution of the task, thus long execution time is no longer a reason to believe workload was high;

This example can be interpreted as follows. For argument C it holds that all its defeaters are labelled *out* (trivial as C is not defeated by any argument), thus C has to be labelled *in*. B has now a defeater labelled *in* thus it has to be labelled *out*. For A, it holds that all its defeaters are labelled *out*, so it has to be labelled *in*. As a consequence the resulting status of each argument is: $Lab(A) =$*in*, $Lab(C) =$ *in* and $Lab(B) =$ *out*. Informally speaking, arguments A and C can be accepted, instead argument B has to be rejected. This means that in the absence of further evidence, the mental workload was likely low.

Definition 3 (Abbreviations)
Let $< Ar, Defeats >$ *be an argumentation framework and let* $A \in Ar$ *and* $Args \subseteq Ar$. *For abbreviation:*

- A^+ *as* $\{B | A\ Defeats\ B\}$
- $Args^+$ *as* $\{B | A\ Defeats\ B\ for\ some\ A \in Args\}$
- A^- *as* $\{B | B\ Defeats\ A\}$
- $Args^-$ *as* $\{B | B\ Defeats\ A\ for\ some\ A \in Args\}$

■

A^+ indicates the arguments defeated by A, A^- indicates the arguments that defeat A. $Args^+$ refers to the arguments that are defeated by the set of arguments $Args$ while $Args^-$ refers to the arguments that defeat the set of arguments $Args$.

A set of arguments is called *conflict-free* if and only if it does not contain any argument A and B such that A defeats B. A set of arguments $Args$ is said to *defend* an argument C if and only if each defeater of C is defeated by an argument in $Args$.

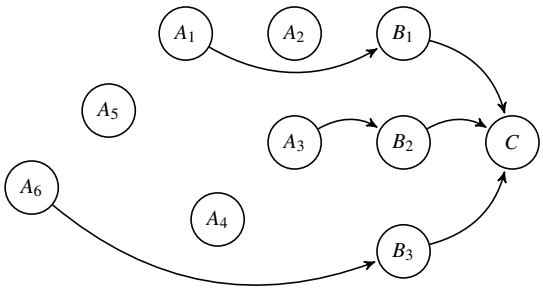


Fig. 3.11: The set of arguments $Args = [A_1, ..., A_6]$ defends argument C

Definition 4 (Conflict-freeness)
Let $< Ar, Defeats >$ be an argumentation framework and let $Args \subseteq Ar$. Args is said to be conflict-free iff $Args \cap Args^{+} = \emptyset$. ■

Definition 5 (Defence)
Let $< Ar, Defeats >$ be an argumentation framework and let $Args \subseteq Ar$ and $B \in Ar$. Args is said to defend B iff $B^{-} \subseteq Args^{+}$. ■

Definition 6 (Defence of arguments)
Let $< Ar, Defeats >$ be an argumentation framework and let $Args \subseteq Ar$ and $B \in Ar$. A function F is introduced

$$F : 2^{Ar} \rightarrow 2^{Ar}$$

such that F(Args) = $\{A | A$ is defended by $Args\}$. ■

F yields the arguments defended by a given set of arguments. It specifies the set of arguments that are acceptable, in line with Dung's definitions (Dung, 1995).

The aforementioned notions are the building blocks of abstract argumentation theory (AAT) useful for the definition of acceptability semantics for the resolution of potential contradictions among arguments. The basic semantic, as proposed in (Dung, 1995) is referred to as *complete semantic* aimed at computing *complete extensions*.

Definition 7 (Complete extension)
Let $< Ar, Defeats >$ be an argumentation framework and let Args be a conflict-free set of Arguments. Args is said to be a complete extension *iff $Args = F(Args)$.* ■

Example 6

In the AF of figure 3.10 there is just one complete extension, $\{A, C\}$, which is conflict-free and defends exactly itself. Note $\{A, B, C\}$ is also a fixpoint of F, but not a complete extension as it is not conflict-free.

The idea behind complete extensions is that a complete labelling might be viewed as a subjective and reasonable point of view that a designer can consider with respect to which arguments are accepted, rejected or considered undecided. Each subjective point of view is internally coherent and if contested, the designer can use its own position to defend it. Although this position can be certainly questioned by someone, its internal inconsistency cannot be pointed out. The set of all the complete labelings coincides with all the possible and reasonable positions available to a designer (Wu et al., 2010). Complete semantics have an important property: more than one complete extension might exist. However, sometimes it is advantageous to consider a semantic that is guaranteed to generate exactly one extension: the *grounded semantic*. This is a skeptical approach a designer can take for the evaluation and acceptance of designed arguments.

Skeptical approach

The idea behind the grounded semantic is to select the complete labelling *Lab* in which the set of *in*-labelled arguments is minimal.

Definition 8 (Grounded Extension)
Let $< Ar, Defeats >$ be an argumentation framework. The grounded extension is the minimal fixpoint of F. ■

> **Note 1**
> The grounded extension coincides with the complete labelling in which *in* is minimised, *out* is minimised and *undec* is maximised.

Example 7

In the AF of figure 3.10, the grounded extension is $\{A, C\}$.

Grounded semantics are useful because they yield always one unique grounded extension (it can be the empty set). However, this skeptical view might be replaced by a more credulous approach, available to a designer, known under the name of *preferred semantic*.

Credulous approach

The idea behind preferred semantic is that, instead of maximising *undec* arguments, it maximises *in* arguments (and also *out*). They are based on the concept of admissibility. A set of arguments is admissible if and only if it is conflict-free and defends at least itself.

Definition 9 (Admissibility)
Let $< Ar, Defeats >$ be an argumentation framework and let $Args \subseteq Ar$. Args is said to be admissible iff Args is conflict-free and $Args \subseteq F(Args)$. ■

Definition 10 (Preferred extension)
Let $< Ar, Defeats >$ be an argumentation framework and $Args \subseteq Ar$. Args is said to be a preferred extension iff Args is a maximal admissible set. ■

> **Note 2**
> The empty set is admissible in every AF as it is conflict-free and trivially defends itself against each of its (none) defeaters. For any AF, there exist at least one preferred extension. Every grounded and every preferred extension is a complete extension.

Example 8

The admissible sets are $\{C\}$, $\{A,C\}$. $\{B\}$ and $\{A\}$ are not admissible as they do not defend themselves respectively against C and B. Only one preferred extension exists: $\{A,C\}$.

Grounded and preferred semantics represent respectively a skeptical and credulous approach for arguments acceptability. The two notions are sufficient for the reminder of this thesis and the reader is referred to (Baroni et al., 2011; Baroni and Giacomin, 2009) and (Dung et al., 2007) for further acceptability semantics. In the following section, the multi-layer argumentative schema for knowledge representation, as emerged so far, is summarised.

3.6 Summary

Argumentation theory (AT) has been proved useful for tackling many knowledge representation problems characterised by partial and incomplete knowledge-base as well as uncertainty and contradictions of the pieces of evidence in this knowledge-base (Toni, 2010). The theory is aimed at modelling *defeasible* reasoning, a form of reasoning with the property of non-monotonicity. *Non-monotonic* activity occurs when the conclusions, drawn with previous evidence, can be retracted in the light of new additional evidence: fallibility and corrigibility of conclusions are acknowledged. This is in contrast to monotonic activity in which conclusions do not change, even if evidence is added to the existing set of premises, because the validity of the conclusions is all embedded in the premises. Argumentation theory has acquired importance because, with the application of arguments, usually expressed as natural language propositions, knowledge-bases can be represented more intuitively. In turn, this modus-operandi has demonstrated that the theory leads to *explanatory reasoning*, increasing the understanding of the knowledge-base being modelled.

The process of argumentation towards the representation of a concept, a situation or a construct starts by the identification of an *underlying language*, driven by the context and the domain of application. Usually, the logic is left unspecified and the argumentative system under consideration is referred to as framework

because it can be instantiated by alternative and different logics. The knowledge representation process starts by the identification of the relevant pieces of evidence, the *arguments*. These arguments can be natural language propositions claiming something or more structured arguments using an underlying language such as first order logic. The internal structure of argument is addressed by *monological models* (top layer of figure 3.12). Usually they are built in the form of inference rules that link a set of premises to a conclusion, the claim. They can also be defined as inference trees that are grounded in the premises as well as deduction that means as consequence of such inferences. Monological models are aimed at internally represent an argument, and they can be complemented by *dialogical models* focused on the interaction among them and aimed at investigating the issue of fallacious arguments, invalid arguments that appear to be valid. These interactions are usually referred to as conflicts, attacks or counterarguments.

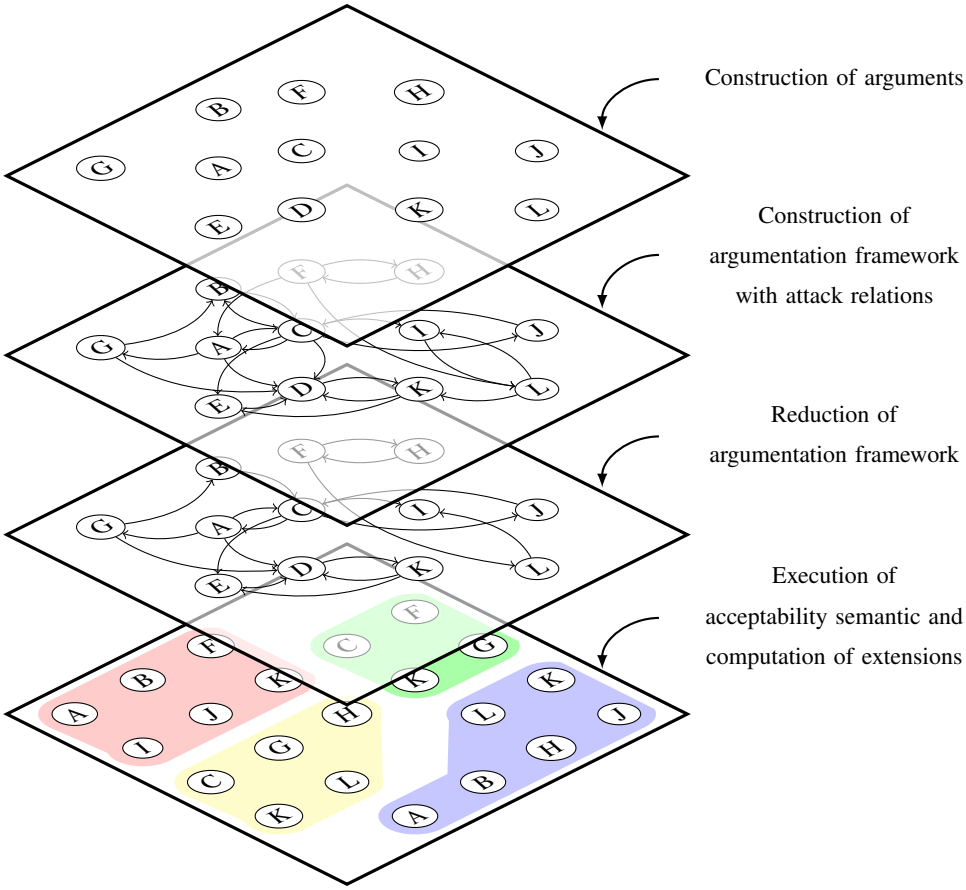


Fig. 3.12: The multi-layer argumentative schema for knowledge representation

Three typologies of conflict are possible:

- *undermining attack*: an argument can be attacked on one of its premises, by another argument having a conclusion that negates those premises;
- *rebutting attack*: an argument negating the conclusion of another argument;
- *undercutting attack*: an argument using a defeasible inference rule, thus open to challenge, can be attacked by another one arguing a special case exists that does not allow the application of that rule.

The second layer of the argumentative process for knowledge representation (figure 3.12) is focused on the definition of those conflicts among the arguments defined in the first layer, forming an *argumentation framework*. This level does not embodies yet any approach for evaluating defined attacks. The determination of the success of an attack from one argument to its target is evaluated in the third layer of figure 3.12. Here proper attacks are usually referred to as *defeats* and they are determined in various ways, depending on the argumentation system and the context of application. To establish whether an attack can be considered a proper defeat there is a trend in the literature of argumentation theory devoted to the consideration of the strength of arguments. The underlying idea is that not all the arguments are equal in strength. As a consequence the notion of *preferentiality* of arguments can be adopted. The simple form of implementation is a preference list of arguments thus if an arguments X undercuts another argument Y, then X is a successful attack (defeat) if Y is not stronger than X. Other approaches adopt preferentiality at a more abstract level so an attack from X to Y is successful only if Y is not preferred to X. These are the most intuitive approaches, easy to be implemented, but they are not the only ones that have been proposed. For instance, an attack from X to Y can be successful only if the value promoted by X is ranked higher or equal than the value promoted by Y, according to a given ordering on values. All these solutions assume that the information required to determine whether an attack is successful (a proper defeat) is pre-specified as a given value ordering or a given preference list, partial or not. However, the information related to preferentiality among arguments might be contradictory as well. For this reason the notion of *meta-level argument*, a special argument about preferences, has beed proposed. A meta-level argument is a simple argument (a node in the argumentation framework) and the application of preferentiality is abstractly characterised by the definition of a new attack relation that originates from this preference argument. The advantage is that this approach does not require any commitment regarding the definition of a preference list or an orderings of values. As opposite to the aforementioned approaches in which preferentially is considered between arguments, another branch of argumentation theory associates weights to attack relations instead to arguments. Here each attack has a value associated to it and, as a consequence, the consideration of all the attacks in the argumentation framework determines which attacks are successful. This value might be a crisp value or a more fuzzy number that allows the representation of the degree to which one arguments attack another one. Eventually, probabilities can be applied both to arguments and attack relations and they refer to the likelihood of the existence of a specific argument or a defeat relation, thus capturing the inherent uncertainties in the argumentation system.

The last layer of the argumentative process for knowledge representation (bottom layer of figure 3.12) is addressed by dialogical models and it is aimed at investigating the *dialectical status of arguments*. In other words, the layer is devoted to the determination of the final justification status of each individual argument. This depends on the interaction with the other arguments and it usually represents the last step for the determination of the outcome of an argumentation system. The terminology used for the dialectical status of arguments includes terms as 'justified' or 'defeated' and they are computed by *acceptability semantics*. A semantic specifies zero or more sets of acceptable arguments, called *extensions*. Modern implementations of dialogical models are based upon the *abstract argumentation* theory proposed in (Dung, 1995). Here, multiple extensions might exist coinciding with possible consistent points of view that can be taken into account for describing the knowledge being modelled. However, sometimes for practical reasons, a single decision has

to be taken. This might consider the strongest extension (according to a given criterion) or the aggregation of the computed extensions of arguments in a way that just one single value is eventually produced. This might be the case of mental workload, where a single index has to be inferred in order to take an action for system design purposes. The modular multi-layer argumentative schema for knowledge representation, as described so far, is appealing for representing and assessing the construct of human mental workload. Next section investigates this point making the starting assumptions behind this thesis (section 2.6.2, page 45) reasonable and valid.

3.7 Discussion on modelling mental workload as a defeasible construct

In order to clarify why defeasible reasoning and argumentation theory seem to be appropriate paradigms for representing the multi-dimensional complex construct of human mental workload, consider the following illustrative reasoning process that might be followed by a mental workload designer.

Example 9

A system designer wants to improve the design of a web-based interface to maximise its usability and optimise user engagement. To do this the concept of human mental workload is applied along with user studies that include the execution of tasks, on that web-interface, and the acquisition of subjective ratings related to different mental workload influencing factors. The designer believes that temporal demand, task mental demand, the psychological state of the user, the effort exerted to the task, his/her skills and the degree of external context bias are all useful factors to represent mental workload. In particular, task mental demand, temporal demand, psychological stress, exerted effort and context bias all have a direct relationship with mental workload: the higher is the quantification of the factor, the higher is the mental workload. Skill, instead, has an inverted relationship with workload: the higher the user's skills, the lower is the mental workload.

Now, assuming that the temporal demand required to complete a given task has been quantified as low and the objective completion-time by the user as low too: the designer might infer that the resulting mental workload imposed by the task was low. If the time dimension is the only evidence available, it is reasonable to propose such an inference. However, if it is also known that the end-user interrupted the execution of the task several times, due to high context bias, then the previous conclusion could be retracted. This new evidence may cause the designer to retract the resulting mental workload by inferring a high degree. Yet, although high context bias, if the user is known to be skilled, then the designer might assume that the execution of the task imposed low mental workload, retracting again the previous workload inference. Similarly, if the user has perceived a high degree of stress during the execution of the task, then the designer might prefer inferring again high workload, giving more importance to stress and less to the skills of the user. The same reasoning is applied to the factors mental demand and effort: the higher their quantification, the higher is the inferred mental workload.

Example 9 clearly emphasises the multi-dimensionality and complexity of the construct of mental workload, being this influenced by heterogeneous factors. The reasoning process followed by the designer is clearly defeasible, having the property of non-monotonicity because the potential representative index of overall mental workload is retracted several times in the light of new information. Formal argumentation logics appear to be useful for representing this kind of non-monotonic reasoning activity in form of interacting arguments. The natural language propositions and statements that vaguely try to link a certain degree, of a given workload factor (example temporal demand), to a certain degree of mental workload (example high) can be translated into formal arguments in form of inference from premises to conclusion. This consideration of the monological structure of arguments coincides withe the first layer of the argumentative multi-layer schema (page 67). From example 9 it also emerges a form of preferentiality among the argument built for the factor 'psychological stress' (a) and the argument built for the factor 'skills' (b). This preference could be either implemented with a partial preference list ($a > b$) or in form of unidirectional attack relation from (a) to (b). The union of designed arguments as well as designed attack relations coincides with the definition of 'argumentation framework' (second layer of figure 3.12, page 67). This framework is the formal translation of the designer's knowledge-base into interactive defeasible arguments. However, in practical settings, some of the premises of designed arguments might not be fully quantifiable, thus invalidating the defeasible argument itself. In turn, the designed argumentation framework can be fully or partially activated. This step coincides with the third layer of the argumentative schema (page 67). Here, even if some designed argument cannot be used for inferring mental workload, the overall reasoning process can still be carried out only with the remaining activated arguments. Eventually, the activated argumentation framework can be evaluated with acceptability semantics to extract consistent and conflict-free sets of arguments (extensions). The application of these semantics coincides with the last layer of figure 3.12 (page 67)Considering the arguments within the computed extension/s, an index of mental workload can be finally computed.

The aforementioned modular process is believed to be appealing for mental workload designers who can have a more structured methodology for representing and assessing mental workload. Defeasible reasoning (DR) and argumentation theory (AT) support the assumptions of section 2.6.2 (page 45) appearing now valid candidates for modelling the construct of human mental workload as a defeasible computational concept. The next chapter is devoted to the proposal of a formal defeasible framework, built in line with the aforementioned multi-layer argumentative schema.

Chapter 4

Design

This chapter is devoted to the design of a computational framework for representing the multi-dimensional construct of mental workload (MWL) and for assessing it by employing formal argumentation theory (AT). The word 'computational' refers to the fact that the framework employs numerical manipulable values and delivers a crisp usable numerical index. The design process is informed with a list of the ideal properties that such a framework should have, according to the author of this thesis, and his interpretation of the literature review of chapter 2. The design approach is in line with the multi-layer argumentative schema which emerged from the review of defeasible reasoning models of chapter 3, and the core tenets of mental workload which emerged from the literature review of chapter 2. It is a methodology which guides a modeller in how to represent mental workload, and which provides a tool for assessment. This methodology starts by building formal, defeasible arguments from natural language propositions, adopting the notion of degree of truth, borrowed from fuzzy logic, and the notion of logical consequence. These defeasible arguments can be connected in a graph using the notion of 'attack,Õ which is useful for modelling inconsistencies. This graph can be fully or partially activated through the quantification of the degrees of truth of each argument. The resulting activated argumentation graph is subsequently abstractly evaluated by applying Dung-style acceptability semantics for the resolution of the potential inconsistencies that might have arisen from the interaction of activated arguments. Eventually, a final index of mental workload is produced by aggregating the single assessments, one for each argument, in the most credible, acceptable extension, as computed by the selected acceptability semantics.

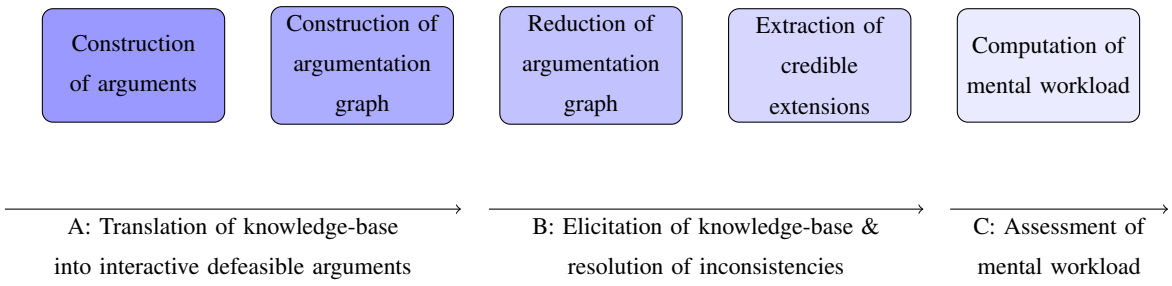


Fig. 4.1: Summary of the process for representing and assessing human mental workload

4.1 An ideal framework for modelling mental workload

According to what has been reviewed so far, it is clear that no single classification, definition or assessment procedure is capable of providing full information and entirely describing mental workload in human-computer interactive environments. Practitioners and designers have different knowledge-bases that influence their view of the construct of mental workload, thus they tend to represent it accordingly, for their contexts of application. In this thesis it is argued that mental workload could be better represented with a more complete framework able to handle several workload factors, their interrelationships and resolving the potential inconsistencies that are derived from their interaction. This formalism should take into account the uncertainty that characterises the definition of each workload factor as well as their aggregation, being as self-explanatory as possible. Furthermore, it is argued that an ideal formal framework for defining and assessing mental workload as a computational construct should have few properties, as proposed in the following paragraphs.

1. *flexibility*: an ideal framework should be open, enabling a designer to create an instance of mental workload representation by incorporating those factors believed to be useful for representing mental workload. This in turn would allow, for example, the extension or reduction of other instances built with the same framework. This flexibility supports adaptability, especially in human-computer interactive environments, characterised by heterogeneity and different constraints. In other words, an instance of the framework should be open to adjustments and refinements (Popper, 1967, page 53), thus being falsifiable (Popper, 1967, 1969).
2. *falsifiability*: flexibility would enable falsifiability. According to Popper, this is not a negative property, rather it is a positive quality because it means that a hypothesis (a set of workload factors and a set of relationships among them) is testable by empirical experimentations, conforming to the standards of scientific methodologies. If something is falsifiable or refutable, it does not mean that it is false, but rather, in the case where it is false, then the observations and experiments carried out will, at some stage, demonstrate its falsehood (Popper, 1967). Given the same inputs, an instance of the framework that leads to a certain workload index can be falsified by another instance which embeds different factors whose aggregation leads to another workload index. Each proper test aimed at defining and shaping/modelling mental workload is an attempt to falsify it.
3. *replicability*: an ideal framework should be replicable and duplicable. A mental workload index, computed by an instance of the framework, regardless if expressed as a scalar number or as a vector value, must be able to be repeated and duplicated. The replication of an instance of the framework, built over the same set of workload attributes, their interrelationships, and activated with the same input set, must always infer the same result (Popper, 1967, page 45). Replicability refers also to the application of an instance of the framework in different contexts.
4. *simplicity*: an ideal framework should be simple and as intuitive as possible to be used even by non-experts. It should allow practitioners, not fully familiar with the concept of mental workload, to design each workload factor and to specify the relationships between them believed to be useful in representing mental workload. As asserted by Popper, simplicity is better than complexity because it allows extreme and multiple tests to be carried out (Popper, 1967, 1969). This in turn supports replicability, allowing several applications to be tested in different settings and environments. Additionally, without attempting

a simple exploration of the concept of mental workload, little can be said about the result obtained. Undoubtedly, a complex formalism based upon advanced mathematics concepts could be extremely precise to model mental workload, but if it works in a particular context, there might be the issue of explaining and justifying why it works. A simple framework, with a better self-explanatory capacity, usually has a higher degree of testability than more complicated ones (Popper, 1969, page 61).

5. *inconsistency-awareness*: a framework should be able to handle contradictions of workload factors and potential inconsistencies arising from their interaction. In other words, an explicit strategy for inferring an index of mental workload from a set of workload factors, that can be contradictory in a given context, should be embedded in the framework.

4.2 Top-down design approach

The design of the framework follows a top-down approach, firstly formulating a multi-layer overview, and subsequently expanding each layer in greater detail. The assumptions behind this approach are:

- Assumption 1: the framework is to be used by an MWL modeller/designer
- Assumption 2: a prior existence of a knowledge-base of MWL, that means a set of attributes believed to influence mental workload
- Assumption 3: a modeller has an expertise in relation to the construct of MWL, thus understanding how accounted attributes interact with each other.

The process that allows a designer to represent and shape mental workload according to their own knowledge-base, and the computational model that yields a mental workload assessment as a crisp numerical index, is summarised in detail in the flow-chart of figure 4.2.

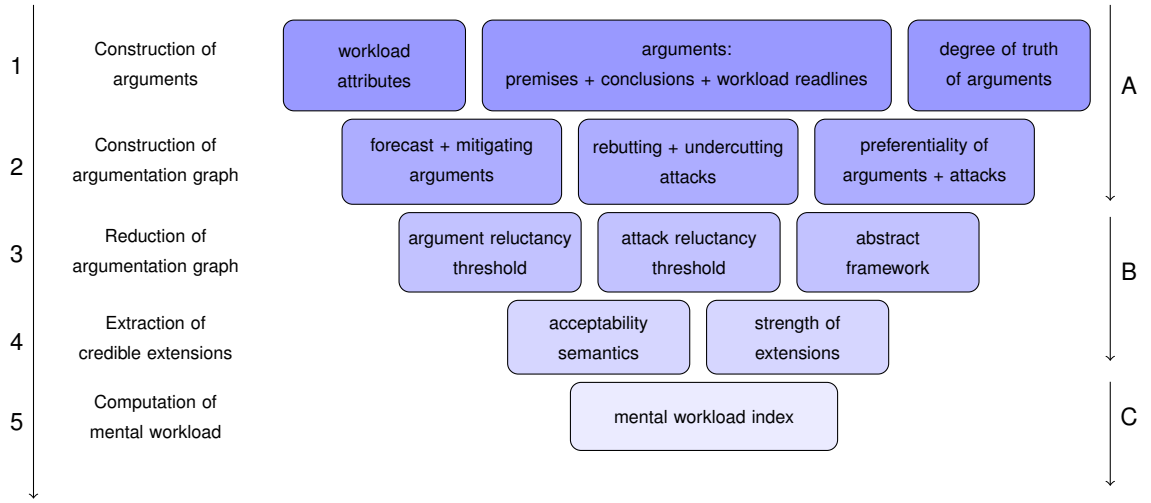


Fig. 4.2: Multi-layer framework for human mental workload: a detailed view

The first two layers are devoted to the translation of a workload designer's knowledge-base into interactive *defeasible arguments*. The first layer focuses on the construction of arguments from natural language

propositions in a *monological structure*. These propositions are based upon a set of *attributes* believed to influence mental workload. An argument links a set of premises, built upon selected workload attributes, to a workload dichotomy: a tentative assessment of mental workload for that argument. Four *dichotomies* are possible: underload, fitting (lower and upper), and overload. These dichotomies are separated by two *redlines*: special thresholds aimed at defining the limits for 'too low' and 'too high' workloads. Dichotomies are mutual and jointly exclusive; a mental workload quantification must belong to one dichotomy and it can not belong simultaneously to more than one. Constructed arguments can be elicited via quantification of their premises, and activated with a certain degree of truth.

The second layer focuses on the *dialogical structure* of arguments, that meaning their *interaction*. Two typologies of arguments are formalised: forecast arguments when in favour of a certain workload dichotomy, and mitigating arguments, when challenging the information and knowledge employed to construct other forecast or mitigating arguments. Interaction is implemented using the notion of *attacks* and it is aimed at modelling logical inconsistencies of arguments. An attack uni-directionally or bi-directionally links two arguments. A *rebutting attack* is always symmetrical and can be used with two arguments known to be incompatible and which cannot coexist. An *undercutting attack* is unidirectional and can be employed to highlight special cases in which an argument, if activated, invalidates another argument. The set of inter-connected arguments through attack forms an *argumentation graph*: a formal representation of a designer's knowledge-base. The process of knowledge-base translation is coupled with a method for acknowledging *preferentiality*, if any, of those workload attributes considered by an MWL designer.

The third and fourth layers are devoted to the elicitations of the knowledge-base previously translated into an argumentation graph. Specifically, the third layer requires an MWL designer to define two *reluctancy thresholds* that are used to activate just those arguments (and in turns attacks), with a certain degree of truth. The resulting graph, a sub-set of the previously designed argumentation graph, is abstractly evaluated, without considering the internal structure of arguments, by employing Dung-style *acceptability semantics*. The output of these semantics is a set of *extensions*, sub-sets of arguments, that can be seen as different but internally coherent points of view available to a designer for assessing mental workload. The fourth layer focuses on the computation of the *strength* of each of these extensions, and on the extraction of the most credible extension.

Eventually, the fifth, and last layer, is devoted to the final *assessment of mental workload* which is a crisp numerical index that can be practically employed for design purposes. The following sections describe, in greater details, each of the aforementioned layers.

4.3 Layer 1 - translation of knowledge-base

It is expected that designers intuitively have their own understanding of mental workload, being this influenced by the their working field, background, beliefs as well as the context of application. Consider example 10 which illustrates a knowledge-base of a designer through a set of natural language propositions.

Example 10

---

1. *'the mental demand required by task T is linearly related to workload: the higher the demand, the higher the mental workload'*
2. *'given a low degree of performance there is a reason to believe that the mental workload exerted by a user on task T is high'*
3. *'although task T is highly mentally demanding, the user has a high degree of past knowledge so there is a reason to believe the mental workload exerted on T is not high'*

It is clear that mental workload is a multi-dimensional construct, influenced by many factors. In the above example these are 'mental demand', 'performance' and 'past knowledge'. These factors are now on referred to as *workload attributes*.

4.3.1 Workload attributes

The first step for translating the knowledge-base of a designer is to define those factors believed to influence mental workload: the workload attributes. These attributes might have a different influence on overall mental workload, either linear or non-linear, positive or negative. In addition, these factors can either support and influence mental workload individually, or as a combination of two or more factors.

Notation 1
The list of workload attributes believed to influence mental workload is a set of labels defined by a workload designer. This set is now on referred to as the *ATTR* set.
In example 10 $ATTR = \{\text{'}mental\ demand\text{'}, \text{'}performance\text{'}, \text{'}past\ knowledge\text{'}\}$.

4.3.2 Translating natural language propositions into formal arguments

In the first proposition of example 10 it is implied that the mental demand of a task linearly relates to mental workload; the higher the degree of mental demand the higher the mental workload elicited by a user for performing a given task. This relation can computationally be formalised with a simple linear monotonic function. However, the second natural language proposition can not be intuitively formalised like the first one. The issue is that it contains linguistic variables such as 'low' performance and 'high' mental workload which are vague terms; this adds complexity in formalising them as computational notions. The third proposition is even more complex, because not only linguistic variables such as 'highly mentally demanding' and 'high degree of past knowledge' are used, but also because the mental workload to be inferred depends on two attributes: the 'mental demand' of the task and the 'past knowledge' of the user. The next section is devoted to the resolution of the aforementioned issues. The natural language propositions of example 10 can be seen as arguments, and, as mentioned in chapter 3, they can be seen as tentative proof for proposition. The understanding of an argument as a tentative proof relates to its internal representation (as described in section

3.2, page 52) and its monological structure. Generally, an argument is composed of a set of premises and a claim which can be derived by the application of some inference rule $\rightarrow$.

$$Argument : premises \rightarrow claim$$

This implication is intrinsically uncertain, and coherent with human reasoning that is uncertain rather than exact. In this thesis, the proposal is to see a workload attribute (for instance the factor 'performance' of example 10 - proposition 2) as the premise of an argument, followed by a claim which is a possible conclusion a designer wishes to infer (for example 'low/high workload'). Therefore a possible monological (internal) informal structure of the arguments that can be built upon the natural language propositions of example 10 might be described as in example 11.

Example 11

- $a : Low\ mental\ demand \rightarrow Underload$,
  $b : Medium\ mental\ demand \rightarrow Fitting\ load$
  $c : High\ mental\ demand \rightarrow Overload$
- $d : Low\ performance \rightarrow Overload$
- $e : High\ mental\ demand$ AND $High\ past\ knowledge \rightarrow Underload$

In turn, the issue is how to computationally represent:

- the vague linguistic terms associated to each attribute in the premise of each informal argument
- the conclusion of each informal argument.

Formalising premises of arguments

In order to formalise the premise of argument, the proposal is to use Fuzzy Set Theory (FST) and *degrees of truth* (Zadeh, 1965) (Zadeh, 1966). Degrees of truth can be computationally modelled using *membership functions*: particular functions useful for formalising vaguely defined sets (fuzzy sets) and human reasoning which is approximate rather than fixed and exact. Membership functions allow the mapping of an attribute's numerical quantification to the relative set with degrees of truth. This process is often referred to as *fuzzification*; it transforms crisp values into grades of membership for linguistic terms. A membership function is subsequently employed to associate a grade to each linguistic term.

Definition 11 (Membership function)

For any set X, a membership function on X is any function

$$f : X \to [0,1] \in \Re$$

Membership functions on X represent fuzzy subsets of X. For an element x of X, the value $f(x)$ is called the 'membership degree' or 'degree of truth' of x in the fuzzy set and quantifies the grade of membership of x to the fuzzy set X. The set of membership functions defined over X is defined as

$$MF_X = \{f | f : X \to [0,1] \in \Re\}$$

■

> **Note 3**
>
> For each fuzzy set X, zero or more membership functions can be defined and usually they can partially overlap, sharing some values of X, but not necessarily returning the same degree of truth for the same input x. A membership value of 0 and 1 indicates respectively non-membership and full membership: intermediate values refer to fuzzy members partially belonging to X.

Example 12

The natural language expression 'low performance' might be expressed by the membership function $f^{Low}_{Performance}(x)$ which quantifies the grade of membership of a level x of performance to the fuzzy subset 'Low' of the attribute (set) 'Performance'. Various membership functions may be used to model the fuzzy subset 'Low' of the fuzzy set 'Performance', as in figure 4.3. Here, a performance value quantified as 28, (on a scale bounded from 0 to 100), has three different degrees of truth, according to the three different membership functions. For example, the straight line function (green) returns a degree of truth of 0.72. For the same function, if the quantified performance would have been 1, then the degree of truth would have been 1, meaning that the designer of the membership function would have been fully confident that a value of 1 for 'performance' fully belongs to the low set.

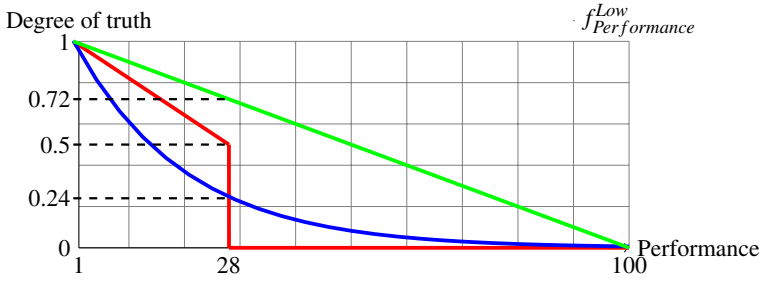


Fig. 4.3: Possible membership functions for the fuzzy set 'Performance' and its fuzzy subset 'Low'

Note 4

A fuzzy membership function can be any function, from the classical straight line to the step-function, or more complex such as sigmoidal functions, logarithmic functions or curves in general. A membership function provides a designer with a flexible tool for modelling an attribute and relative sub-classes.

Each of the membership functions designed for a workload attribute maps an input space to an output space of degrees of truth. In isolation or combined with other membership functions (of other workload attributes), they form the premises of an argument. These premises have to be tentatively linked to a conclusion representing the index of mental workload that is to be inferred (examples are *underload, fitting workload or overload*.

Formalising the conclusion of arguments

In order to model the conclusion of an argument, the proposal is to design a function that, given a degree of truth of the argument's premise, it returns an index of workload. Consider possible translations of the natural language proposition of example 11 as proposed in figure 4.4. The value returned by the function that models a conclusion has to be unique; from a premise, one and only one index of workload must be inferred. This property is clearly violated by argument *b* (figure 4.4) where two possible indexes of mental workload can be inferred from a degree of truth of the premise. In order to solve this issue, the proposal is to force a designer to use a *strict monotonic function* [1] for the conclusion of an argument. Clearly, the function associated with the conclusion of argument *b* (*fitting workload*) is not a strict monotonic function. A possible solution is to split the range of workload indexes, covered by the function *fitting workload*, into two non-overlapping ranges: *fitting lower* and *fitting upper*. The set of functions used for the conclusion of arguments, aimed at covering the range of possible mental workload indexes, contrarily to the premise of arguments, cannot overlap (sharing part of the output space). For this reason from now on, they are referred to as *workload dichotomies*. These dichotomies are both *jointly exclusive* and *mutually exclusive*; the premise of an argument must be associated with one partition and it can not be associated simultaneously with more than one partition. The proposal is to model the mental workload range of inferrable indexes with a continuous space bounded in the range 0 to 100 (as per definition 12) and to use 4 dichotomies to partition this space. The remaining issue is how to set the boundaries of each dichotomy. As mentioned in section 2.2.1 (page 17), these are not static boundaries, rather they depend on the context of application, the task and the operator. In the literature of mental workload, these boundaries are sometimes referred to as *redlines*. They are aimed at defining the areas of low and high mental workload, and separating these from the area of optimal workload (as in figure 2.2, page 11). The proposal is to use the same terminology, and allow a designer to set them according to their own knowledge-base applied to a given experimental context. The are two boundaries involved, as depicted in figure 4.5: one that divides the dichotomies *underload* and $fittingLower$, and one that separates the dichotomies *fitting upper* and *overload* (definition 13). The following paragraphs provide the reader with the formal definitions of overall mental workload, redlines and dichotomies.

[1] A strict monotonic function has the property that for two different inputs, the former being greater than the latter, its output, given the former input, is greater than its output, given the latter input. In other words, just one unique output corresponds to an input.

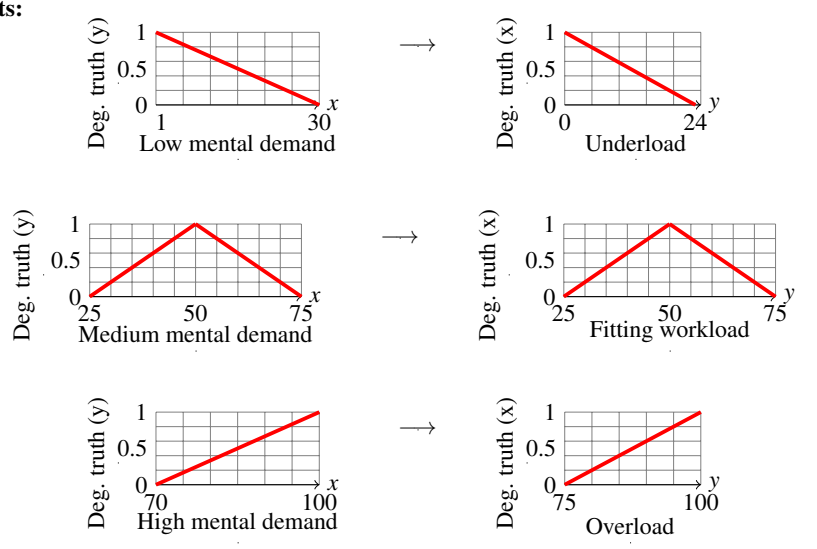
Informal arguments:

a: (*Low mental demand* → *Underload*)

b: (*Medium mental demand* → *Fitting workload*)

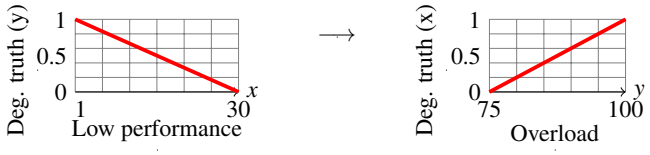
c: (*High mental demand* → *Overload*)

Formal arguments:



Informal argument: d: (*Low performance* → *Overload*)

Formal Argument:



Informal argument: e: (*High mental demand* AND *High past knowledge* → *Underload*)

Formal Argument:

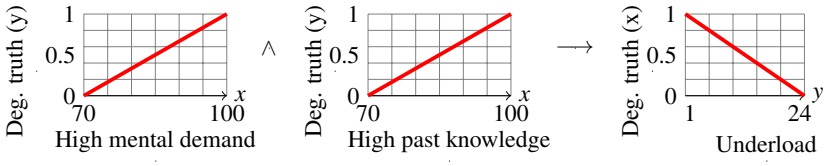


Fig. 4.4: A possible translation of natural language propositions into formal arguments

Definition 12 (Overall mental workload)
The overall mental workload is a real number $MWL : [0..100] \in \Re$ ■

Definition 13 (Redlines)

Let MWL the overall mental workload space. The redlines used for separating MWL are two:

$$RedLine^{fitting}_{underload}, RedLine^{overload}_{fitting} \in MWL$$

with $0 < RedLine^{fitting}_{underload} < 50 \leq RedLine^{overload}_{fitting} < 100$ ■

Definition 14 (Workload dichotomies)
Let MWL the overall mental workload space. Let $RedLine^{fitting}_{underload}$ *and* $RedLine^{overload}_{fitting}$ *the two redlines. A workload dichotomy is a sub-set A of MWL denoted as a function that takes a degree of truth:*

$$f : [0..1] \rightarrow A \subset MWL$$

Four dichotomies are defined:

$$UNDERLOAD = f_{UNDERLOAD} : [0..1] \rightarrow [0..RedLine^{fitting}_{underload})$$

$$FITTING^{-} = f_{FITTING^{-}} : [0..1] \rightarrow [RedLine^{fitting}_{underload}..50)$$

$$FITTING^{+} = f_{FITTING^{+}} : [0..1] \rightarrow [50..RedLine^{overload}_{fitting}]$$

$$OVERLOAD = f_{OVERLOAD} : [0..1] \rightarrow (RedLine^{overload}_{fitting}..100]$$

■

Property 1
The four functions are strictly monotonic thus it holds that:

- $\forall x, y$ with $x < y$ then $f(x) < f(y)$

Property 2
The sets follow a mutual exclusive property thus it holds that:

- $UNDERLOAD \cap FITTING^{-} \cap FITTING^{+} \cap OVERLOAD = \emptyset$ and
- $UNDERLOAD \cup FITTING^{-} \cup FITTING^{+} \cup OVERLOAD = MWL$

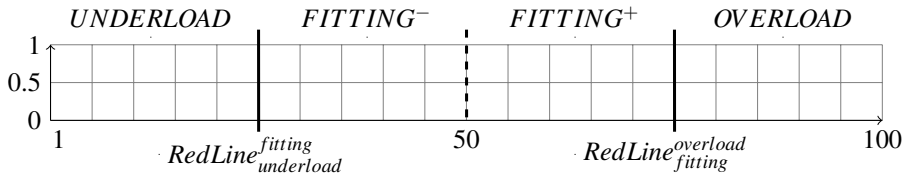


Fig. 4.5: Workload space separated into four dichotomies by redlines

Definition of argument

According to what has been designed so far, having a formal tool for representing a premise of an argument (membership functions) and its conclusion (dichotomy), an argument can be formally defined as well.

Definition 15 (Argument)
An argument A is a tentative inference $\rightarrow$ *that links premises* $P_1, ..., P_n$ *to a claim C.*

$$A : P_1, ..., P_n \rightarrow C$$

with $P_X : f_X \in MF_X$ *and* $C \in \{UNDERLOAD, FITTING^{-}, FITTING^{+}, OVERLOAD\}$. *Each* P_X *is a premise built upon a given workload attribute X and it is modelled with a membership function* f_X. *A claim C is the conclusion of the argument and it is modelled with a workload dichotomy.* ■

Modelling an argument as just described has some advantages:

- *simplicity*: it provides a simple way for representing vague natural language propositions and beliefs;
- *structure*: it affords a detailed structure for aggregating different beliefs and pieces of knowledge;
- *quantification*: it offers a method (membership functions) able to handle uncertainty in the definition of workload attributes and the quantification of their degree of truth;
- *inference*: it deliveres a method for tentatively inferring a mental workload index from the degree of truth of beliefs and knowledge computed in a given experimental context.

The notion of argument is already very powerful because it enables the translation of the uni-dimensional Rating Scale Mental Effort (RSME) workload assessment procedure (Zijlstra, 1993). RSME considers the workload attribute 'effort' as the unique source of information for predicting mental workload. The more the effort devoted to a task, the more the mental workload exerted, as depicted in the original scale (figure 4.6). Example 13 shows how the attribute 'effort' can be modelled using four membership functions (figure 4.7) and how four arguments can be built upon these functions (figure 4.8).

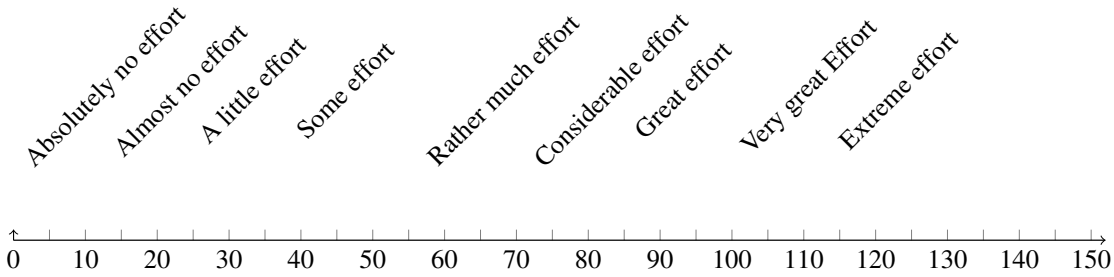


Fig. 4.6: Rating scale mental effort

Example 13

The arguments that might be designed for the translation of the RSME uni-dimensional workload assessment procedure are:

- A: *Low Effort* $\rightarrow UNDERLOAD$
- B$^-$: *Medium lower Effort* $\rightarrow FITTING^-$
- B$^+$: *Medium upper Effort* $\rightarrow FITTING^+$
- C: *High Effort* $\rightarrow OVERLOAD$

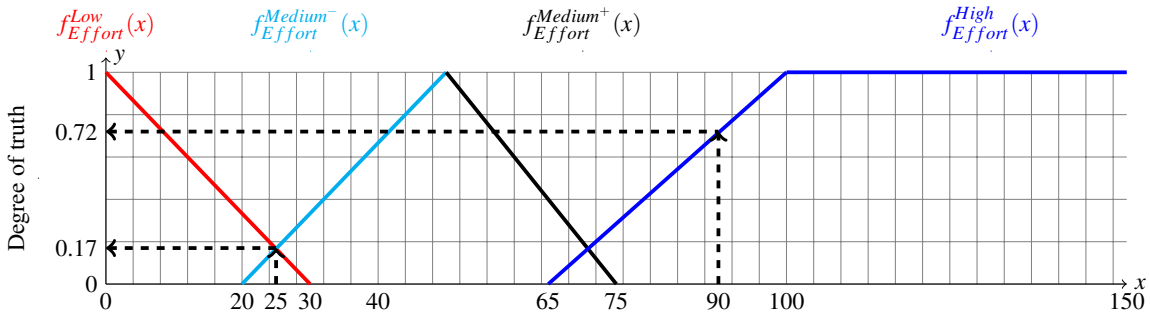


Fig. 4.7: Illustrative membership functions for attribute 'effort' of the Rating Scale Mental Effort instrument

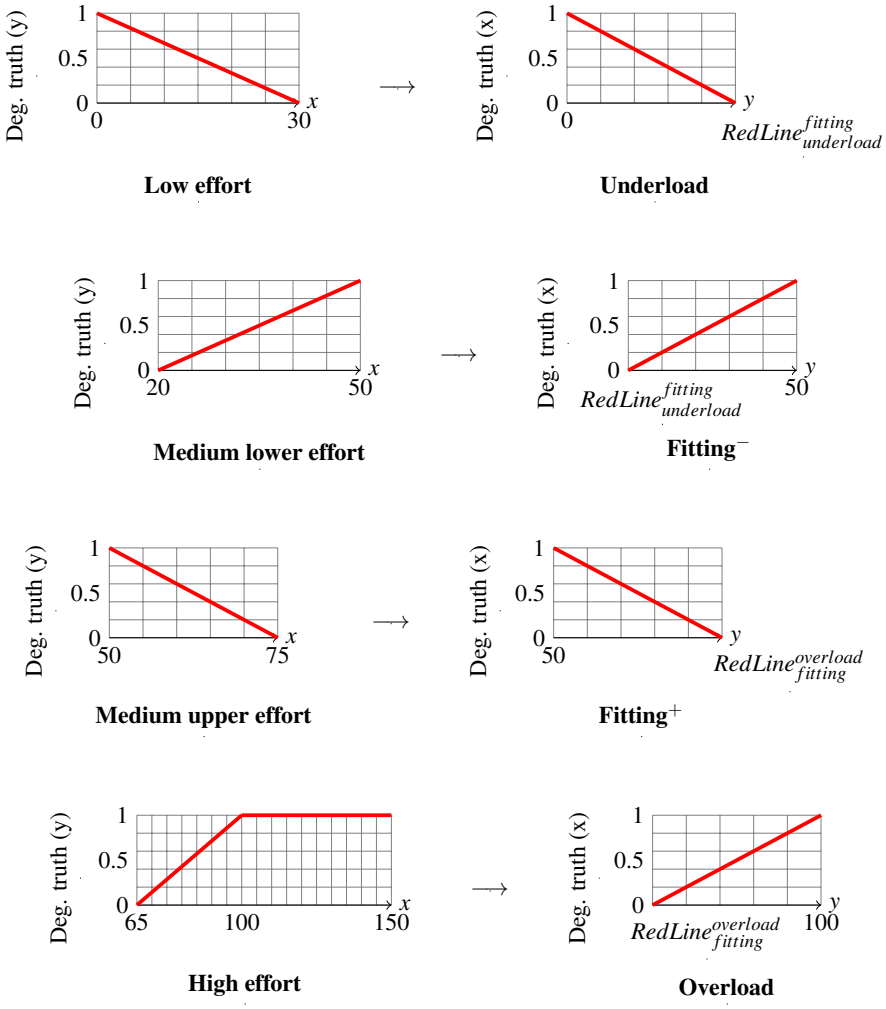


Fig. 4.8: Illustrative arguments for modelling the Rating Scale Mental Effort instrument

4.3.3 Computing the degree of truth of argument

In example 13, four arguments were designed, each having just one attribute embedded in their premise: the 'effort'. In turn, only one membership function needs to be evaluated to compute the degree of truth of that attribute. This degree can also be employed as the representative degree of truth of the argument itself. However, the premise of an argument can contain multiple attributes, resulting in multiple membership functions, one for each attribute. It turns out that, in order to produce a unique representative *degree of truth of an argument*, the single degrees of truth associated with each attribute in the premise have to be aggregated. Several methods for aggregating degrees of truth exist in the literature of multi-valued logics such as the fuzzy-AND [2], the fuzzy-OR[3] or other simpler methods such as the sum or average. Deciding which method is the most appropriate is a non-trivial problem. This decision might be influenced by the expertise of a

[2]Fuzzy-AND (intersection): given two membership functions with two input $f_1(a)$, $f_2(b)$, their fuzzy AND is the minimum degree of truth $min(f_1(a), f_2(b))$

[3]Fuzzy-OR (union): given two membership functions with two inputs $f_1(a)$, $f_2(b)$, their fuzzy OR is the maximum degree of truth $max(f_1(a), f_2(b))$

designer and the context of application, along with other factors. For this reason, in this thesis the proposal is to let a workload designer decide the most suitable operators for the unification of more attributes within an argument's premise.

Definition 16 (Argument's degree of truth)
Let A be an argument. The degree of truth A_{deg} of A coincides with the aggregation of the degree of truth P^{deg} of each premise P, by employing one or more Φ operators.

$$A_{deg} = P_1^{deg}\ \Phi^1, ..., \Phi^n\ P_n^{deg}$$

with $P_i^{deg} = f_X(\alpha)$ and X a workload attribute, $f_X \in MF_X$ and α a crisp quantification of X. ■

To clarify the computation of the degree of truth of an argument, consider example 14 that deals with the illustrative arguments built for the RSME of example 13.

Example 14

Scenario 1
Given an 'effort' quantified as 90, the arguments designed in example 13 (and formalised in figure 4.8) have the following degrees of truth:

- $A_{deg} : f_{Effort}^{Low}(90) = 0$
- $B_{deg}^{-} : f_{Effort}^{MediumLower}(90) = 0$
- $B_{deg}^{+} : f_{Effort}^{MediumUpper}(90) = 0$
- $C_{deg} : f_{Effort}^{High}(90) = 0.72$

Scenario 2
Given an 'effort' quantified 25, the arguments designed in example 13 (and formalised in figure 4.8) have the following degrees of truth:

- $A_{deg} : f_{Effort}^{Low}(25) = 0.17$
- $B_{deg}^{-} : f_{Effort}^{MediumLower}(25) = 0.17$
- $B_{deg}^{+} : f_{Effort}^{MediumUpper}(5) = 0$
- $C_{deg} : f_{Effort}^{High}(25) = 0$

It is evident that argument *C* of scenario 1 is the only candidate for inferring an overall index of mental workload, with a degree of truth of 0.72. In scenario 2 there is more than one candidate argument, thus multiple workload indexes can be generated. Argument *A* and *B* in this specific case have the same degree of truth. However, arguments can be activated with different degrees of truth, thus a strategy for the accrual of these degrees is needed in order to infer a unique final index of mental workload that can be used for practical purposes. Additionally, the Rating Scale Mental Effort workload assessment procedure is a uni-dimensional instrument based upon just one single workload attribute ('effort'). Mental workload is strongly believed being a multi-dimensional construct; multiple attributes can be considered and in turn multiple arguments can be built. These arguments can interact, conflicting with each other and creating contradictory scenarios. For this reason they are referred to as *defeasible arguments*. One of the assumptions behind this thesis is that

accounting and understanding the relationships of workload attributes as well as resolving the inconsistencies that might arise from their interaction is essential in modelling human mental workload. As emerged in the literature review of defeasible reasoning of chapter 3, the analysis of the interaction of defeasible arguments is addressed by *dialogical models*. In the next section, various typologies of argument as well as different ways of attacking each other are described, leading towards the construction of an *argumentation framework*: the dialogical translation of a knowledge-base.

4.4 Layer 2 - Construction of the argumentation graph

Constructing an argumentation framework is addressed by dialogical models aimed at formalising the interaction among different arguments and complementing monological models. From the illustrative scenario of example 11 (page 76), each argument is aimed at tentatively inferring mental workload because it has a given claim in its conclusion part (a workload dichotomy). However, argument e is somehow contradics argument c. Both have *high mental workload* within their premises, but the former claims *overload* while the latter claims *underload*. Argument e might be rewritten also as: 'e_1: *high past knowledge* $\rightarrow$ not c'. In this case, the fact the user has a high degree of past knowledge mitigates argument 'c' that is no longer valid: e_1 attacks what is claimed by argument c. According to a previous study (Matt et al., 2010), arguments can be divided into 2 classes:

- *forecast arguments* when they are in favour or against a certain claim (workload dichotomy), but justification is not infallible. This coincides with definition 15 of argument.
- *mitigating arguments* when defeating forecast or other mitigating arguments, undermining their justification.

4.4.1 Forecast and mitigating arguments

Forecast arguments are tentative defeasible inferences and they can be seen as justified claims concerning the expected or the anticipated behaviour of the target (mental workload assessment). They represents hints or clues given by a designer under uncertainty and not mathematical proofs, in line with (Krause et al., 1995). In turn the validity of these arguments has to be carefully evaluated by a mental workload designer. A forecast argument '$P_1, ..., P_n \rightarrow c$' can be read as 'there is a reason to believe c from $P_1, ..., P_n$' or 'c is what reasonably follows from $P_1, ..., P_n$'. As anticipated, the definition of forecast argument coincides with the definition 15 of argument (page 80).

Mitigating arguments are used to express the uncertainties of a designer concerning the validity of forecast arguments or other mitigating arguments. In other words, they have the effect of undermining the validity of other arguments. Different from forecast arguments, mitigating arguments link a set of premises to another argument, negating its conclusion or challenging its inference link. In other words, the set of premises of a mitigating argument undermine what is claimed by another argument, either forecast or mitigating. Mitigating arguments are useful for modelling special cases and conflicts that might arise during the reasoning process followed by a designer, to shape mental workload.

Definition 17 (Mitigating argument)
A mitigating argument A is an undermining inference $\Rightarrow$ that links a set of premises $P_1,...,P_n$ to argument B

$$A : P_1,...,P_n \Rightarrow B$$

where each premise P_X represents a given workload attribute X and it is modelled with a membership function $f_X \in MF_X$ and B is either a forecast argument or another mitigating argument. ■

Notation 2 (Sets of forecast and mitigating arguments)
For the reminder of this thesis, the sets of forecast arguments and mitigating arguments are respectively referred to as AR^F and AR^M.

Mitigating arguments are forms of conflicts and, as emerged in the literature review of defeasible reasoning of chapter 3, conflicts can be modelled as attack relations between arguments.

4.4.2 Rebutting and undercutting attack relations

Attack relations between arguments (as mentioned in the literature review of section 3.3, page 56), can be *rebuttal*, *undercutting* or *undermining*. The first type occurs between two forecast arguments contradicting each other because they support contradicting claims. In other words, arguments affected by a rebutting attack cannot coexist because they support conflicting knowledge. These arguments follow a bi-directional attack: given two contradicting forecast arguments, they both attack each other. An undercutting attack occurs when a mitigating argument challenges the claim of a forecast or another mitigating argument. It attacks the link of the target argument by claiming that there is a special case that does not allow the application of its inference link from premises to conclusion. The third type is the undermining attack: an argument can be attacked on one of its premises by another argument having a conclusion that negates those premises. Undercutting and undermining attacks are uni-directional (not symmetrical). These can be used by a workload designer, either to undermine what is claimed by a forecast argument, or by challenging the attack of another mitigating argument. In this thesis, only the notion of undercutting attacks is adopted. The rationale for this is to keep the formalism as simple as possible. An undercutting attack is always generated by a mitigating argument.

Definition 18 (Rebutting attack)
Let A, $B \in AR^F$ with $A \neq B$ be two distinct forecast arguments. A is a rebuttal of B if they logically contradict each other. This attack is denoted as (A,B). ■

Property 3
A rebuttal attack is symmetrical so it holds that

- iff (A,B) then $\exists(A,B)$

Definition 19 (Undercutting attack)
Let $A \in AR^M$ be a mitigating argument that challenges some or all of the information used to construct a different argument $B \in AR^F \cup AR^M$, either forecast or mitigating with $A \neq B$. $A : P_1,...,P_n \Rightarrow B$ is an undercutting of B if $P_1,...,P_n$ attacks the inference link of B ($\rightarrow$ if $B \in AR^F$ or $\Rightarrow$ if $B \in AR^M$) by claiming there is a special case that does not allow the application of such an inference. This attack is denoted as (A,B). ■

Note 5

In the definitions of rebutting and undercutting attacks (18, 19), the attacker and the attacked arguments must be distinct. As a consequence, this excludes situations of self-defeating in which an argument attacks itself. In this thesis is it assumed that a workload designer does not deal with self-defeating propositions and pieces of knowledge.

Notation 3

For clarification purposes, the following notations are adopted for the typologies of argument and attacks:

- *forecast argument*: a 1-border circle;
- *mitigating argument*: a 2-borders circle;
- *rebutting attack*: a directed arrow line. As a rebutting attack is always symmetrical, 2 directed arrow lines exist. Another notation can be interchangeably used: a single double-direction line (with two arrows);
- *undercutting attack*: a 1-dashed-line.

4.4.3 Preferentiality of arguments and attacks

According to the literature of chapter 2, preferentiality of workload attributes is an important property for assessing mental workload. A designer may prefer one workload attribute to another, thus one attribute may have a greater influence on the overall assessment. Preferentiality, in argumentation theory (as discussed in section 3.4.1, page 58), can be implemented in two ways:

- explicitly preferring arguments to others
- adding a strength to the attack relations between arguments

In the former case, preferentiality is usually implemented as a preference-list of arguments, total or partial. This list can either be a simple ranking list, hence without any indication of the relative difference between two arguments, or a list with an explicit numerical strength attached to arguments. In the latter case, preferentiality is achieved by adding a numerical value on the attacks relation between two arguments indicating its strength.

Preferentiality of arguments

Preferentiality of arguments, in the context of mental workload modelling, and according to what has emerged in the literature review of mental workload assessment techniques of chapter 2, can originate from:

- the reasoning process followed by a workload designer, while considering a set of attributes believed to influence mental workload;
- the subjective judgements of workload attributes quantified by raters (example through a questionnaire).

In the former case, a workload designer provides static preferences, extracted from their own knowledge-base, that do not change during the assessment of mental workload. In other words, a workload designer and expert already knows the importance of some arguments before mental workload assessment. In the latter case, preferences are dynamically quantified during the assessment of mental workload, thus being different for a given task performed by a given user. For example, in the NASA-TLX (Hart and Staveland, 1988),

end-users are provided with a given task, and after its execution, they are required to perform a pair-wise comparison of 6 workload attributes. This comparison is a boolean preference of two attributes, for each possible pair of attributes, which eventually generates an absolute preference list of the 6 workload attributes (details in section 2.4.2, 33 and appendix A.1.1, 173). Regardless of the method employed for preferentiality, sometimes it is not always intuitive and straightforward to build a complete preference list among those workload attributes believed to influence the assessment of mental workload. For this reason, in this study, the proposal is to implement the idea of preferentiality as a partial preference list of those workload attributes a designer is willing to consider in own representation of mental workload. A partial list is a flexible option that can become useful in the extreme cases in which preferentiality is either undefined or defined for all the workload attributes, resulting in an empty list or a total list respectively.

Definition 20 (Attribute preference)
Let $ATTR$ a finite set of workload attributes. f_{pref} is a partial function that maps an attribute to an importance value:

$$f_{pref} : ATTR \rightarrow [0..1] \in \Re$$

If $f_{pref}(x) \leq f_{pref}(y)$ with $x \neq y$ and $x, y \in ATTR$ then the attribute y is said to be preferred or it has equal importance than the attribute x. ■

Property 4
f_{pref} is a partial function so for any $x \in ATTR$, either:

- $f_{pref}(x) = [0..1] \in \Re$ or
- f_{pref} is undefined

Property 5
It holds that for any defined $f_{pref}(x)$, $f_{pref}(y)$, $f_{pref}(z)$ with $x \neq y \neq z$, and $x, y, z \in ATTR$

- If $f_{pref}(x) \leq f_{pref}(y) \leq f_{pref}(z)$ then z is preferred to or is equally important than x (transitivity);

The definition of preferentiality as a partial function enables the definition of preferentiality of arguments. According to definition 15 (page 80), an argument might have different premises built upon different workload attributes. If the premise of an argument contains just one attribute, then the importance of that argument coincides with the importance of that attribute. In the case where multiple attributes are employed in the premise of an argument, the average of their importance is proposed to be the overall importance of that argument.

Definition 21 (Argument importance)
Let $A : P_1, ..., P_n \rightarrow C$ be an argument, $ATTR$ the finite set of workload attributes a designer is accounting for in own knowledge-base, f_{pref} the preference partial function over the attributes in $ATTR$ and each of the n premises of A is built upon an attribute $x_i \in ATTR$. The importance of argument A is

$$A_{imp} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} f_{pref}(x_i), & if\ \forall x_i \in ATTR,\ f_{pref}(x_i) \neq undefined \\ undefined, & otherwise. \end{cases}$$ ■

According to the above definition, the importance of an argument is defined if and only if the importance of every single attribute embedded in its premise is defined as well in the preference list. If there exists even one single attribute within the premise of that argument that has an undefined importance, then it is not possible to assign a clear importance to that argument.

Preferentiality of attacks

An alternative method for implementing preferentiality, in the field of argumentation theory, is to assign weights to attack relations. A weighted attack underlines the strength of the attack, thus having an influence on the attacked (target) argument. In this study, it is argued that the method of attaching a value to an attack relation might not be straightforward and intuitive, and it adds a burden to a mental designer, while modelling mental workload. In addition, a designer might not have a meaningful number to assign to each attack relation, or he might not even have any clue about the strength of an attack. For these reasons, the proposal is to provide a designer only with the aforementioned idea of importance of arguments, and implicitly assume the strength of attack relations by employing such a idea. It turns out that the importance of arguments is a primitive notion, as computed by considering the explicit preferences of workload attributes, while the weights of attacks is a derived notion, as implicitly inferred by considering the importance of arguments. In this study, it is believed that this design choice would minimise the burden on a workload designer imposed by the reasoning process of translating own knowledge-base into interactive defeasible arguments.

4.4.4 Argumentation graph

Implementation of preferentiality is the last step for the translation of a designer's knowledge-base into interactive defeasible arguments, completing the layer A of the schema of figure 4.1. The output of this layer is an argumentation graph where each node is an argument and each link is an attack relation. The argumentation graph can now be elicited with a quantification of each workload attribute. Example 15 illustrates an argumentation graph with 3 arguments employing the terms of notations 3 (page 86.

Example 15

$ATTR = \{`skill', `completion\ time', `interruptions'\}$
$ARGS = \{\ A : high\ skill \rightarrow underload, B : high\ completion\ time \rightarrow overload, C : high\ interruptions \Rightarrow B\ \}$
with $A, B \in AR^F$ and $C \in AR^M$
$ATTACKS = \{(A,B),(B,A),(C,A)\}$

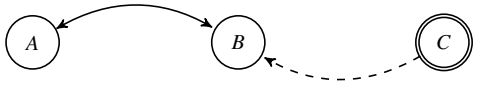


Fig. 4.9: Argumentation graph with forecast, mitigating arguments and rebutting, undercutting attacks

---

4.5 Layer 3 - Reduction of the argumentation graph

The third layer of the argumentative schema of figure 4.2 (page 73) is the first step towards the assessment of a final index of mental workload. In this layer, the knowledge-base of a designer, previously translated into interactive defeasible arguments organised in a graph, has to be activated with objective inputs (the quantification of the workload attributes) gathered from a given user who interacted on a given task. These inputs activate designed arguments with certain degrees of truth, making them more or less credible. As a consequence, the attacks starting from these arguments can also be more or less credible. According to this, few questions arise:

- What are the arguments that should be really accounted in the assessment of mental workload?
- What are the attacks that should be considered valid for assessing mental workload?
- When is an attack, from a less credible attacker to a more credible attacked argument, still valid?

The following sections are aimed at answering these questions by providing formal means for instantiating arguments and attack relations, forming a new graph of interconnected arguments, equal or smaller than the argumentation graph which emerged from the second layer of the schema of figure 4.2 (page 73).

4.5.1 Evaluating the importance of arguments and strength of attacks

It is important to note that the arguments designed so far have two associated notions: degree of truth and importance. The former refers to the activation of the argument, which means the degree to which the argument is instantiated according to the given inputs, evaluated by the membership functions for each attribute in its premise. The latter refers to the importance, if any, of the attributes accounted in the premises, independently of their degree of activation. The former is context-dependent, based on the inputs objectively gathered in a given context. The latter can either be knowledge-dependent, in which case, importance is statically assigned by a workload designer, or user-dependent, in the case where end-users (raters) subjectively express the importance of the workload attributes. The two notions are distinct but both play an important role for the consideration of an argument (and its outgoing and incoming attacks) within an argumentation graph. The issue now is how to consider an argument is strong enough to be part of an argumentation graph and how to determine the minimum strength required by an attack for it to succeed. The proposal here is to use the degree of truth of an argument, as in definition 16 (page 83) for the two problems. In particular, two *reluctancy thresholds* are designed: one for the degree of truth of an argument and one for the degree of truth of the two arguments involved in an attack relation. This notion of reluctancy threshold is similar to the *inconsistency budget* proposed in (Dunne et al., 2011) that allows a finer investigation of the interaction of arguments, generating different solutions. These thresholds respectively indicate how reluctant a designer would be to disregard:

- an argument (and all its outgoing and incoming attacks)
- an attack relation, either rebutting or undercutting.

Definition 22 (Argument reluctancy threshold)
The argument reluctancy threshold $Reluct^{th}_{Arg}$ indicates the minimum degree of truth that an argument must have in order to be activated and thus included in an argumentation graph.

$$Reluct^{th}_{Arg} : [0..1] \in \Re$$

■

A value of 1 indicates that just those arguments with a full degree of truth are activated and included in an argumentation graph. 1 is indeed too strict and restrictive. On the other hand, a value of 0 indicates no reluctancy at all, thus all the arguments will be considered in an argumentation graph, independently of their degree of truth. The application of the argument reluctancy threshold defines a new argument set that is referred to as the *set of activated arguments*.

Definition 23 (Set of activated arguments)
Let Args be a set of arguments and $Reluct^{th}_{Arg}$ the argument reluctancy threshold. The set of activated arguments is:

$$Arg_{act} = \begin{cases} A|A \in Args \wedge (1 \geq A_{deg} \geq Reluct^{th}_{Arg}), & if\ A \in AR^F \\ A|A, B \in Args \wedge (1 \geq A_{deg},\ \ B_{deg} \geq Reluct^{th}_{Arg}), & if\ A : P_1, ..., P_n \Rightarrow B\ \ with\ B \in AR^M \end{cases}$$

■

The first line of the formula refers to those forecast arguments whose degree of truth is greater than the arguments reluctancy threshold and less or equal than the upper limit (1). The second line indicates those mitigating arguments having a degree of truth of the premises and a degree of truth of the attacked argument, both greater than the argument reluctancy threshold, and less than or equal to the upper limit (1).

Given the set of activated arguments, now the issue is to determine which attacks are not strong enough to succeed and hence disregarded. If the degree of truth of the attacker argument is higher than the degree of truth of the attacked argument, then there is no doubt that the attack can be considered a proper attack. Even if the difference of their degree of truth is minimal, the attack makes sense because it was conceptualised by a designer. However, if the attacker has a lower degree of truth than the attacked argument, the issue is now when to consider it a proper attack and when to disregard it. A new threshold is then designed, similar to the argument reluctancy threshold, and referred to as *attack reluctancy threshold*. A value of 0 indicates null reluctance, that means if an attacker has a lower degree of truth than the attacked argument, the attack can still be considered a proper attack, regardless of the difference of degrees of truth between the arguments. In this scenario, every designed attack succeeds because the designer is not reluctant, at all, of the designed attacks. Intermediate values indicate partial reluctancy, so a value of 0.6, for example, indicates that the designer is reluctant for 60%, with the willingness to tolerate an attack from an argument with a lower degree of truth to an argument with an higher degree of truth, if and only if the difference of their degrees of truth is less than or equal to $1 - 0.6 = 0.4$. In the other case, the attack is disregarded because the attacker has not enough high degree of truth (not enough credibility) to perform the attack. A value of 1 indicates total reluctance, so the designer is not willing, at all, to tolerate an attack from an argument with a lower degree of truth to an argument with an higher degree of truth.

Definition 24 (Attack reluctancy threshold)
The attack reluctancy threshold $Reluct^{th}_{Att}$ indicates the reluctancy to tolerate an attack from a less to a more credible argument.

$$Reluct^{th}_{Att} : [0..1] \in \Re$$ ■

The application of the attack reluctancy threshold along with the degree of truth of arguments define a new attack set that is referred to as *set of activated attacks*. This set contains attacks that coincide with the notion of *defeats* (as reviewed in section 3.4, page 57). In other words, this set is composed by all those attacks that logically succeed and are not disregarded because they are considered too weak and not credible.

Definition 25 (Set of activated attacks)
Let Arg_{act} the set of activated arguments, $Atts$ the set of attack relations and $Reluct^{th}_{Att}$ the attack reluctancy threshold. The set of activated attacks is defined as:

$$Attack_{act} : \Big\{ \{(A,B) \mid (A,B) \in Atts \ \wedge \ A,B \in Arg_{act} \quad if\, A_{deg} \geq B_{deg} \ \vee \ 0 \leq Abs(A_{deg} - B_{deg}) < 1 - Reluct^{th}_{Att}$$

with Abs the absolute function. ■

This set contains all the attacks that have not been disregarded because either the attacker has a higher degree of truth than the attacked argument (so its attack is valid as designed), or because the attacker's degree of truth is not credible enough to perform the attack (its degree of truth is much lower than the attacked argument's degree of truth). Applications of the two thresholds can be found in example 16.

Example 16

- Forecast arguments: $A_{deg} = 0.8$, $B_{deg} = 0.9$
- Mitigating arguments: $C_{deg} = 0.3$, $D_{deg} = 0.2$
- $Reluct^{th}_{Arg} = 0.25$
- rebutting attack: (A, B), (B, A)
- undercutting attack: (C, A), (D, B)
- $Reluct^{th}_{Att} = 0.15$

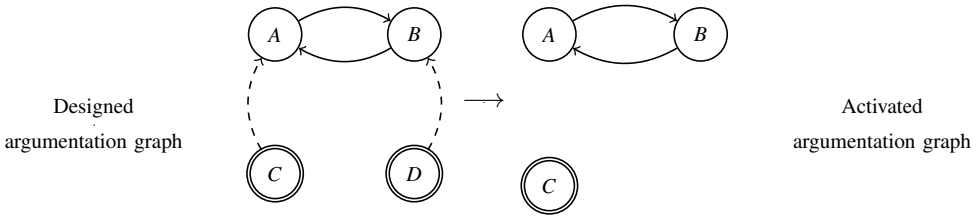


Fig. 4.10: Example of activated arguments and attack relations

The arguments with a degree of truth greater than 0.25 are A, B, C, and so, they are activated. A, although having a lower degree of truth than B, can still attack B because its degree of truth is strong enough compared to B's degree of truth ($Abs(0.8-0.9) < 0.15$). B has a higher degree of truth than A so its attack towards A is a proper attack as designed. C's attack towards A is disregarded because A's degree of truth is much stronger than C's degree of truth ($Abs(0.3-0.8) > 0.15$).

Given the set of activated arguments as well as the set of activated attack relations, the third step of the multi-layer argumentative schema of figure 4.2 (page 73) is complete. From this step, a new argumentation graph emerges, this being a strict sub-set of the graph which emerged from layer 2. This new argumentation graph coincides with the portion of the knowledge-base of a designer objectively activated with numerical inputs, gathered from a given user, in a given context, and with a given task. This sub-argumentation framework can now be evaluated by applying acceptability semantics in order to extract consistent and conflict-free extensions of arguments, and eliminating the inconsistencies that might arise from their interaction.

4.6 Layer 4 - Extraction of credible extensions

In order to investigate the potential inconsistencies that might emerge from the interaction of activated arguments, acceptability semantics (as described in section 3.5.3, page 62), can be applied. The skeptical approach of the grounded semantics, as proposed by (Dung, 1995), always returns one extension that can be empty. The more credulous approach of the preferred semantics, instead, might return a set of extensions that can be seen as different but reasonable points of view from which mental workload can be assessed. In this last case, the issue is how to select the most credible extension, and so a strategy for quantifying the strength (credibility) of each computed preferred extension, is needed. In the following section, such a quantification strategy is introduced.

4.6.1 Computing strength of acceptable extensions

Preferred semantics can produce one or more extensions (set of arguments). In the case where just one extension is produced, this coincides with the grounded extension, thus there is no need to compute the strength of that extension. However, in the case where multiple extensions of arguments are computed (as per definition 10, page 66), quantification of strength is necessary to decide which extension is the most credible. Here, it is argued that the cardinality of an extension (set of arguments) is an important factor to consider for computing its credibility. Intuitively, an extension with a higher cardinality might be seen as more credible than extensions with lower cardinality, as it contains more pieces of evidence that are consistent with each other (an extension is a conflict-free set of arguments) and that support the same claim (workload dichotomy). However, considering just the cardinality might be reductive in the case where, for instance, an extension with several arguments has a combined degree of truth lower than an extension with fewer arguments. For these reasons the proposal is to use the cardinality of an extension jointly with the degree of truth of its arguments, for the quantification of its credibility.

Definition 26 (Strength of acceptable extension)
Let Arg_{act} be the set of activated arguments and E an acceptable extension, as computed by an acceptability semantics. The strength of E is defined as:

$$E_{Strength} : \frac{Card(E)}{Card(Arg_{act})} + \frac{1}{Card(E)} \sum_{Arg \in E} Arg_{deg}$$

with Card the cardinality function.

■

The strength of an acceptable extension is the combination of its cardinality (compared to the cardinality of the set of activated arguments) with the average of the degrees of truth of its arguments. The cardinality of the extension is divided by the cardinality of the activated arguments set so it can be normalised on the same scale as the average of degrees of truth of arguments, which are both in the range $[0..1] \in \Re$. Once the strength of each extension is computed, the strongest extension can be selected or, in the less probable case of multiple equally stronger extensions, a representative crisp index for overall mental workload can be computed. In the following section, such a computation is introduced and this computation represents the last step for mental workload assessment.

4.7 Layer 5 - Assessment of mental workload

Given a set of extensions, as computed by acceptability semantics, the final step towards mental workload assessment is the identification of the strongest extension, meaning the one that maximises cardinality of arguments and their degrees of truth. The strongest extension should be unique, but there might be the case that multiple, equally stronger extensions are computed. In this case, they are all taken into account to assess a final crisp index of mental workload. According to definition 4.4.1 (page 84), two typologies of arguments have been designed: forecast and mitigating arguments. Both of them can exist within the strongest computed extension/s, however, just forecast arguments have a claim (workload dichotomy) that can be taken into account to infer an overall mental workload index. Mitigating arguments already played their role, contributing to the identification of the acceptable extensions. In order to infer an index of mental workload of a single forecast argument, the proposal is to project the degree of truth of the premise of that argument to its conclusion using the notion of *logical consequence*. Specifically, the degree of truth of the argument, as per definition 16 (page 83) is used as the input of the workload dichotomy associated with the conclusion of that argument.

In addition, each argument in the strongest extension/s might either have an associated numerical importance, or an 'undefined' importance, according to definition 21 (page 87). As a consequence two sub-sets of arguments within a stronger extension might occur: arguments with 'defined' or 'undefined' importance. The former set contains arguments with a numerical value whose aim is to give importance to the computed degree of truth of the argument itself, while the latter contains arguments where importance is undefined, thus their computed degree of truth cannot be weighted, and it should intuitively be taken as it is. However, this intuitive way of operating is equivalent to associating a value of importance of 1 to those arguments with undefined value. In turn, this is equivalent to say that arguments with 'undefined' importance are always more important than or equally as important as those arguments with an associated importance value, which is always less than or equal to 1 ($[0..1] \in \Re$). Clearly, this is not accurate and precise, being counterintuitive. To handle this issue and mitigate the aforementioned effect, the proposal is to weight the two sub-sets of arguments (defined and undefined importance) by their respective cardinality. If the set 'defined' is thus bigger than the set 'undefined', its internal arguments contribute more to the final computation of overall mental workload than the arguments belonging to the set 'undefined'. These considerations can be summarised as in the following formal definition.

Definition 27 (Overall mental workload index)

Let AE a set containing the n acceptable extensions computed by an acceptability semantics, and SE the set containing the strongest extension/s

$$SE = \{A \mid A \in AE,\ A_{Strength} = max(E^1_{Strength}, ..., E^n_{Strength})\ with\ E^1, ..., E^n \in AE\}$$

The overall mental workload index is:

$$MWL = [0..100] \in \Re$$

$$MWL = \frac{\sum_{A \in SE} \left(\frac{Card(Undef)}{Card(A)} \cdot \frac{\sum_{arg \in Undef} arg_c(arg_{deg})}{Card(Undef)} + \frac{Card(Def)}{Card(A)} \cdot \frac{\sum_{arg \in Def} arg_c(arg_{deg}) \cdot arg_{imp}}{\sum_{arg \in Def} arg_{imp}} \right)}{Card(SE)}$$

with arg_c a workload dichotomy associated to the conclusion of each forecast argument within the extension $A \in SE$, Card the cardinality function and $Def, Undef \subseteq A$ respectively the subsets of arguments with defined and undefined importance. ■

The computation of the overall index of mental workload lies in the range [0..100] in line with definition 12 (page 78). This is the last step of the multi-layer schema (figure 4.2, page 73) and it represents the final mental workload assessment.

The following chapter is firstly devoted to the implementation of the framework for mental workload representation and assessiment by describing the pseudo-code for mental workload assessments, in line with the definitions provided in this chapter. Practical uses of the framework follow, demonstrating how the NASA-TLX assessment procedure (Hart and Staveland, 1988), as well as the Workload Profile WP instrument (Tsang and Velazquez, 1996), can be translated into two particular instances of the framework itself.

Chapter 5

Implementation and instantiation

The defeasible framework designed in chapter 4 has been implemented using the programming language Java. Although the actual code is not provided, the pseudo-code describing the overall methodology for assessing mental workload is described in this chapter as well as the instantiation of the framework. In detail, two state-of-the-art subjective mental workload assessment techniques are translated into two instances of the framework, demonstrating its practical usage. The assessment of mental workload of these two particular instances will be evaluated in the next chapter.

5.1 The defeasible framework as a formal tuple

As designed in chapter 4, the defeasible framework for human mental workload assessment can be summarised as a 9-tuple:

$$DEF-MWL: < ATTR,\ f_{Pref},\ MF,\ RL,\ DMF,\ ARGS,\ ATTACKS,\ RT,\ INPUTS >$$

where

- **ATTR** is a finite set of workload attributes that a designer wishes to consider in the representation of the construct of mental workload. They are usually expressed in natural language proposition and $ATTR$ contains the representative labels.
- $f_{\textbf{Pref}}$ is the partial function that assigns importance values to the attributes in $ATTR$, as per definition 20 (page 87), thus delineating preferences of the attributes themselves. Each attribute in $ATTR$ has at most one importance value and if $f_{Pref}(a) \leq f_{Pref}(b)$ then attribute b is preferred or is equally important than attribute a. Transitivity applies to the importance of attributes so if $f_{Pref}(a) \leq f_{Pref}(b) \leq f_{Pref}(c)$ then c is preferred or has equal importance than a.
- **MF** is the set of membership functions defined for each attribute in $ATTR$. Each attribute can be described by different membership functions and the same membership functions can also be associated to other attributes as per definition 11 (page 77).
- **RL** are the two redlines: $RedLine^{fitting}_{underload}$ *and* $RedLine^{overload}_{fitting} \in [0..100]\aleph$ with $RedLine^{fitting}_{underload} \leq RedLine^{overload}_{fitting}$ as per definition 13 (page 79). These are boundaries used for partitioning the overall mental workload assessment space into 4 dichotomies.

- **DMF** is the set containing the strict monotonic functions used for modelling the four workload dichotomies that are separated by the redlines as per definition 14 (79).
- **ARGS** is a finite set of defeasible arguments built according to a designer's knowledge-base. They can be forecast or mitigating as per definitions 15 and 17 (pages 80 and 85). Both have a set of premises, built upon some attribute in $ATTR$ (and modelled with some membership functions in MF). The former links premises to a workload dichotomy in DMF, while the latter links premises to the negation of another argument either forecast or another mitigating argument.
- **ATTACKS** is a finite set of attack relations among arguments in *ARGS*. It is a binary relation defined on $ARGS \times ARGS$. Attack can be rebutting as per definition 18 (page 85) or undercutting/undermining as per definition 19 (page 85). The former occurs when two forecast arguments logically contradict each other, while the latter occur when a mitigating argument undermines the justification of a forecast or another mitigating argument.
- **RT** are the two reluctancy thresholds $Reluct^{th}_{Arg}$ and $Reluct^{th}_{Att} : [0..1] \in \Re$ as per definition 22 and 24 (page 90). They are used to activate a sub-set of arguments in *ARGS* and a sub-set of attack relations in $ATTACKS$.
- **INPUTS** is a finite set of input values in $[0..100] \in \aleph$, one for each of the attribute in $ATTR$. $Card(INPUTS) = Card(ATTR)$: one and only one input value exists for each defined attribute in $ATTR$.

The first step of the algorithm towards the assessment of mental workload is to activate some or all of the arguments in *ARGS* by using the values in the $INPUTS$. These values are evaluated by the membership functions associated to the premises of each argument producing a degree of truth. All the arguments that satisfy definition 23 (page 90), and meet the *argument reluctancy threshold* ($Reluct^{th}_{Arg} \in RT$), are activated, generating a finite *set of activated arguments* (Arg_{act}).

The second step is to activate attack relations. Given the finite set of activated arguments Arg_{act}, their degree of truth as per definition 16 (page 83), the set of attacks relations $ATTACKS$ as well as the *attack reluctancy threshold* $Reluct^{th}_{Att} \in RT$, those arguments that satisfy definition 25 (page 91) are activated, generating the *set of activated attacks* $Attack_{act}$. This is a finite set containing those attacks that have not been disregarded and that are considered valid.

Thirdly, given the finite set of activated arguments Arg_{act} and the finite set of activated attacks $Attack_{act}$, an *abstract argumentation graph* $< Arg_{act}, Attack_{act} >$ can be defined (in line of Dung's proposal, as in section 3.5.2 and definition 1, page 62). This framework is a graph of interconnected arguments that is abstractly evaluated (not considering the internal representation of arguments) by running the *grounded and the preferred acceptability semantics*, as per definitions 8 and 10 (page 65). These acceptability semantics produce a set *acceptable extensions of arguments*: one by the grounded semantics (it can be empty) and one or more by the preferred semantics. Extensions are conflict-free sub-sets of arguments in Arg_{act} that are seen as coherent points of view that can be employed for mental workload assessment.

The subsequent step, given a finite set of acceptable extensions of arguments, is to compute their strength according to definition 26 (page 92) and extract the *strongest extension/s*. Eventually, the last step is to generate a crisp index, the defeasible assessment of overall mental workload, from the strongest extension/s. Given the arguments in the strongest extension/s, the overall mental workload index is computed by using the strategy proposed in definition 27 (page 94).

5.1.1 Pseudo-code of the algorithm towards mental workload assessment

For clarification purpose, the pseudo-code of the algorithm towards the assessment of overall mental workload is presented in figure 5.1.

---

1 For each A^i in ARG (A^i of the form $P_1\ \Phi\ P_2\ \Phi,...,\Phi\ P_n \rightarrow c$)

2 For each $mf_x^j \in MF$ built upon attribute $x \in ATTR$ and associated to premise P^j of argument A^i

3 compute degree of truth of premise P^j with input $I_x \in INPUTS$ related to attribute $x : degTruthP^j = mf_x^j(I_x)$

4 Compute degree of truth of argument A^i aggregating all the degrees of truth of each premise $A^i_{deg} = degTruthP^1\ \Phi^1\ ,...,\Phi^n\ degTruthP^n$

5 If $A^i_{deg} \geq Reluct^{th}_{Arg}$ then $A^i \in Arg_{Act}$

6 For each $(A,B) \in ATTACKS$

7 If $A,B \in Arg_{Act}$ then

8 $if(A_{deg} \geq B_{deg}\ \vee\ 0 \leq Abs(A_{deg} - B_{deg}) < 1 - Reluct^{th}_{Att})$ then $(A,B) \in Attack_{act}$

10 Execute acceptability semantics (Grounded and/or Preferred) on abstract argumentation graph $< Arg_{Act},\ Attack_{Act} >$

11 For each extension E^i computed by acceptability semantics

12 compute strength: $E^i_{Strength} = \frac{n}{Card(Arg_{act})} + \left(\frac{1}{n}\sum_{j=1}^{n} Arg^j_{deg}\right)$ with $Arg^j \in E^i$

13 Compute strongest extension(s) set: $SE = \{E^i |\ E^i_{Strength} = max(E^1_{Strength},..,E^n_{Strength})\}$

14 Compute overall mental workload with arguments in each $E^i \in SE$

$$MWL = \frac{\sum_{A \in SE}\left(\frac{Card(Undef)}{Card(A)} \cdot \frac{\sum_{arg \in Undef} arg_c(arg_{deg})}{Card(Undef)} + \frac{Card(Def)}{Card(A)} \cdot \frac{\sum_{arg \in Def} arg_c(arg_{deg}) \cdot arg_{imp}}{\sum_{arg \in Def} arg_{imp}}\right)}{Card(SE)}$$

Fig. 5.1: Pseudo-code of the algorithm for mental workload assessment

5.2 Instantiation of the framework

In the following sections practical instantiations of the framework designed in chapter 4 are described. Specifically, it is demonstrated how two well-known subjective mental workload assessment techniques can be translated into two particular instances of the framework, following step-by-step the multi-layer schema of figure 4.2)page 73). These are the following:

- The Nasa Task Load Index Hart and Staveland (1988)
- The Workload Profile Tsang and Velazquez (1996)

5.2.1 Translating NASA-TLX as an instance of the framework

Layer 1: translation of knowledge-base

The NASA-TLX mental workload assessment procedure is based upon six attributes: mental (MD), physical (PD) and temporal demand (TD), effort (EF), performance (PE) and frustration (FR). These are the workload attributes believe to influence mental workload and form the finite set of attributes $ATTR$. The membership functions of figure 5.2 are designed to partition each of the six attributes in four areas (Low, Medium lower/upper and High).

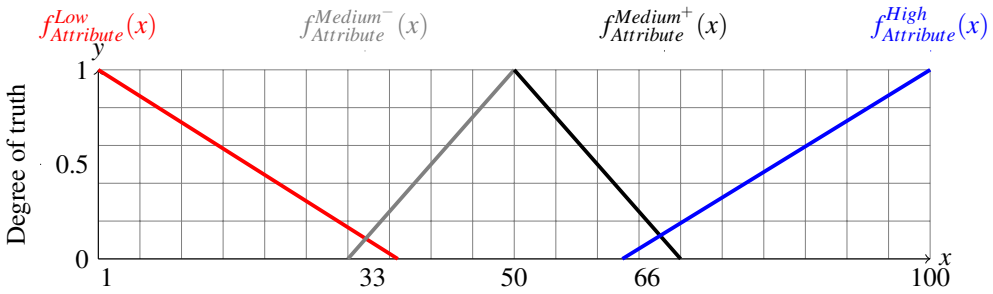


Fig. 5.2: A NASA-TLX defeasible translation: membership functions for every attribute

As described in chapter 2 (page 23), in the original NASA-TLX instrument a form of preferentiality of workload attributes is accounted. This is manifested by a pair-wise comparison of the 6 attributes (appendix A.1, page 173) generating 15 preferences and thus defining a preference rank of the 6 attributes themselves. Let us suppose an end-user has expressed the following preferences among the attributes, as conceived in the pair-wise comparison procedure of the original NASA-TLX instrument:

- MD: 5 preferences;
- PH: 0 preferences;
- TD: 4 preferences;
- EF: 3 preferences;
- PE: 2 preferences;
- FR: 1 preferences;

These preferences need to be re-scaled within the range $[0..1] \in \Re$ according to definition 20 (page 87).

Thus the partial function that returns the importance of each attribute is as it follows:

$$f_{pref}(x) = \begin{cases} 1.0 & \text{if x='MD'} \\ 0.0 & \text{if x='PH'} \\ 0.8 & \text{if x='TD'} \\ 0.6 & \text{if x='EF'} \\ 0.4 & \text{if x='PE'} \\ 0.2 & \text{if x='FR'} \end{cases}$$

Note 6

In the case of the NASA-TLX, all the attributes have an associated preference, derived from the pairwise comparisons of the original instrument. As a consequence, the function f_{pref} is a full-defined function, always returning an importance value for any designed workload attribute.

The functions of figure 5.3 model the four workload dichotomies, according to definition 14 (page 79), divided by the following redlines:

- $RedLine^{fitting}_{underload} = 33$
- $RedLine^{overload}_{fitting} = 66$

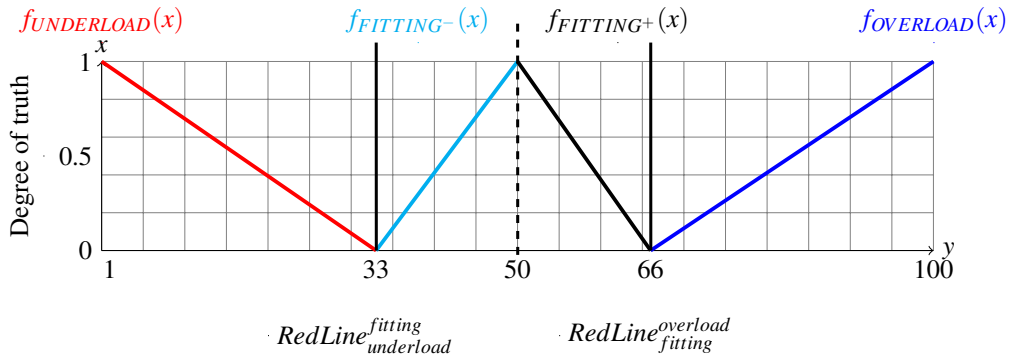


Fig. 5.3: A NASA-TLX defeasible translation: workload dichotomies partitioned by redlines

The forecast arguments that might be designed for the NASA-TLX are those expressed in table 5.1, both in natural language and formally.

Layer 2: construction of argumentation graph

Note 7

The attribute 'Performance' has an inversely proportional effect on mental workload, in line with the original NASA-TLX instrument: decrements in performance correspond to increments in workload. This is reflected in arguments Q, R, S, T. All the other arguments have a directly proportional effect on workload. All the arguments in natural language (column 2) imply a certain workload level, thus they are considered forecast and correctly formalised (column 3) in line with definition 15 (page 80).

Attribute $\in ATTR$	Arguments (natural language)	Arguments (formal)
Mental demand (MD)	A: low MD *implies* underload	A: low MD $\rightarrow f_{UNDERLOAD}$
	B: medium low MD *implies* low fit MWL	B: medium low MD $\rightarrow f_{FITTING^-}$
	C: medium high MD *implies* high fit MWL	C: medium high MD $\rightarrow f_{FITTING^+}$
	D: high MD *implies* overload	D:high MD $\rightarrow f_{OVERLOAD}$
Physical demand (PD)	E: low PD *implies* underload	E: low MD $\rightarrow f_{UNDERLOAD}$
	F: medium low PD *implies* low fit MWL	F: medium low PD $\rightarrow f_{FITTING^-}$
	G: medium high PD *implies* high fit MWL	H: medium high PD $\rightarrow f_{FITTING^+}$
	H: high PD *implies* overload	H:high PD $\rightarrow f_{OVERLOAD}$
Temporal demand (TD)	I: low TD *implies* underload	I: low TD $\rightarrow f_{UNDERLOAD}$
	J: medium low TD *implies* low fit MWL	J: medium low TD $\rightarrow f_{FITTING^-}$
	K: medium high TD *implies* high fit MWL	K: medium high TD $\rightarrow f_{FITTING^+}$
	L: high TD *implies* overload	L:high TD $\rightarrow f_{OVERLOAD}$
Effort (EF)	M: low EF *implies* underload	M: low EF $\rightarrow f_{UNDERLOAD}$
	N medium low EF *implies* low fit MWL	N: medium low EF $\rightarrow f_{FITTING^-}$
	O: medium high EF *implies* high fit MWL	O: medium high EF $\rightarrow f_{FITTING^+}$
	P: high EF *implies* overload	P:high EF $\rightarrow f_{OVERLOAD}$
Performance (PE)	Q: low PE *implies* overload	Q: low PE $\rightarrow f_{OVERLOAD}$
	R: medium low PE *implies* high fit MWL	R: medium low PE $\rightarrow f_{FITTING^+}$
	S: medium high PE *implies* low fit MWL	S: medium high PE $\rightarrow f_{FITTING^-}$
	T: high PE *implies* underload	T:high PE $\rightarrow f_{UNDERLOAD}$
Frustration (FR)	U: low FR *implies* underload	U: low FR $\rightarrow f_{UNDERLOAD}$
	V: medium low FR *implies* low fit MWL	V: medium low FR $\rightarrow f_{FITTING^-}$
	W: medium high FR *implies* high fit MWL	W: medium high FR $\rightarrow f_{FITTING^+}$
	X: high FR *implies* overload	X:high FR $\rightarrow f_{OVERLOAD}$

Table 5.1: The NASA-TLX defeasible translation: natural language and formal arguments

In the original NASA-TLX instrument no logical relationship of attributes is taken into account, as a consequence no interaction of arguments can be designed and, in turn, no rebutting or mitigating attack can be defined. Indeed the model might be extended in the case a designer is aware of some theoretical relationship among the six original workload attributes. The resulting argumentation graph containing arguments and attacks (none) is depicted in figure 5.4.

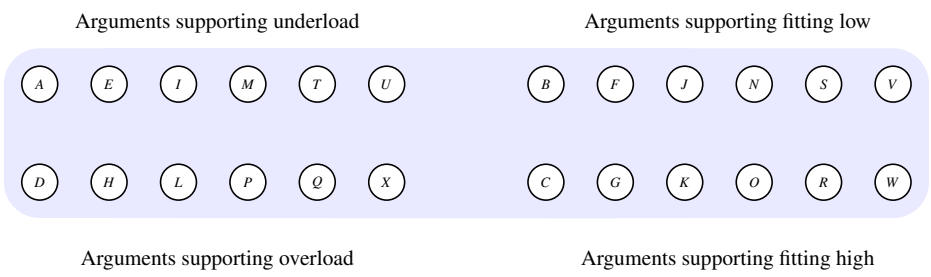


Fig. 5.4: The NASA-TLX defeasible translation: the argumentation graph (with no attack)

Given $f_{pref}(x)$, the importance of each argument can be computed, according to definition 21 (page 87). As each premises of each designed argument contains one and only one attribute, the importance of an argument coincides exactly with the importance of each attribute:

- $A_{imp} = B_{imp} = C_{imp} = D_{imp} = 1$;
- $E_{imp} = F_{imp} = G_{imp} = H_{imp} = 0$;
- $I_{imp} = J_{imp} = K_{imp} = L_{imp} = 0.8$;
- $M_{imp} = N_{imp} = O_{imp} = P_{imp} = 0.6$;
- $Q_{imp} = R_{imp} = S_{imp} = T_{imp} = 0.4$;
- $U_{imp} = V_{imp} = W_{imp} = X_{imp} = 0.2$;

Let us suppose that an end-user has answered the questions associated to the NASA-TLX, as in appendix A.1, producing the following rates (in the scale 0 to 100):

- MD: 90;
- PD: 0;
- TD: 70;
- EF: 60;
- PE: 72;
- FR: 15;

The degrees of truth of the arguments, according to definition 16 (page 83), using the membership functions defined in figure 5.2, are listed in table 5.2.

Attribute $\in ATTR$	Argument's degree of truth	
Mental demand (MD)	$A_{deg} = f_{MD}^{Low}(90) = 0$	$B_{deg} = f_{MD}^{Medium^-}(90) = 0$,
	$C_{deg} = f_{MD}^{Medium^+}(90) = 0$	$D_{deg} = f_{MD}^{High}(90) = 0.66$
Physical demand (PD)	$E_{deg} = f_{PD}^{Low}(0) = 1$	$F_{deg} = f_{PD}^{Medium^-}(0) = 0$,
	$G_{deg} = f_{PD}^{Medium^+}(0) = 0$	$H_{deg} = f_{PD}^{High}(0) = 0$
Temporal demand (TD)	$I_{deg} = f_{TD}^{Low}(70) = 0$	$J_{deg} = f_{TD}^{Medium^-}(70) = 0$,
	$K_{deg} = f_{TD}^{Medium^+}(70) = 0.33$	$L_{deg} = f_{TD}^{High}(70) = 0$
Effort (EF)	$M_{deg} = f_{EF}^{Low}(60) = 0$	$N_{deg} = f_{EF}^{Medium^-}(60) = 0$,
	$O_{deg} = f_{EF}^{Medium^+}(60) = 0.66$	$P_{deg} = f_{EF}^{High}(60) = 0$
Performance (PE)	$Q_{deg} = f_{PE}^{Low}(72) = 0$	$R_{deg} = f_{PE}^{Medium^-}(72) = 0$,
	$S_{deg} = f_{PE}^{Medium^+}(72) = 0.26$	$T_{deg} = f_{PE}^{High}(72) = 0.06$
Frustration (FR)	$U_{deg} = f_{FR}^{Low}(15) = 0.5$	$V_{deg} = f_{FR}^{Medium^-}(15) = 0$,
	$W_{deg} = f_{FR}^{Medium^+}(15) = 0$	$X_{deg} = f_{FR}^{High}(15) = 0$

Table 5.2: The NASA-TLX defeasible translation: degree of truth of arguments

Layer 3: reduction of argumentation graph

Let us suppose a designer is willing to initialise the argument and attack reluctancy thresholds as:

- $Reluct^{th}_{Arg} = 0$. Willingness to consider all those arguments whose degree of truth is greater than 0 (def. 22).
- $Reluct^{th}_{Att} = 0$. Willingness to tolerate an attack from a less to a more credible argument - no reluctancy (def. 24).

The set of activated arguments Arg_{act}, computed according to definition 23 (page 90) is: {D, E, K, O, S, T, U} The set of activated attack relations $Attack_{act}$, according to definition 25 (page 91), is empty, as no designed attack relation exists. Thus the reduced argumentation graph on which Dung-style grounded and preferred acceptability semantics, as described in section 3.5.3, can be executed, is the one depicted in figure 5.5.

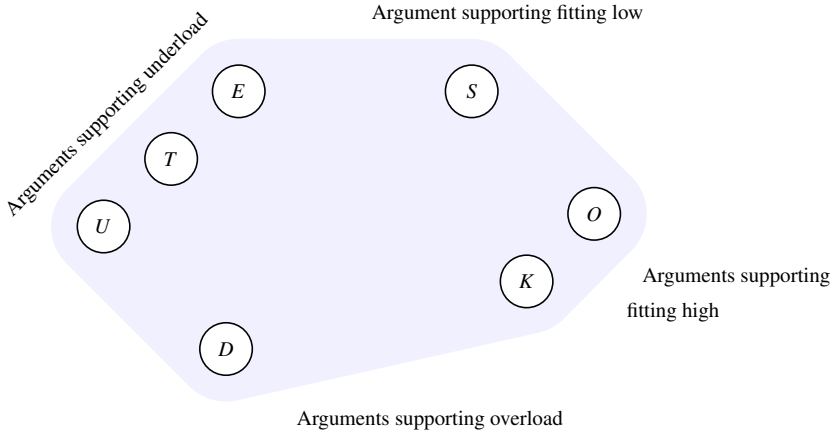


Fig. 5.5: The NASA-TLX defeasible translation: reduced argumentation graph

The grounded extension produced by the grounded acceptability semantics (definition 8, page 65) coincides exactly with the original abstract framework of figure 5.5, thus containing all the arguments (D, E, K, O, S, T, U). The preferred extensions produced by the preferred acceptability semantics (definition 10, page 66), in this case coincides with the grounded extension:

- E^1: D, E, K, O, S, T, U (grounded extension = preferred extension)

Layer 4: extraction of credible extensions

The strength of the unique extension computed according to definition 26 (page 92) is:

$$E^1_{Strength} = \frac{Card(E^1)}{Card(Arc_{act})} + \frac{D_{deg} + E_{deg} + K_{deg} + O_{deg} + S_{deg} + T_{deg} + U_{deg)}}{Card(E_1)}$$
$$= \frac{7}{7} + \frac{0.66 + 1 + 0.33 + 0.66 + 0.26 + 0.06 + 0.5}{7} = 0.495$$

E^1 is also the strongest unique extension thus the overall index of mental workload is computed considering the forecast arguments within this extension, according to definition 27 (page 94). The equations of the straight lines associated to the four workload dichotomies of figure 5.3 are:

- $f_{underload} : [0..1] \in \Re \rightarrow [1..32] \in \Re$ $\quad f_{underload}(x) = -31(x-1)+1$
- $f_{fitting^-} : [0..1] \in \Re \rightarrow [33..49] \in \Re$ $\quad f_{fitting^-}(x) = 16x+33$
- $f_{fitting^+} : [0..1] \in \Re \rightarrow [50..66] \in \Re$ $\quad f_{fitting^+}(x) = -16(x-1)+50$
- $f_{overload} : [0..1] \in \Re \rightarrow [67..100] \in \Re$ $\quad f_{overload}(x) = 23x+77$

Layer 5: assessment of mental workload

The unique stronger extension carries just forecast arguments and as each of them has an associated preference, the overall mental workload score can be computed just using the right part (within big brackets) of the formula of definition 27 (page 94) as it follows:

$$MWL = \frac{Card(def)}{Card(E^1)} \cdot \frac{\begin{array}{c} D_c(D_{deg}) \cdot D_{imp} + E_c(E_{deg}) \cdot E_{imp} + K_c(K_{deg}) \cdot K_{imp} + O_c(O_{deg}) \cdot O_{imp} + \\ S_c(S_{deg}) \cdot S_{imp} + T_c(T_{deg}) \cdot T_{imp} + U_c(U_{deg}) \cdot U_{imp} \end{array}}{D_{imp} + E_{imp} + K_{imp} + O_{imp} + S_{imp} + T_{imp} + U_{imp}}$$

$$= \frac{7}{7} \cdot \frac{\begin{array}{c} (f_{overload}(0.66) \cdot 1) + (f_{underload}(1) \cdot 0) + (f_{fitting^+}(0.33) \cdot 0.8) + (f_{fitting^+}(0.66) \cdot 0.6) + \\ (f_{fitting^-}(0.26) \cdot 0.4) + (f_{underload}(0.06) \cdot 0.4) + (f_{underload}(0.5) \cdot 0.2) \end{array}}{1+0+0.8+0.6+0.4+0.4+0.2}$$

$$= \frac{(92.18 \cdot 1) + (1 \cdot 0) + (60.72 \cdot 0.8) + (55.44 \cdot 0.6) + (37.66 \cdot 0.4) + (29.14 \cdot 0.4) + (16.5 \cdot 0.2)}{3.4} = 60.00$$

The overall mental workload score is 60.00 which clearly falls within the workload dichotomy $fitting^+$, closer to the redline that separates it with the dichotomy overload rather than the optimal middle point of 50. A designer can interpret this workload score positively, and if a system under examination has to be tested, it imposed an optimal workload (tending to high) on single the tested end-user. The above assessment can be repeated several times with different users in order to achieve a better and more robust indication of the actual workload imposed by the system under evaluation. In the case the average of the computed workload indexes of all the tested subjects falls significantly within the dichotomies *underload* or *overload*, the designer might structurally modify the system and repeat the workload assessment procedure. In turn, if the new changes will bring the new average of workload indexes within the optimal ranges (Fitting lower or upper dichotomies), then the system can be considered workload-optimised to end-user interaction.

5.2.2 Translating WP as an instance of the defeasible framework

Layer 1: translation of knowledge-base

The Workload Profile (WP)assessment procedure (Tsang and Velazquez, 1996) is based upon 8 attributes built upon the multiple-resource model of Wickens (Wickens, 2008; Wickens and Hollands, 1999).

- processing stage: perceptual/central (SPPC)
- processing stage: response (SPR)
- processing code: spatial (CPS)
- processing code: verbal (CPV)
- input: visual (IV)
- input: auditory (IA)
- output: manual (OM)
- output: speech (OS)

These attributes form the set $ATTR$ and each of them is modelled with the membership function of figure 5.2. In contrast to the NASA-TLX instrument, in the Workload Profile procedure, preferentiality of attributes is not taken into account, so each attribute has an 'undefined' importance. It turns out that the partial function that returns the importance of each attribute returns always an undefined value, as per definition 20 (page 87).

$$f_{pref}(x) = \Big\{ undefined \quad \forall x$$

The functions of figure 5.3 are used for the four workload dichotomies divided by redlines. All these functions as well as the two redlines are the same as the ones employed in the translation of the NASA-TLX instrument into a defeasible instance (5.2.1).

Layer 2: construction of argumentation graph

The forecast arguments that might be designed for the WP are those expressed in table 5.3, both in natural language and formally.

> **Note 8**
> All the designed arguments have a directly proportional effect on mental workload.

In the original WP workload assessment instrument, no theoretical relationship between attributes is considered, as a consequence no interaction between arguments can be designed and thus no rebutting or mitigating attack can be defined. Indeed the model might be extended in the case a designer is aware of further theoretical relationships among the eight attributes. The resulting graph containing arguments and attacks (none), is depicted in figure 5.6.

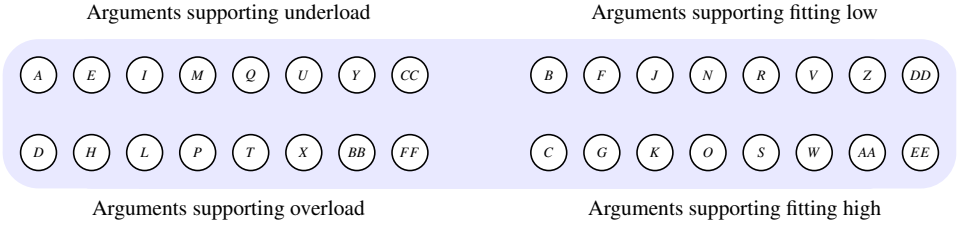


Fig. 5.6: The Workload Profile defeasible translation: arguments framework (with no attack)

Attribute $\in ATTR$	Arguments (natural language)	Arguments (formal)
processing stage: perceptual/central (SPPC)	A: low SPPC *implies* underload B: medium low SPPC *implies* low fit MWL C: medium high SPPC *implies* high fit MWL D: high SPPC *implies* overload	A: low SPPC $\rightarrow f_{UNDERLOAD}$ B: medium low SPPC $\rightarrow f_{FITTING^-}$ C: medium high SPPC $\rightarrow f_{FITTING^+}$ D: high SPPC $\rightarrow f_{OVERLOAD}$
processing stage: response (SPPC)	E: low SPR *implies* underload F: medium low SPR *implies* low fit MWL G: medium high SPR *implies* high fit MWL H: high SPR *implies* overload	E: low SPR $\rightarrow f_{UNDERLOAD}$ F: medium low SPR $\rightarrow f_{FITTING^-}$ G: medium high SPR $\rightarrow f_{FITTING^+}$ H: high SPR $\rightarrow f_{OVERLOAD}$
processing code: spatial (CPS)	I: low CPS *implies* underload J: medium low CPS *implies* low fit MWL K: medium high CPS *implies* high fit MWL L: high CPS *implies* overload	I: low CPS $\rightarrow f_{UNDERLOAD}$ J: medium low CPS $\rightarrow f_{FITTING^-}$ K: medium high CPS $\rightarrow f_{FITTING^+}$ L: high CPS $\rightarrow f_{OVERLOAD}$
processing code: verbal (CPV)	M: low CPV *implies* underload N: medium low CPV *implies* low fit MWL O: medium high CPV *implies* high fit MWL P: high CPV *implies* overload	M: low CPS CPV $f_{UNDERLOAD}$ N: medium low CPV $\rightarrow f_{FITTING^-}$ O: medium high CPV $\rightarrow f_{FITTING^+}$ P: high CPV $\rightarrow f_{OVERLOAD}$
input: visual (IV)	Q: low IV *implies* underload R: medium low IV *implies* low fit MWL S: medium high IV *implies* high fit MWL T: high IV *implies* overload	Q: low IV $f_{UNDERLOAD}$ R: medium low IV $\rightarrow f_{FITTING^-}$ S: medium high IV $\rightarrow f_{FITTING^+}$ T: high IV $\rightarrow f_{OVERLOAD}$
input: auditory (IA)	U: low IA *implies* underload V: medium low IA *implies* low fit MWL W: medium high IA *implies* high fit MWL X: high IA *implies* overload	U: low IA $f_{UNDERLOAD}$ V: medium low IA $\rightarrow f_{FITTING^-}$ W: medium high IA $\rightarrow f_{FITTING^+}$ X: high IA $\rightarrow f_{OVERLOAD}$
output: manual (OM)	Y: low OM *implies* underload Z: medium low OM *implies* low fit MWL AA: medium high OM *implies* high fit MWL BB: high OM *implies* overload	Y: low OM $f_{UNDERLOAD}$ Z: medium low OM $\rightarrow f_{FITTING^-}$ AA: medium high OM $\rightarrow f_{FITTING^+}$ BB: high OM $\rightarrow f_{OVERLOAD}$
output: speech (OS)	CC: low OS *implies* underload DD: medium low OS *implies* low fit MWL EE: medium high OS *implies* high fit MWL FF: high OS *implies* overload	CC: low OS $f_{UNDERLOAD}$ DD: medium low OS $\rightarrow f_{FITTING^-}$ EE: medium high OS $\rightarrow f_{FITTING^+}$ FF: high OS $\rightarrow f_{OVERLOAD}$

Table 5.3: The Workload Profile defeasible translation: natural language & formal arguments

As mentioned before, in contrast to the NASA-TLX instrument, in the Workload Profile procedure, preferentiality is not accounted. In turn all the designed arguments have an 'undefined' importance. Let us suppose that an end-user has answered the WP questions of appendix A.3 (page 175) with the following values (in the scale 0 to 100):

- SPPC: 88;
- CPS: 15;
- IV: 90;
- OM: 85;
- SPR: 81;
- CPV: 2;
- IA: 75;
- OS: 3;

The degrees of truth of each argument, according to definition 16 (page 83), are listed in table 5.4.

Attribute $\in ATTR$	Argument's degree of truth	
processing stage:	$A_{deg} = f^{Low}_{SPPC}(45) = 0$	$B_{deg} = f^{Medium^-}_{SPPC}(45) = 0.86$
perceptual/central (SP-PC)	$C_{deg} = f^{Medium^+}_{SPPC}(45) = 0$	$D_{deg} = f^{High^+}_{SPPC}(45) = 0$
processing stage:	$E_{deg} = f^{Low}_{SPR}(31) = 0$	$F_{deg} = f^{Medium^-}_{S-R}(31) = 0.37$
response (SP-R)	$G_{deg} = f^{Medium^+}_{SPR}(31) = 0$	$H_{deg} = f^{High^+}_{SPR}(31) = 0$
processing code:	$I_{deg} = f^{Low}_{CPS}(24) = 0.2$	$J_{deg} = f^{Medium^-}_{CPS}(24) = 0$
spatial (CP-S)	$K_{deg} = f^{Medium^+}_{CPS}(24) = 0$	$L_{deg} = f^{High^+}_{CPS}(24) = 0$
processing code:	$M_{deg} = f^{Low}_{(CPV)}(53) = 0$	$N_{deg} = f^{Medium^-}_{(CPV)}(53) = 0$
verbal (CP-V)	$O_{deg} = f^{Medium^+}_{(CPV)}(53) = 0.9$	$P_{deg} = f^{High^+}_{(CPV)}(53) = 0$
input:	$Q_{deg} = f^{Low}_{IV}(59) = 0$	$R_{deg} = f^{Medium^-}_{IV}(59) = 0$
visual (I-V)	$S_{deg} = f^{Medium^+}_{IV}(59) = 0.7$	$T_{deg} = f^{High^+}_{IV}(59) = 0$
input:	$U_{deg} = f^{Low}_{IA}(28) = 0.07$	$V_{deg} = f^{Medium^-}_{IA}(28) = 0.27$
auditory (I-A)	$W_{deg} = f^{Medium^+}_{IA}(28) = 0$	$X_{deg} = f^{High^+}_{IA}(28) = 0$
output:	$Y_{deg} = f^{Low}_{(OM)}(29) = 0.04$	$Z_{deg} = f^{Medium^-}_{(OM)}(29) = 0.31$
manual (O-M)	$AA_{deg} = f^{Medium^+}_{(OM)}(29) = 0$	$BB_{deg} = f^{High^+}_{(OM)}(29) = 0$
output:	$CC_{deg} = f^{Low}_{(OS)}(45) = 0$	$DD_{deg} = f^{Medium^-}_{(OS)}(45) = 0.86$
speech (O-S)	$EE_{deg} = f^{Medium^+}_{(OS)}(45) = 0$	$FF_{deg} = f^{High^+}_{(OS)}(45) = 0$

Table 5.4: The Workload Profile defeasible translation: degree of truth of arguments

Layer 3: reduction of argumentation graph

Let us initialise the argument and attack reluctancy thresholds as:

- $Reluct^{th}_{Arg} = 0$. Willingness to consider all the arguments whose degree of truth is greater than 0 (def. 16).
- $Reluct^{th}_{Att} = 0$. Willingness to tolerate an attack from a less to a more credible argument - no reluctancy (def. 24).

The set of activated arguments Arg_{act}, computed according to definition 23 (page 90) are: B, F, I, O, S, U, V, Y, Z, DD. The set of activated attack relations $Attack_{act}$, according to definition 25 (page 91), is empty, as no designed attack relation exists. Thus the abstract argumentation graph on which Dung-style grounded and preferred acceptability semantics, as described in section 3.5.3, can be executed, is as in figure 5.7.

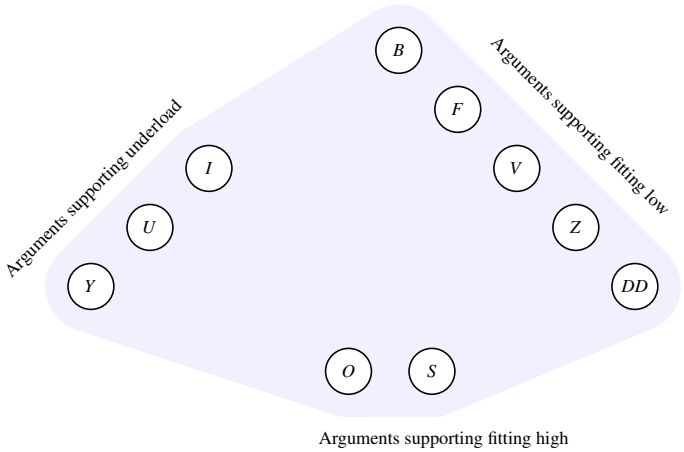


Fig. 5.7: The Workload Profile defeasible translation: reduced argumentation graph

The grounded extension produced by the grounded acceptability semantics (definition 8, page 65) coincides exactly with the reduced graph of figure 5.7, thus containing all the arguments (B, F, I, O, S, U, V, Y, Z, DD). The preferred extensions produced by the preferred acceptability semantics (definition 10, page 66), in this case coincides with the grounded extension:

- E^1: B, F, I, O, S, U, V, Y, Z, DD (grounded extension = preferred extension)

The strength of the unique extension computed according to definition 26 (page 92) is:

Layer 4: extraction of credible extensions

$$E^1_{Strength} = \frac{Card(E^1)}{Card(Arc_{act})} + \frac{B_{deg}+F_{deg}+I_{deg}+O_{deg}+S_{deg}+U_{deg}+V_{deg}+Y_{deg}+Z_{deg}+DD_{deg}}{Card(E_1)}$$
$$= \frac{10}{10} + \frac{0.86+0.37+0.2+0.9+0.7+0.07+0.27+0.04+0.31+0.86}{10} = 0.458$$

E^1 is also the strongest unique extension thus the overall index of mental workload is computed considering the forecast arguments within that extension, as in definition 27 (page 94). The equations of the straight lines associated to the four workload dichotomies are the same as the one used in the translation of the NASA-TLX, as in figure 5.3.

The unique strongest extension carries just forecast arguments and as each of them do not have an associated preference, the overall mental workload score can be computed just using the left part (within big brackets) of the formula of definition 27 (page 94) as it follows:

Layer 5: assessment of mental workload

$$MWL =$$
$$= \frac{Card(Undef)}{Card(E^1)} \cdot \frac{\begin{array}{c} f_{fitting^-}(0.86)+f_{fitting^-}(0.37)+f_{underload}(0.2)+f_{fitting^+}(0.9)+f_{fitting^+}(0.7) \\ f_{underload}(0.07)+f_{fitting^-}(0.27)+f_{underload}(0.04)+f_{fitting^-}(0.31)+f_{fitting^-}(0.86) \end{array}}{Card(Undef)}$$

$$= \frac{10}{10} + \frac{46.76+38.92+25.8+51.6+54.8+29.83+37.32+30.76+37.96+46.76}{10} = 36.97$$

The overall mental workload score is 36.97 which clearly falls within the workload dichotomy $fitting^{-}$. A designer might interpret this workload score as not really optimal, as it is closer to the redline that separates the dichotomy fitting low and underload. This means that the system under examination imposed a moderately fit workload on the tested end-user, and although not generating underload situations, it could be better optimised, according to that user.

In the next section, the goal is to show how to create a brand new instance of the defeasible framework that accounts for a wider set of workload attributes MWL^{NI}_{def}. This is subsequently extended into another instance MWL_{def}, by adding theoretical and logical interaction of designed arguments in form of attack relations.

5.2.3 Definition of a brand new instance of the defeasible framework

A new instance of the defeasible framework is built according to the knowledge-base of the author of this thesis, driven by his subjective interpretation of the literature of mental workload and his beliefs. Each pieces of evidence in this knowledge-base is defeasible thus open to invalidation by other evidence. In addition it does not aim to be fully exhaustive and the final ultimate set of pieces of evidence to consider for representing mental workload, but just a subjective proposal open to criticisms that can be extended, reduced or discarded as a whole. This knowledge-base is summarised with the following natural language propositions, classified as endogenous and exogenous factors. The former refers to those variables inherent in a person's ability and skills, while the latter refers to those variables inherent in the situation.

Layer 1: translation of knowledge-base

- task demands (exogenous factors) - from the NASA-TLX (Hart, 2006; Reid and Nygren, 1988)
    1. *mental demand* has a direct relationship with mental workload: the higher the perceived mental demand of the task, the higher the mental workload.
    2. *temporal demand* has a direct relationship with mental workload: the higher the perceived temporal demand of the task, the higher the mental workload.
    3. *physical demand* has a direct relationship with mental workload: the higher the perceived physical demand of the task the higher the mental workload.
- task features/complexity and interaction with the user (exogenous factors) - from the MRT (Tsang and Velazquez, 1996; Wickens, 2008; Wickens and Hollands, 1999)
    4. *solving and deciding* are notions that have a direct relationship with mental workload: the higher the degree of attention required for decision-making, problem-solving and remembering, the higher the mental workload.
    5. *selection of response* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for selecting the proper response channel, the higher the mental workload.
    6. *task and space* are notions that have a direct relationship with mental workload: the higher the degree of attention required for spatially paying attention around, the higher the mental workload.

7. *verbal material* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for processing linguistic material or listening to verbal conversation or reading, the higher the mental workload.
8. *visual resources* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for the execution of the task based on the information visually received, the higher the mental workload.
9. *auditory resources* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for the execution of the task based on the information auditorily received, the higher the mental workload.
10. *manual response* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for manually respond to the task, the higher the mental workload.
11. *speech response* is a notion that has a direct relationship with mental workload: the higher the degree of attention required for producing the speech response, the higher the mental workload.

- user's state (endogenous factors)

12. *psychological stress* has been thought having a direct relationship with mental workload: the higher the stress felt by the user the higher the mental workload (Hart, 2006; Hart and Staveland, 1988; Reid and Nygren, 1988). However, here the belief is that the psychological stress perceived by the user influences mental workload only when it is too low or too high. In these two cases the operator's state is significantly affected. In the former case, mental workload is at a minimum level, underlying underload, while in the latter case, it is at a maximum level, denoting overload.
13. *arousal* has a complex relationship with performance, following a curve that changes due to task differences. For simple or well-learned tasks, the relationship can be considered linear with improvements in performance as arousal increases. For complex/difficult or unfamiliar tasks, the relationship between arousal and performance becomes inverse, with declines in performance as arousal increases (Yerkes and Dodson, 1908). The effect of task complexity/difficulty leads to the hypothesis that the Yerkes-Dodson law (as depicted in figure 2.7) might be sliced into two distinct parts. The upward part of the inverted 'U' curve might be believed to be as the energising effect of arousal while the downward part is caused by the negative effects of arousal on cognitive processes like attention, memory and problem-solving.

- user intentions (endogenous factors)

14. *effort* has a direct relationship with mental workload: the higher the effort exerted by the user the higher the mental workload (Hart, 2006; Hart and Staveland, 1988).
15. *motivation* is related to effort and performance: the higher the user's motivation to attend to the task the higher is the willingness to exert effort to improve task performance. When motivation is moderate, here it is believed it does not have a significant influence on mental workload. On the other hand, when motivation is too low, it might have a direct relationship with mental workload. In this case, the user's state is affected and workload is hypothesised to be at a minimum level.

- context/domain (exogenous factors)

16. *parallelism* has a direct relationship with mental workload: the higher the parallelism regarding the execution of multiple tasks, the higher the mental workload. In addition harder tasks are harder to perform in parallel as they require more attention and cognitive resources. On the other hand, easier tasks can be concurrently executed more easily. Analogously, tasks which are similar to each other are harder to be executed in parallel than more distinct ones. Similarity of tasks could be measured by employing the dimensions accounted in the multiple resource theory, as previously mentioned (Tsang and Velazquez, 1996; Wickens, 2008; Wickens and Hollands, 1999).

17. *context bias* has a direct relationship with mental workload, when bias are not too low: the higher the bias and distraction degree is, the higher the mental workload. Here it is believed that when bias are too low, workload is not influenced. On the other hand, when a moderate or high degree of bias and interruptions occurs during primary task, users can take longer time to complete the task, committing more errors and experiencing even double negative effects with a significative increment in mental workload (Bailey and Konstan, 2006). In addition, it is reasonable to assume that when the degree of context bias is too high, the psychological stress of a subject is likely not to be low.

- user's features (endogenous factors)

18. *past knowledge* has an inverted relationship with mental workload: the higher the user's knowledge of the task or the context/domain, the lower the mental workload. This is related to the notion of learning as described by Kahneman whose model explains why learning helps, as it makes execution of tasks easier (Kahneman, 1973). When past knowledge is too low, the user has likely never dealt with the task under consideration, thus the mental workload is likely to be high. On the other hand, when past knowledge is high, the user has already learnt the task or similar ones in the past, thus the resulting mental workload is likely to be low. Past knowledge is an important factor that contributes to develop the skill of a person. In addition, if past knowledge is too low, it is very unlikely that a subject exerted no effort to perform a task. Similarly, if past knowledge is too high, it is unlikely that a subject exerted high effort to perform a task.

19. *skill* has an inverted relationship with mental workload: the higher the user's skill the lower the mental workload. Skills incorporate the notion of strategy (heuristic) used for dealing with more difficult and complex tasks in the same context/domain. Heuristic might be seen as mental shortcuts which could provide a reasonable performance without investing too much effort (Wickens and Hollands, 1999). User's skill is important when it is too low or too high. In the former case, the user is not skilled enough to perform the task, experiencing high workload, while in the latter case, the user's skill play a significant role in reducing mental workload on task. Skill is related to past knowledge: if a subject has already dealt with a task or similar tasks, the degree of skill is likely not to be low. In addition, if the degree of skill is too low, it is very unlikely that a subject exerted no effort to perform a task. Similarly, if the degree of skill is too high, it is unlikely that a subject exerted high effort to perform a task.

20. *performance* has an inverted relationship with mental workload: the higher the performance perceived by the user the lower the mental workload and vice-versa (Hart, 2006; Hart and Staveland, 1988; O' Donnel and Eggemeier, 1986).

The above knowledge-base was in practice quantified by designing a subjective questionnaire as in appendix appendix A.5. During the completion of this questionnaire of appendix by experimental volunteers (as described in the next chapter), it has been noted that subjects could better indicate low levels rather than high levels of a workload attribute. In other words, they could better move the sliders of figure 6.2 towards the extreme left side, to indicate a low degree, than moving it towards the extreme right side, to indicate a high degree. This means that in general, subjects could easily quantify the null impact of a workload attribute rather than the full impact, showing more uncertainty in indicating higher levels. This tendency can be understood by taking a closer look at the distributions of the answers of volunteers for each attribute (appendix C, section C.1.3). In these distributions, the frequencies between 66 and 100 tend to be closer to 66. For these reasons, the membership functions for each workload attribute introduced in section 5.2.3, are designed as in figure 5.8.

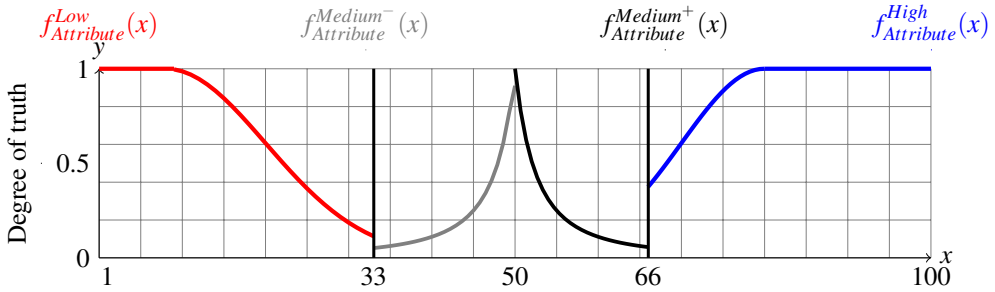


Fig. 5.8: The new defeasible instance: membership functions used for every MWL attribute

These membership functions are generalised Bell curves or Gaussian curves commonly used in Fuzzy Logic FL. In order to model the 'low' and 'high' subsets of each attributes, a composite function has been adopted, that depends on other two Gaussian functions, each with two parameters *sig* and *c* as given by:

$$f(x, sig, c) = e^{\frac{-(x-c)^2}{2sig^2}}$$

The first function, specified by sig_1 and c_1, determines the shape of the left-most curve. The second function specified by sig_2 and c_2 determines the shape of the right-most curve. In the case $c1 < c2$, the total function reaches a maximum value of 1 else the maximum value is less than 1.

$$f_1(x, sig_1, c_1) = \begin{cases} e^{\frac{-(x-c)^2}{2*sig^2}} & \text{if } x \leq c_1 \\ 1 & \text{otherwise} \end{cases} \qquad f_2(x, sig_2, c_2) = \begin{cases} 1 & \text{if } x \leq c_2 \\ e^{\frac{-(x-c_2)^2}{2*sig_2^2}} & \text{otherwise} \end{cases}$$

$$f(x, sig_1, c_1, sig_2, c_2) = f_1(x, sig_1, c_1) * f_2(x, sig_2, c_2)$$

$$f^{Low}_{attribute} : [0..32] \in \aleph \rightarrow [0..1] \in \Re \qquad f^{Low}_{attribute} = f(x,10,1,12,8)$$
$$f^{High}_{attribute} : [67..100] \in \aleph \rightarrow [0..1] \in \Re \qquad f^{High}_{attribute} = f(x,10,80,1,100)$$

As it is possible to see, the extreme left part of the membership function for the 'low' subset returns 1 for inputs less than 8: in these cases there is no doubt the interviewed subjects wanted to rate the attribute under examination as really low. Similarly, the extreme right part of the membership function for the 'high' subset returns 1 for inputs greater than 80: in these cases there is no doubt the subjects wanted to rate the attribute under examination as really high. The extreme right part is more extended than the extreme left part because of the aforementioned issue of subjects being able to indicate 'low' levels better than 'high' levels for a workload attribute.

The function used for the middle subsets of an attribute (lower and upper parts) is a generalised Bell-shaped membership function as follows:

$$f(x,a,b,c) = \frac{1}{1+|\frac{x-c}{a}|^{2}b}$$

with a controlling the width of the curve at $f(x) = 0.5$, b controlling the slope of the curve at $x = c-a$ and $x = c+a$, and c represents the centre of the curve.

$$f^{Medium^{-}}_{attribute}, : [33..50] \in \aleph \rightarrow [0..1] \in \Re \qquad f^{Medium^{-}}_{attribute} = f(x,5,1,50)$$
$$f^{Medium^{+}}_{attribute}, : [51..66] \in \aleph \rightarrow [0..1] \in \Re \qquad f^{Medium^{+}}_{attribute} = f(x,5,1,50)$$

In this new defeasible instance, the designer (author of this thesis) is not willing to specify any preference of considered attributes, thus the function that returns the importance of each attribute is undefined, as per definition 20 (page 87).

$$f_{pref}(x) = \Big\{ undefined \quad \forall x$$

The *workload dichotomies* the author is willing to define for the new instance are the same functions used in section 5.2.1 and depicted in figure 5.3, partitioned by the following redlines:

- $RedLine^{fitting}_{underload} = 33$
- $RedLine^{overload}_{fitting} = 66$

Layer 2: construction of argumentation graph

The forecast arguments that might be designed for the aforementioned knowledge-base are as it follows:

1. *mental demand*:
   - MD1: [low mental demand $\rightarrow$ $UNDERLOAD$]
   - MD2: [medium lower mental demand $\rightarrow$ $FITTING^{-}$]
   - MD3: [medium upper mental demand $\rightarrow$ $FITTING^{+}$]
   - MD4: [high mental demand $\rightarrow$ $OVERLOAD$]

2. *temporal demand*:
   - TD1: [low temporal demand → $UNDERLOAD$]
   - TD2: [medium lower temporal demand → $FITTING^{-}$]
   - TD3: [medium upper temporal demand → $FITTING^{+}$]
   - TD4: [high temporal demand → $OVERLOAD$]

3. *physical demand*:
   - PD1: [low physical demand → $UNDERLOAD$]
   - PD2: [medium lower physical demand → $FITTING^{-}$]
   - PD3: [medium upper physical demand → $FITTING^{+}$]
   - PD4: [high physical demand → $OVERLOAD$]

4. *solving and deciding*:
   - SD1: [low solving/deciding degree → $UNDERLOAD$]
   - SD2: [medium lower solving/deciding degree → $FITTING^{-}$]
   - SD3: [medium upper solving/deciding degree → $FITTING^{+}$]
   - SD4: [high solving/deciding degree → $OVERLOAD$]

5. *selection of response*:
   - SR1: [low selection of response degree → $UNDERLOAD$]
   - SR2: [medium lower selection of response degree → $FITTING^{-}$]
   - SR3: [medium upper selection of response degree → $FITTING^{+}$]
   - SR4: [high selection of response degree → $OVERLOAD$]

6. *task and space*:
   - TS1: [low task/space degree → $UNDERLOAD$]
   - TS2: [medium lower task/space degree→ $FITTING^{-}$]
   - TS3: [medium upper task/space degree → $FITTING^{+}$]
   - TS4: [high task/space degree → $OVERLOAD$]

7. *verbal material*:
   - VM1: [low verbal material degree → $UNDERLOAD$]
   - VM2: [medium lower verbal material degree → $FITTING^{-}$]
   - VM3: [medium upper verbal material degree → $FITTING^{+}$]
   - VM4: [high verbal material degree → $OVERLOAD$]

8. *visual resources*:
   - VR1: [low visual resources degree → $UNDERLOAD$]
   - VR2: [medium lower visual resources degree → $FITTING^{-}$]
   - VR3: [medium upper visual resources degree → $FITTING^{+}$]
   - VR4: [high visual resources degree → $OVERLOAD$]

9. *auditory resources*:
   - AR1: [low auditory resources degree → $UNDERLOAD$]
   - AR2: [medium auditory resources degree → $FITTING^{-}$]
   - AR3: [medium auditory resources degree → $FITTING^{+}$]
   - AR4: [high auditory resources degree → $OVERLOAD$]

10. *manual response*:
    - MR1: [low manual response degree → $UNDERLOAD$]
    - MR2: [medium lower manual response degree → $FITTING^{-}$]
    - MR3: [medium upper manual response degree → $FITTING^{+}$]
    - MR4: [high manual response degree → $OVERLOAD$]

11. *speech response*:
    - SP1: [low speech response degree → $UNDERLOAD$]
    - SP2: [medium lower speech response degree → $FITTING^{-}$]
    - SP3: [medium upper speech response degree → $FITTING^{+}$]
    - SP4: [high speech response degree → $OVERLOAD$]
12. *psychological stress*:
    - PS1: [low psychological stress → $UNDERLOAD$]
    - PS2: [high psychological stress → $OVERLOAD$]
13. *arousal*: none
14. *effort*:
    - EF1: [low effort → $UNDERLOAD$]
    - EF2: [medium lower effort → $FITTING^{-}$]
    - EF3: [medium upper effort → $FITTING^{+}$]
    - EF4: [high effort → $OVERLOAD$]
15. *motivation*:
    - MV1: [low motivation → $UNDERLOAD$]
16. *parallelism*:
    - PA1: [low parallelism degree → $UNDERLOAD$]
    - PA2: [medium lower parallelism degree→ $FITTING^{-}$]
    - PA3: [medium upper parallelism degree → $FITTING^{+}$]
    - PA4: [high parallelism degree → $OVERLOAD$]
17. *context bias*:
    - CB1: [low context bias degree→ $UNDERLOAD$]
    - CB2: [medium lower context bias degree→ $FITTING^{-}$]
    - CB3: [medium upper context bias degree → $FITTING^{+}$]
    - CB4: [high context bias degree → $OVERLOAD$]
18. *past knowledge*:
    - PK1: [low past knowledge → $OVERLOAD$]
    - PK2: [high past knowledge → $UNDERLOAD$]
19. *skills*:
    - SK1: [low skills → $OVERLOAD$]
    - SK2: [high skills → $UNDERLOAD$]
20. *performance*:
    - PF1: [low perceived performance → $OVERLOAD$]
    - PF2: [medium lower perceived performance → $FITTING^{+}$]
    - PF3: [medium upper perceived performance → $FITTING^{-}$]
    - PF4: [high perceived performance → $UNDERLOAD$]

The mitigating arguments that might be designed considering the aforementioned knowledge-base are:

13 *arousal*: (the arguments are built upon the relationships between task difficulty, arousal and performance of figure 2.7 and summarised in figure 5.9)
- AD1a: [low arousal & easy task → PF4]
- AD1b: [low arousal & easy task → PF3]
- AD1c: [low arousal & easy task → PF2]
- AD2a: [low arousal & difficult task → PF4]
- AD2b: [low arousal & difficult task → PF3]
- AD2c: [low arousal & difficult task → PF2]
- AD3a: [medium lower arousal & easy task → PF1]
- AD3b: [medium lower arousal & easy task → PF4]

- AD4a: [medium lower arousal & difficult task → PF1]
- AD4b: [medium lower arousal & difficult task → PF3]
- AD4c: [medium lower arousal & difficult task → PF4]
- AD4d: [medium upper arousal & difficult task → PF1]
- AD4e: [medium upper arousal & difficult task → PF3]
- AD4f: [medium upper arousal & difficult task → PF4]
- AD5a: [medium upper arousal & easy task → PF1]
- AD5b: [medium upper arousal & easy task → PF2]
- AD5c: [medium upper arousal & easy task → PF3]
- AD5d: [high arousal & easy task → PF1]
- AD5e: [high arousal & easy task → PF2]
- AD5f: [high arousal & easy task → PF3]
- AD6a: [high arousal & difficult task → PF2]
- AD6b: [high arousal & difficult task → PF3]
- AD6c: [high arousal & difficult task → PF4]

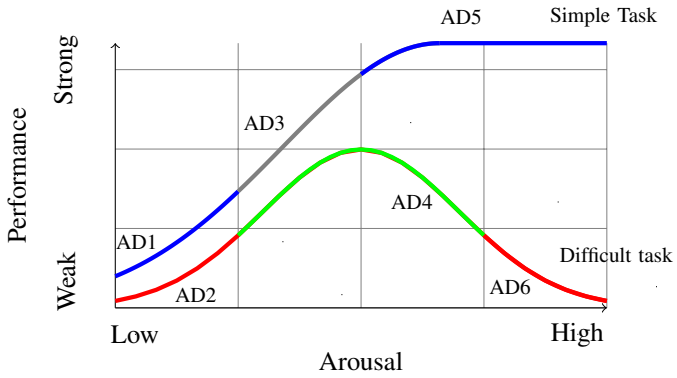


Fig. 5.9: Arguments extracted from the relationships between task difficulty, arousal and performance

15 *motivation*:

- MV2: [low motivation → EF3]
- MV3: [low motivation → EF4]
- MV4: [high motivation → EF1]
- MV5: [high motivation → EF2]

19 *skills*:

- DS1 [difficult task & high skills → EF4]
- DS2 [difficult task & high skills & low effort → PF1]
- DS3 [difficult task & high skills & medium lower effort → PF1]
- DS4 [difficult task & high skills & medium upper effort → PF1]

The rebutting attack relations that might be built from the aforementioned knowledge-base are:

- Rebutting
  - The attributes ‘mental demand’ and ‘solving and deciding’ model similar notions, therefore they contradict each other if they support totally different conclusions (two total opposite workload dichotomies). In this case rebutting attacks model the inconsistency of the available evidence:
    - r1: (MD1, SD4)
    - r2: (MD4, SD1)
  - From the knowledge-base (points 18, 19) high skills and low past knowledge (and vice-versa) are situations that should not occur. Therefore rebutting attacks between these two extreme opposite degrees of skill and past knowledge are aimed at model such inconsistency:
    - r3: (PK1, SK4)
    - r4: (PK4, SK1)

- From the knowledge-base (points 18, 19) high skills and high effort, low skills and low effort, are situations that should not occur. Similarly, between high past knowledge and high effort (and low past knowledge and low effort). These inconsistent cases are modelled with the following rebutting attacks:
  - r5: (PK1, EF1)
  - r6: (PK2, EF4)
  - r7: (SK1, EF1)
  - r8: (SK4, EF4)
- From the knowledge-base (point 17) a higher degree of context bias is in contradiction with a lower degree of psychological stress. Thus to model this inconsistency, the following rebutting attack might be designed:
  - r9: (CB4, PS1)

The undercutting attack relations that follow (according to definition 19, page 85) from the designed mitigating arguments are as it follows:

- Undermining
  - um1: (AD1a, PF4), um2: (AD1b, PF3), um3: (AD1c, PF2)
  - um4: (AD2a, PF4), um5: (AD2b, PF3), um6: (AD2c, PF2)
  - um7: (AD3a, PF1), um8: (AD3b, PF4)
  - um9: (AD4a, PF1), um10: (AD4b, PF3), um11: (AD4c, PF4), um12: (AD4d, PF1), um13: (AD4e, PF3), um14: (AD4f, PF4)
  - um15: (AD5a, PF1), um16: (AD5b, PF2), um17: (AD5c, PF3) um18: (AD5d, PF1), um19: (AD5e, PF2), um20: (AD5f, PF3)
  - um21: (AD6a, PF2), um22: (AD6b, PF3), um23: (AD6c, PF4)
  - uc1: (MV2, EF3), uc2: (MV3, EF4), uc3: (MV4, EF1), uc4: (MV5, EF2)
  - uc5: (DS1, EF4), uc6: (DS2, PF1), uc7: (DS3, PF1), uc8: (DS4, PF1)

The argumentation graph that results by joining all the forecast and mitigating arguments is part of an instance of the defeasible framework now on referred to as MWL_{def}^{NI} (no interactions). Extending this argumentation graph by adding the designed rebutting and undercutting attacks, a new instance of the framework emerges, now on referred to as MWL_{def}. These two instances are treated as different instances of the defeasible framework because, as described in the following chapter, they are separately evaluated. The argumentation graph emerged in MWL_{def} (including interactions) is depicted in figure 5.10.

In addition, as it is possible to see in the list of attributes of list 5.2.3, the attribute *arousal* is based on *task difficulty* for which no question has been designed in the questionnaire of table A.5 of appendix. As a consequence an explicit mechanism to quantify task difficulty is needed for representing the workload attribute *arousal* and also the mitigating arguments designed upon it (list 5.2.3). Here the proposal is to model *task difficulty* as the average of the workload attributes accounted in the Workload Profile WP instrument which can be quantified because an explicit question has been designed for each of them (questionnaire A.5).

$$Task_{difficulty} : [0..100] \in \Re$$

$$Task_{difficulty} = \frac{1}{8}((solving/deciding) + (response) + (task/space) + (verbal \quad material) + (visual\ resources) + (auditory\ resources) + (manual\ response) + (speech\ response))$$

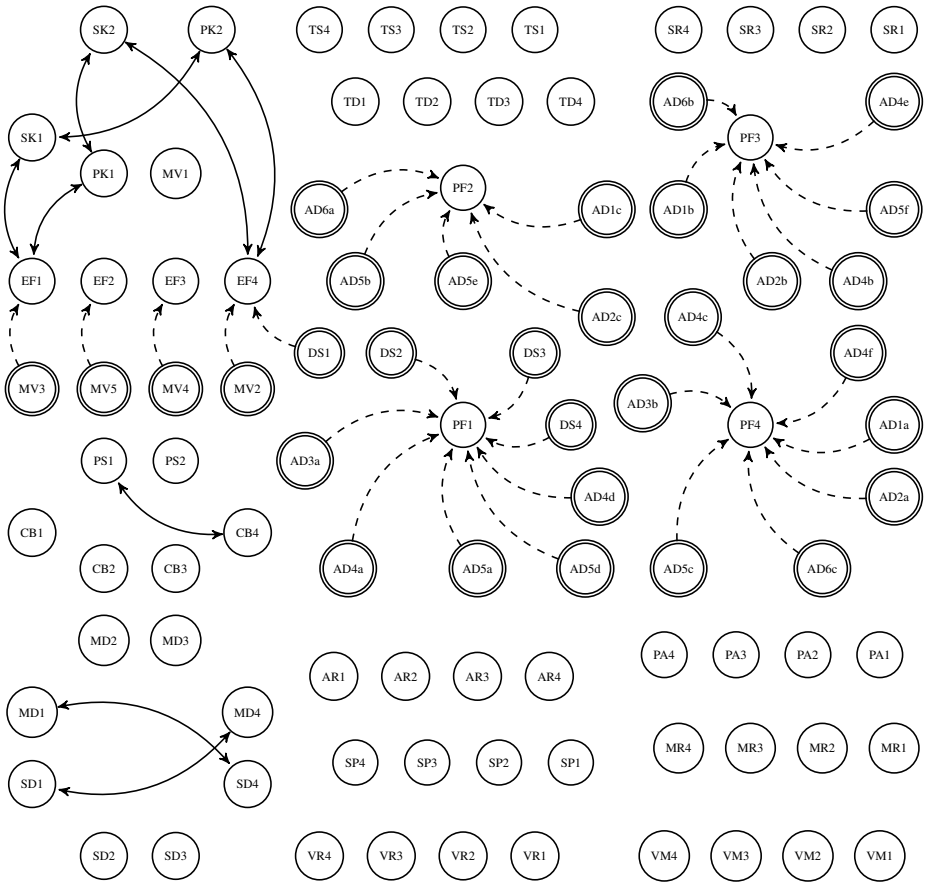


Fig. 5.10: The brand new instance: Knowledge-base translated into an argumentation graph

Layer 3: reduction of argumentation graph

The *argument reluctancy threshold* and the *attack reluctancy threshold* are defined as it follows:

- $Reluct^{th}_{Arg} = 0$. Willingness to consider all the arguments whose degree of truth is greater than 0 (def. 16).
- $Reluct^{th}_{Att} = 0.5$. Willingness to tolerate an attack from a less to a more credible argument just if their difference of degree of truth is not more than 0.5 (def. 24).

The instance MWL_{def} can be summarised with the following tuple using an illustrative set of inputs (answers of the questionnaire of table A.5 in appendix) is used:

$$MWL_{def} = \{ATTR,\ f_{Pref},\ MF,\ RL,\ DMF,\ ARGS,\ ATTACKS,\ RT,\ INPUTS\}$$

with

- **ATTR**: { Mental demand, temporal demand, effort, performance, frustration, solving and deciding, selection of response, task and space, verbal material, visual resources, auditory resources, manual response, speech response, context bias, past knowledge, skill, motivation, parallelism, arousal, task difficulty}.
- **Pref**: $f_{pref}(x)$ is $undefined$ (no preferentiality considered)

- **MF**: the membership function for the attributes are the ones defined in figure 5.8
- **RL**: $\{ RedLine^{fitting}_{underload} = 33, RedLine^{overload}_{fitting} = 66 \}$
- **DMF**: workload dichotomies of figure 5.3
- **ARGS**: the designed arguments built upon the attributes in $ATTR$ are the ones listed in section 5.2.3 (page 112)
- **ATTACKS**: the designed attack relationships are the ones defined in page 115
- **RT**: $\{ Reluct^{th}_{Arg} = 0, Reluct^{th}_{Att} = 0.5 \}$
- **INPUTS**: {70, 15, 78, 76, 12 18, 17, 14, 78, 82, 9, 13, 0, 9, 71, 64, 16, 7, 21, 30}

Note 9

The instance MWL^{NI}_{def} coincides with the above tuple describing the instance MWL_{def} with the only difference that the $ATTACK$ set is empty.

Note 10

The values in the **INPUTS** set are random and just for illustrative purposes. These values are the inputs responsible for the activation of the argumentation graph behind the instances MWL^{NI}_{def} and MWL_{def}

The instances MWL^{NI}_{def} and MWL_{def} can now be evaluated by applying the algorithm of section 5.1, starting with the activation of arguments and attacks relations (using the values in the $INPUTS$ set of the tuple). Table 5.5 lists which arguments are activated with the correspondent degree of truth (according to definition 23, page 90).

Argument	Internal representation	Degree of truth
MD4	high mental demand → $OVERLOAD$	0.606
TD1	low temporal demand → $UNDERLOAD$	0.843
EF4	high effort → $OVERLOAD$	0.980
PF4	high perceived performance → $UNDERLOAD$	0.923
PS1	low psychological stress → $UNDERLOAD$	0.945
SD1	low solving/deciding degree → $UNDERLOAD$	0.706
SR1	low selection of response degree → $UNDERLOAD$	0.754
TS1	low task and space degree → $UNDERLOAD$	0.882
VM4	high verbal material degree → $OVERLOAD$	0.980
VR4	high visual resources degree → $OVERLOAD$	1.000
AR1	low auditory resources degree → $UNDERLOAD$	0.996
MR1	low manual response degree → $UNDERLOAD$	0.916
SP1	low speech response degree → $UNDERLOAD$	0.916
MV1	low motivation → $UNDERLOAD$	0.800
PA1	low parallelism degree → $UNDERLOAD$	1.000
CB1	low context bias degree→ $UNDERLOAD$	0.996
PK2	high past knowledge → $UNDERLOAD$	0.666
MV3	low motivation → EF4	0.800
ADa1	low arousal and easy task → PF4	0.371

Table 5.5: An illustrative scenario: activated arguments and degree of truth for MWL_{def}

Table 5.6 lists the activated attacks (according to definition 25, page 91). The union of the set of activated arguments and the set of activated attacks forms the argumentation graph depicted in figure 5.11 that can now be evaluated by applying the Dung's preferred semantic, as described in section 3.5.3.

Attack	Internal representation
uc2	(MV3, EF4)
r2	(MD4, SD1)
r6	(PK2, EF4)

Table 5.6: An illustrative scenario: activated attack relations for MWL_{def}

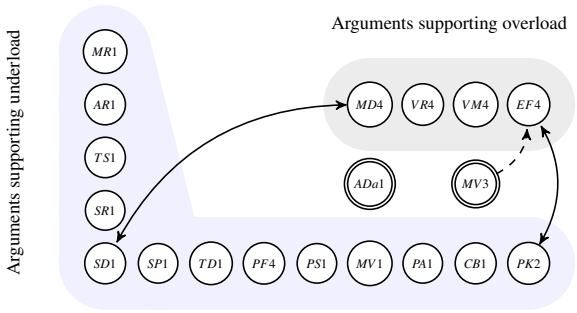


Fig. 5.11: An illustrative scenario: activated arguments and attack relations for MWL_{def}

Layer 4: extraction of credible extensions

Multiple extensions of arguments might be computed by the preferred semantic, as per definition 10 (page 66). In this case, their strength is separately computed as per definition 26 (page 92). From the reduced argumentation graph of figure 5.11, two preferred extensions are computed (with the values in the $INPUTS$ set of the tuple), and according to definition 26 (page 92) their strength is as it follows:

- Extension 1: 1.673
- Extension 2: 1.679

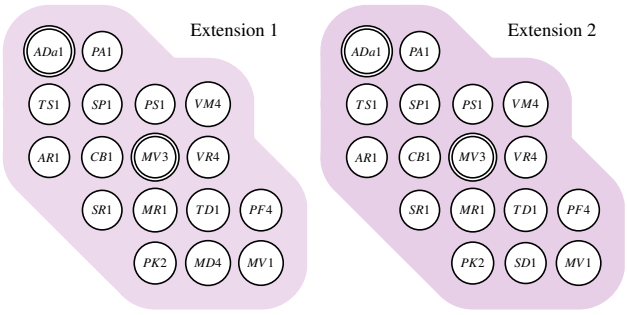


Fig. 5.12: An illustrative scenario: computed preferred extensions for MWL_{def}

As a consequence, extension 2, although really similar to extension 1, is the strongest preferred extension that can be used to compute the final index of mental workload, according to definition 27 (page 94).

Layer 5: assessment of mental workload

The degree of truth of each forecast argument in the stronger extension (ex. 2) is used as the input of the workload dichotomy supported by the argument itself to compute a partial workload score. The average of these scores represents the final index of mental workload, which in this case is 16.81.

Note 11

It is important to recall that definition 27 (page 94) can handle multiple strongest extensions, and it accounts for the importance associated to each arguments that, however, it is undefined in the instance MWL_{def} ($f_{pref}(x) = undefined$).

The two created instances built for the translation of the NASA-TLX and the WP instruments, (sections 5.2.1 and 5.2.2), as well as the brand new instances MWL_{def} and MWL_{def}^{NI} (with and without interaction of arguments) will be part of the evaluation strategy of the defeasible framework, as described in the next chapter.

Chapter 6

Evaluation

This chapter is devoted to the description and execution of a user study for the evaluation of the defeasible framework for mental workload representation and assessment designed in chapter 4. The first goal, as stated in objective 4 of the introduction chapter, is to evaluate the capacity of the proposed framework to reproduce the NASA-TLX and the Workload profile WP assessment procedures. The hypothesis is that the assessments produced by these two state-of-the-art original instruments, which are then used as the baseline, can be replicated by the two defeasible instances of the framework ($NASA-TLX_{def}$ and WP_{def}), as constructed in the previous chapter 5. This capacity is tested by investigating the convergent validity of the two instances with the two original instruments. The second goal of the introduction chapter, as stated in objective 5, is to investigate the quality of the assessments produced by the brand new instances of the defeasible framework MWL_{def} and MWL^{NI}_{def}, as built in previous chapter 5. The hypothesis is that the assessments produced by MWL_{def} have a similar sensitivity, but a better diagnosticity and validity than the original NASA-TLX and the original Workload Profile WP instruments. Additionally, MWL_{def} is hypothesised to overperform the other instance $MWLdef^{NI}$, showing how interaction of arguments can actually play a significant role in the assessment of mental workload.

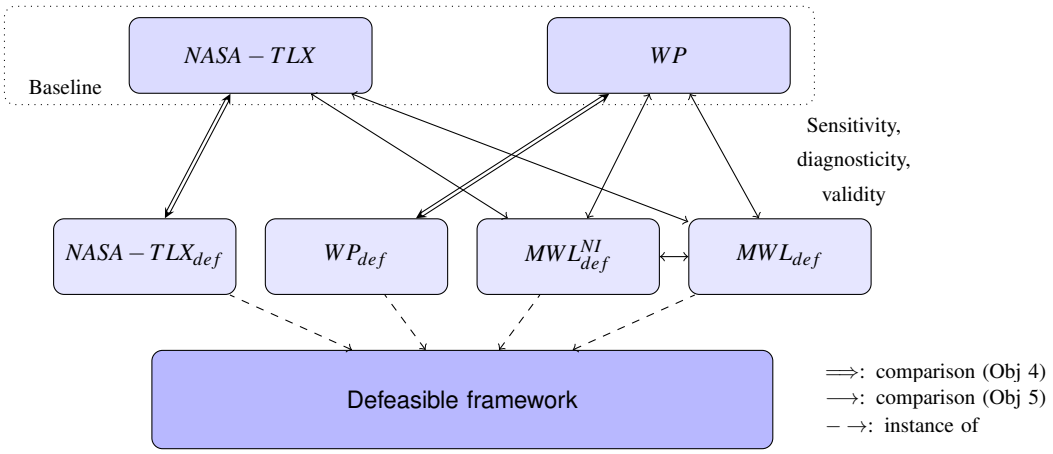


Fig. 6.1: Evaluation strategy schema

6.1 Resolution of objectives and experimental studies

The concept of mental workload has been applied in computer science with an increasing focus in the WWW domain because it is believed to be an important construct for analysing user experience over web-based tasks as well as a design criterion (Albers, 2011; Gwizdka, 2009, 2010; Kokini et al., 2012; Longo et al., 2012b; Schultheis and Jameson, 2004; Tremoulet et al., 2009; Wiebe et al., 2010). For these reasons the WWW domain has been selected as the context for conducting experimental studies, the WWW being part of the wider field of HCI. The evaluation of the framework is the last step for answering the research question of section 1.5, and it deals with objectives 4 and 5 defined in the introduction chapter:

4 investigation of the capacity of the designed framework to reproduce state-of-the-art workload assessment techniques
5 investigation of the quality of the assessments produced by brand new instances of the defeasible framework in the field of HCI

The hypothesis behind objective 4 is that the workload assessments (crisp scores) computed by the two selected baseline instruments ($NASA-TLX$ and WP) can be replicated by their correspondent translations, the instances of the defeasible framework $NASATLX_{def}$ and WP_{def}, as built in chapter 5. The expectation is to obtain a high *convergent validity* between the outcomes of the baseline instruments and the correspondent translations, over a set of designed web-tasks. In other words, the defeasible instances of the framework have to detect high workload when the correspondent base-line instruments assess high workload, and low workload when the correspondent baseline instruments assess low workload. Statistically speaking, the outcomes (workload assessment) of the defeasible instances and the baseline instruments have to highly correlate.

In order to solve objective 5 and to demonstrate that the defeasible framework can be used for enhancing the quality of mental workload assessments, the proposal is to investigate the *sensitivity*, the *diagnosticity* capacity and the *validity* of the brand new instances of the framework MWL_{def} and MWL^{NI}_{def}, as designed in chapter 5. These criteria are in line with those criteria which emerged in the literature review and are listed in section 2.2 (page 16). *Sensitivity* refers to the reliability of an instance of the framework to detect changes in resource demand, task difficulty, user features and environmental influence, as well as its discriminatory capacity between significant variations of workload. *Diagnosticity* refers to the capacity of the instance of the framework to indicate the sources that cause variations in workload, and to quantify the contributions to the workload by the type, resource demand or the capabilities of the human operator. *Validity* refers to the extent to which a measure, produced by an instance of the framework, is actually measuring the construct in question. In the literature of mental workload, different types of validity have emerged (Rubio et al., 2004; Zhang and Luximon, 2005). Here the focus is on the *convergent validity* between different workload scores produced by different assessment techniques, and the *concurrent validity* between workload scores and objective performance measures. As already described in chapter 5, the brand new instance MWL_{def} is an extension of the instance MWL^{NI}_{def} (no interaction of arguments). These instances incorporate the workload attributes considered both in the NASA-TLX instrument (appendix A.1, page 173) and the Workload Profile instrument (WP - appendix A.2, page 174), as well as a new set of attributes believed, by the author of this thesis, to influence mental workload. The former instance extends the latter by adding relationships (in the

form of attacks) among designed arguments built according to the theoretical knowledge which emerged from the literature review of mental workload of chapter 2, and according to the author's understanding of it. The hypothesis is that MWL_{def}^{NI} will underperform, in general, with regards to the original NASA-TLX and WP, but MWL_{def} will show a better sensitivity, a higher diagnosticity capacity, a high convergent validity, and a better concurrent validity than the two baseline assessment instruments over a set of designed web-tasks. In turn, this will confirm the positive impact of adding interactions of arguments, meaning the interaction of pieces of knowledge in the representation and assessment of mental workload. Objectives 4 and 5 and the aforementioned hypotheses are summarised in table 6.1.

Objective 4	
Description	Investigation of the capacity of the designed framework to reproduce state-of-the-art workload assessment techniques
Method	• Translation of the NASA-TLX and WP into instances of the defeasible framework ($NASA\text{-}TLX_{def}$ and WP_{def}) for the investigation of the convergent validity over a set of designed web-tasks
Hypothesis	1 - high convergent validity between NASA-TLX & $NASA\text{-}TLX_{def}$ 2- high convergent validity between WP & WP_{def}

Objective 5	
Description	Investigation of the quality of the assessments produced by new instances of the defeasible framework in the field of HCI
Method	• creation of an instance of the framework (MWL_{def}^{NI}) with no interaction among arguments • extension of MWL_{def}^{NI} to MWL_{def} by adding interactions of arguments • comparison of the sensitivity and the validity of MWL_{def}^{NI}m $NASA\text{-}TLX$ and WP over a set of designed web-tasks • comparison of sensitivity, diagnosticity, validity of MWL_{def}, $NASA\text{-}TLX$ and WP over a set of designed web-tasks
Hypothesis	1 - higher sensitivity of MWL_{def} over $NASA\text{-}TLX$ and WP 2 - higher diagnosticity of MWL_{def} over $NASA\text{-}TLX$ and WP 3 - positive convergent validity of MWL_{def}, $NASA\text{-}TLX$ and WP 4 - better concurrent validity (with time) of MWL_{def} over $NASA\text{-}TLX$ and WP 5 - worse sensitivity and concurrent validity (with time) of MWL_{def}^{NI} over $NASA\text{-}TLX$, WP and MWL_{def}

Table 6.1: Evaluation objectives and hypothesis

6.2 Design of experiments

A set of 11 web-tasks has been defined (table 6.2) on three popular websites: google, wikipedia, youtube. Tasks were designed with different levels of objective difficulty, as shown in table 6.3, manipulating task demands, time pressure, resource demands, duality and concurrency, forming different task conditions. Participants where divided into two groups (A and B) and each person executed the 11 tasks over 2 or 3 sessions, spread over different days. The order of tasks was kept the same for each volunteer. The difference between the 2 groups is that, if people in group A executed a task on the original web-interface, people in group B executed the same task on an altered version of the same web-interface. The details of this experimental design are depicted in table 6.3, with references to screenshots of the interfaces used along with the typology of task, according to the categorisation provided in (Kellar et al., 2006). Task were mainly information seeking tasks, where information had to be found using different cognitive modalities (eyes, touch, ears), in line with modern technologies and web-sites. People in both the groups were exposed to the use of some original and some modified interface, altered at run-time through CSS and HTML manipulation. Original web-interfaces were modified because, as it is going to be shown in the next chapter, the aim was to study the impact of the structural changes of each interface on imposed mental workload on end-users. This also explains why people were divided into two groups.

Task	Description	Notes	Web-site
T_1	Find out how many people live in Sidney		Wikipedia
T_2	Read the content of `simple.wikipedia.org/wiki/Grammar`	No time imposed (user can exit at any time)	Wikipedia
T_3	Using `youtube.com` play your favourite song and while listening to it, search the related lyrics	90 seconds limit	Youtube + Google
T_4	Find out the difference (in years) between the year of the foundation of the Apple Computer Inc. and the year of the 14^{th} FIFA world cup		Google
T_5	Find out the difference (in years) between the foundation of the Microsoft Corporation and the year of the 23^{rd} Olympic games		Google
T_6	Find out the year of birth of the 1^{st} wife of the founder of playboy	2 minutes available. Each 30 seconds the participant is warned of how much time is left	Google
T_1	Find out the name of the man (interpreted by Johnny Deep) in the video `www.youtube.com/watch?v=FfTPS-TFQ_c`	Participant can replay the video if required	Youtube
T_1	a) Play the following song `www.youtube.com/watch?v=Rb5G1eRIj6c` and while listening to it, b) find out the result of the polynomial equation $p(x)$, with $x = 7$ contained in the wikipedia article `http://it.wikipedia.org/wiki/Polinomi`	The song is extremely irritating	Wikipedia
T_1	Find out how many times Stewie jumps in the video `www.youtube.com/watch?v=TSe9gbdkQ8s`	Participant is distracted twice by examiner & can replay the video	Youtube
T_{10}	a) find out (using google.com) the difference (in years) between the foundation of the Alfa Romeo and the year of the 15^{th} New York City marathon b) find out (using wikipedia.com) the capital of Namibia c) find out the two common words that appear in the titles of every referenced paper of the author Longo L. within the wikipedia article `en.wikipedia.org/wiki/Collaborative_search_engine`	Every 30 seconds the participant is forced to switch to the subsequent task (in a different browser's tab) in a loop until the three tasks are completed	Google + Wikipedia
T_{11}	Find out the age of the blue fish in the video `www.youtube.com/watch?v=H4BNbHBcnDI`	150 seconds available. Participant can replay the video. There is no answer.	Youtube

Table 6.2: List of experimental tasks

Task	Typology of tasks, features and task conditions	Web-site	Group A interface	Group B interface	Appendix screenshots
T_1	Fact finding: simple search	Wikipedia		*altered*	B.1
T_2	Browsing: not goal-oriented + time limit	Wikipedia	*altered*		B.2
T_3	Browsing: goal-oriented task	Youtube	*altered*		B.3
T_4	Fact finding: dual task + arithmetic	Google		*altered*	B.4
T_5	Fact finding: dual task + arithmetic	Google	*altered*		B.5
T_6	Fact finding: single task + time pressure	Google		*altered*	B.6
T_7	Fact finding: constant demand on visual + auditory resource	Youtube		*altered*	B.7
T_8	Fact finding: Simultaneous demand on auditory resource + visual resource + arithmetic	Youtube + Wikipedia	*altered*		B.8
T_9	Fact finding: Single tasks on visual resource + external interference	Youtube	*altered*		B.9
T_{10}	Fact finding: Multiple concurrent tasks + time pressure	Google + Wikipedia	*altered*		B.10
T_{11}	Fact finding: demands on auditory + visual resources + verbal processing	Youtube		*altered*	B.11

Table 6.3: Interfaces used in experimental tasks by the two groups

6.2.1 Participants and procedure

A sample of 40 people volunteered to participate in the study, 20 for each group. Ages ranges from 20 to 35 years (Total - Mean: 28.6, Standard deviation 3.98; Group A - Mean 28.35, Standard deviation: 4.22; Group B - Mean: 28.85, Standard deviation: 3.70). 20 were females and 20 males, all with an Internet daily usage of at least 2 hours. Native languages of participants were English, German, French, Polish, Portuguese, Spanish, Chinese, Italian, Arabian, Thai, Armenian, Czech, Hindi, as detailed in table 6.4, and all were relatively fluent in English, so it was assumed they could understand administered questionnaires without major problems. Subjects were instructed about the study and they were required to sign a consent form (see appendix C, page 188) both for data protection and for obtaining detailed study information. Participants were required to execute the tasks of table 6.2 as naturally as they could, over 2 or 3 sessions of approximately 45/70 minutes each, in different days, not consecutive. This experimental design was dictated by subjects unavailability to execute all tasks in one session and also to minimise the learning effect generated by the experiment itself. The subjects in group A were all different than the subjects in group B. Participants could not interact with examiners during the task and a printed version of the instructions was available to them at all time. The order of the tasks administered over the sessions was the same for all the participants:

- 8, 1, 3, 10, 9, 6, 11, 4, 5, 2, 7

In each session, mental workload measures were taken immediately after the task was completed. Specifically, participants were asked to fill in the questionnaires associated respectively to the NASA-TLX and the Workload Profile WP) instruments, (appendix A.1 and A.3, pages 173 and 175). In addition, other 6 questions were administered, aimed at modelling other attributes believed to be related to mental workload (by the author of this thesis), consistently with the attributes accounted in the knowledge-base of the brand new instances MWL_{def} and MWL^{NI}_{def} constructed in chapter 5. The overall set of questions administered

to participants is summarised in table A.5 (appendix, page 177). The web-based tasks were executed on an iMac with 21.5-inch screen, using Mozilla Firefox as browser. Data were collected in a laboratory of the Department of Computer Science and Statistics at Trinity College Dublin. Questions were administered through a desktop software and the order of the questionnaires (NASA-TLX, WP, new 6 questions) was random. Since questions were administered in English and not all the participants were native speakers, each volunteer was invited to ask clarifications to examiners at any time while answering them. Each question had to be rated with a numerical value within the range 0 to 100, by moving a slider. The default value was 50 and the range of values were divided in three parts of equal size, guided by two separation lines, generating 3 regions (low, medium, high), aimed at orientating the user, as in figure 6.2.

	Group A		Group B	
Native language	males	females	males	females
English	1		5	1
German	2	2		
French	2	3		1
Polish		1		
Portuguese	1	2		2
Spanish		1		1
Chinese	1			1
Italian	2	1	4	1
Arabian			1	
Thai			1	
Armenian				1
Czech				1
Hindi	1			

Table 6.4: Demographics of volunteers for the user-study

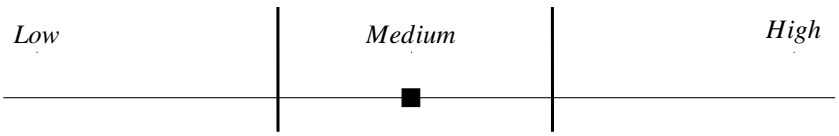


Fig. 6.2: The scale of the answer used for experimental questionnaire

6.3 Evaluating the convergent validity of the NASA-TLX, WP instruments and their defeasible translations

The first step towards the resolution of objective 4 is to investigate the capacity of the designed defeasible framework to reproduce two state-of-the-art workload assessment techniques. In detail, these are the NASA-TLX and the WP mental workload assessment instruments. Chapter 5 demonstrated how these two instruments, here used as baseline, can be translated into particular instances of the defeasible framework, namely the $NASA-TLX_{def}$ and the WP_{def}. The goal is to investigate whether the baseline instruments and the respective defeasible translations have a *convergent validity* over the set of defined web-tasks of table 6.2. In other words, the goal is to investigate whether the workload assessments produced by the baseline instruments and the correspondent defeasible instances of the framework have a significant convergent validity. Convergent validity is a parameter often used in psychology, sociology and other behavioural sciences that indicates the degree to which two measures of constructs, that are supposed to be theoretically related, are in fact related. The parameter can be measured by using the *Pearson Correlation Coefficient* ρ[1]. The hypothesis is to obtain a high convergent validity across any comparison.

6.3.1 Results and discussion

Experiments with the 40 volunteers who executed the 11 designed tasks, generated 440 cases (220 cases for each group). Each case was evaluated individually with each of the following computational models:

- NASATLX (baseline model as described in subsection 2.4.2)
- WP (baseline model as described in subsection 2.4.1)
- $NASATLX_{def}$ (NASA-TLX defeasible translation as designed in 5.2.1)
- WP_{def} (Workload Profile defeasible translation as designed in 5.2.2)

The Pearson's correlation coefficient can be applied if few assumptions are met:

- *continuity:* the variables under consideration should be measured at the interval or ratio level. In other words, they have to be continuous;
- *linearity:* the variables have to show a linear relationship;
- *outliers:* the outliers, that means single data points within the dataset that do not follow the usual pattern, should be removed;
- *normality:* the variables under consideration should be approximately normally distributed.

Instead of taking them for granted, they are verified as it follows. The assumption of *continuity* is met because the mental workload assessment of each tested instrument (computational model) is a value in the range $[0..100] \in \Re$. The assumption of *linearity* is checked by analysing the scatterplots of section C.1.2 (appendix C): all the points are close to the straight line, both in the total scatterplots and in the individual

[1] The Pearson correlation coefficient is a measure of the strength of the linear dependence (correlation) between two variables X and Y returning a value in the range $[-1..1] \in \Re$. A value of -1 indicates perfect negative relationship with one variable increasing and the other decreasing, while a value of $+1$ indicates a perfect positive relationship with both the variables behaving the same way, either both increasing or decreasing. Values closer to 0 imply there is no linear correlation between the two variables. $\rho = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$

scatterplots for each group of users (A and B). The third assumption is met as there are not significant *outliers* in the scatterplots. The final assumption of *normality* is checked by analysing the distributions of the computed workload scores of section C.1.1 (Appendix C). All of them look normally distributed and this is formally confirmed by the Shapiro-Wilk [2] tests of table 6.5, run with a 95% confidence interval. Here, the significance value for each instrument is greater than 0.05, underlying the normality of the data. In the case it would have been below 0.05, the data would significantly have deviated from a normal distribution.

	Total			Group A			Group B		
Model	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.
$NASATLX$	0.995	440	0.163	0.995	220	0.720	0.990	220	0.158
$NASATLX_{def}$	0.995	440	0.164	0.995	220	0.656	0.991	220	0.196
WP	0.995	440	0.140	0.993	220	0.277	0.988	220	0.067
WP_{def}	0.995	440	0.176	0.992	220	0.418	0.991	220	0.167

Table 6.5: Shapiro-Wilk normality test of $NASATLX$, WP and their defeasible translations

The descriptive statistics for each computational model are described in table 6.6. The Pearson coefficients are listed in table 6.7 and the comparisons of the mean of each instrument are shown in figures 6.3 and 6.4.

	Total				Group A				Group B			
Model	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std
$NASATLX_{def}$	0.98	99.9	46.4	19.2	0.98	99.9	46.9	18.8	1.4	99.3	45.9	19.6
$NASATLX$	0.80	99.9	46.2	19.6	0.80	99.9	46.6	19.3	1.4	99.3	45.7	19.8
WP_{def}	0.97	77.3	38.9	14.3	0.97	77.3	39.6	15.3	4.6	66.6	38.2	13.1
WP	1.00	77.0	38.3	14.7	1.00	77.0	38.9	15.9	4.7	67.5	37.6	13.4

Table 6.6: Descriptive statistics for the baseline instruments $NASATLX$, WP and their defeasible translations

All the Pearson's correlation coefficients of table 6.7 are statistically significant ($p < 0.001$) and extremely positive. Several authors have offered guidelines toward the interpretation of a correlation coefficient, however each criterion is arbitrary and context-dependent [3]. The obtained correlations are nearly perfect (close to 1) and this can be considered extremely positive and encouraging, underlining the nearly perfect *convergent validity* of the assessments produced by baseline instruments (NASA-TLX, WP) and the correspondent instances of the defeasible framework ($NASATLX_{def}$, WP_{def}). It can be also noted that the assessments of the sbaseline instruments moderately correlate to each other (Total 0.584) and as a consequence their defeasible translations (Total 0.585). This underlines the positive *face validity* [4] of the original instruments and the correspondent defeasible translations. In this respect, figures 6.3 and 6.4 demonstrate face validity

[2] The Shapiro-Wilk Test is more appropriate when the sample sizes is small (< 50 samples) rather than the classical Kolmogorov-Smirnov Test

[3] Generally, in social and behavioural sciences, there may be a greater contribution from complicating factors, as in the case of subjective ratings, thus correlations above 0.5 are regarded as very high (Cohen, 1988, page 82). Similarly, values within 0.1 and 0.3 are regarded as small correlations and values within 0.3 and 0.5 as medium correlation (ranges apply symmetrically to negative correlations)

[4] Face validity is the extent to which each instrument is viewed as covering and representing mental workload, the construct it purports to measure

		$NASATLX_{def}$		$NASATLX$		WP_{def}		WP	
Model	Cases	ρ	p_{val}	ρ	p_{val}	ρ	p_{val}	ρ	p_{val}
$NASATLX_{def}$	Total	1	0.000	0.995**	0.000	0.585**	0.000	0.586**	0.000
	Group A	1	0.000	0.994**	0.000	0.536**	0.000	0.537**	0.000
	Group B	1	0.000	0.996**	0.000	0.645**	0.000	0.648**	0.000
$NASATLX$	Total			1	0.000	0.582**	0.000	0.584**	0.000
	Group A			1	0.000	0.536**	0.000	0.537**	0.000
	Group B			1	0.000	0.638**	0.000	0.643**	0.000
WP_{def}	Total					1	0.000	0.992**	0.000
	Group A					1	0.000	0.992**	0.000
	Group B					1	0.000	0.991**	0.000

**Correlation is significant at the 0.01 level (2-tailed)

Table 6.7: Pearson correlation coefficient for the baseline instruments $NASATLX$, WP and their defeasible translations

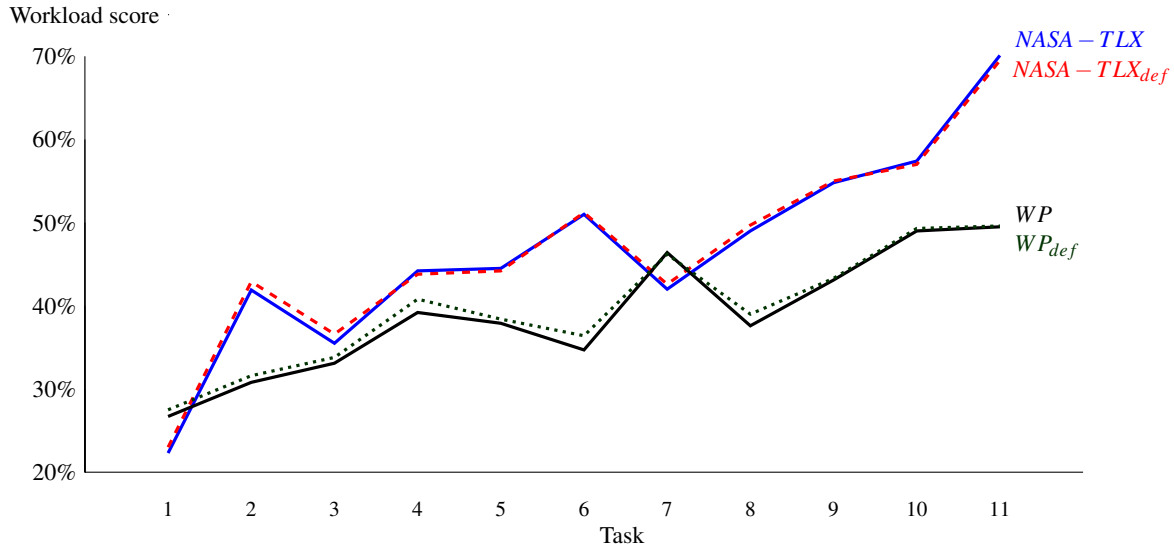


Fig. 6.3: Comparisons of the means of the workload scores produced by the baseline instruments $NASATLX$, WP and their defeasible translations - Group A

by plotting the means of the mental workload indexes, computed by each instrument, for each task and for each group of volunteers. A further note is that preferentiality among workload attributes is considered in the original NASA-TLX but not in the WP instrument: this difference is coherently accounted in the defeasible translations and coherently supported by the high convergent validity of the computed workload scores and the correspondent scores assessed by the original instruments.

Having a nearly perfect *convergent validity* suggests that the defeasible framework designed in chapter 4 can be successfully employed to abstract the two baseline mental workload assessment instruments that can be now represented with a common underlying structure employing the notion of defeasible interactive arguments.

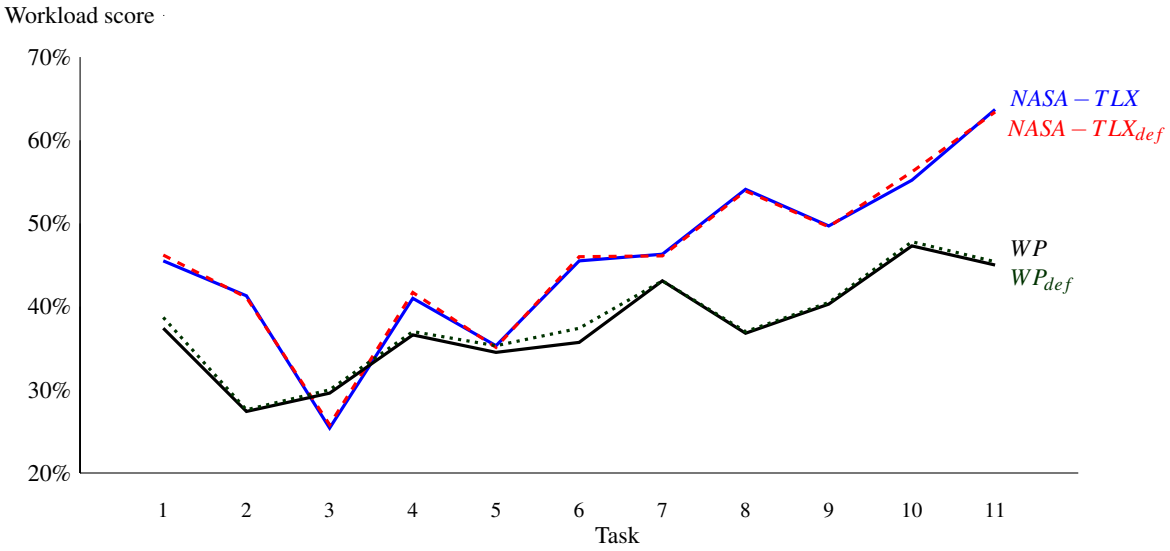


Fig. 6.4: Comparisons of the means of the workload scores produced by the baseline instruments $NASATLX$, WP and their defeasible translations - Group B

6.4 Evaluating the sensitivity, diagnosticity and validity of the brand new defeasible instances

The last step for completing the evaluation of the defeasible framework is to accomplish objective 5 by investigating the quality of the assessments produced by the brand new instances of the defeasible framework MWL_{def}. The goal is to demonstrate how this new instance, as previously constructed in chapter 5, can achieve a better *sensitivity* and *diagnosticity* than the baseline NASA-TLX and WP instruments, and is able to achieve a positive convergent validity with these two but a better concurrent validity with task completion time, being this a performance objective measure. Before introducing the results, each of these properties is clarified with a detailed description along with the formal procedure to measure it, as in table 6.8.

Property/Method	Description/Goal
Sensitivity	The reliability of a model to detect changes in resource demand, task difficulty, user features and environmental influence
ANOVA + PostHoc comparisons	To find out to what extent the global indices of mental workload varied as a function of objective changes and manipulation of tasks
Diagnosticity	The capacity of the model to quantify the contributions to workload by the type or resource demand or the capabilities of the human operator
Multinomial logistic regression	To determine to what extent mental workload attributes allow discrimination between tasks.
Validity	The capacity of the model to measure mental workload.
Pearson/Spearman correlations	*Convergent validity*: to determine to what extent the model measures what is supposed to measure; *Concurrent validity*: to determine to what extent the model is able to explain objective performance measure.

Table 6.8: Definition of mental workload properties and statistical methods applied

6.4.1 Results and discussion

The instance MWL_{def} of the framework, as built in section 5.2.3, is elicited multiple times, with different sets of inputs (answers of questionnaire A.5 of appendix) provided by each volunteer who participated in the study. As stated in objective 5, the aim is to compare the *sensitivity* and the *diagnosticity* capacities of the designed defeasible instance MWL_{def} against the baseline instruments NASA-TLX and WP, as well as comparing their *validity*.

Sensitivity

In order to test the *sensitivity* of the new instance MWL_{def}, a one-way analysis of variance (ANOVA) is adopted to determine whether there are any significant differences between the means of the independent tasks designed in table 6.2. This statistical procedure has a set of assumptions that have to be met before proceeding with the actual test. Instead of taking them as valid, they are explicitly tested.

- *continuity:* the dependent variables should be measured at the interval or ratio level. In other words, they have to be continuous;
- *independency of variables:* the independent variable should consist of two or more categorical, independent groups;
- *independency of observations:* there should be no relationship between the observations in each group or between the group themselves;
- *outliers:* there should be no outliers, that means single data points within the dataset that do not follow the usual pattern. In case these exist, they should be removed;
- *normality:* the dependent variables under consideration should be approximately normally distributed for each category of the independent variable.
- *homogeneity of variance:* there is the need of having homogeneity of variances.

The assumption of *continuity* is met because the indexes of mental workload computed by the new instance MWL_{def}, the NASA-TLX and the WP baseline instruments are in the scale $[0..100] \in \Re$. The assumption of *independency of variables* is met because there are more than 2 groups for the independent variables, represented by the task of table 6.2 (11 for group A and 11 for group B of volunteers, for a total of 22 groups). The assumption of *independency of observation* is met, despite the fact that each participant is in each cluster (task). The study conducted was not designed as a repeated measure study, because each task (each cluster) is different and independent to each other. So each participant did not execute the same task in different conditions, but each executed 11 different tasks in one unique condition. This was more a study design issue because of the difficulty of recruiting volunteers. Hence the study has been designed to gather more data with a less amount of volunteers. The assumption of no significant *outliers* is met by removing few outliers from the distributions of the computed mental workload scores for each instrument, as depicted in the box plots of section C.4 (appendix C). Here the circles outside a box plot, one for each task, has been removed from the dataset. In addition, if the resulting box plots still presented other outliers, the process was repeated until no outliers were found. The assumption of *normality* is met by conducting, for each distribution of the mental workload scores produced by each instrument, each task and group of volunteers, a Shapiro-Wilk normality test, with a 95% confidence interval. Results are listed in table C.3 (appendix C) and as it is

possible to observe, the significance value for each instrument, each task and each group is greater than 0.05, underlying the normality of the data. In the case it would have been below 0.05, the data would significantly have deviated from a normal distribution. Eventually, in order to check the assumption of *homogeneity of variances*, the Levene's test has been employed for each workload assessment instrument: results are listed in table 6.9. The test of homogeneity of variances test the null hypothesis that all the tasks have same variance:

$$H_0 : \delta^2{}_{T_1} = \delta^2{}_{T_2} = \delta^2{}_{T_3} = \delta^2{}_{T_4} = \delta^2{}_{T_5} = \delta^2{}_{T_6} = \delta^2{}_{T_7} = \delta^2{}_{T_8} = \delta^2{}_{T_9} = \delta^2{}_{T_{10}} = \delta^2{}_{T_{11}}$$

To interpret the outcomes of the Levene's test, the *Sig*. value (p-value) has been used. In the case it is less than or equal to the $\alpha = 0.05$ level for this test, then the hypothesis H_0 that the variances are equal can be rejected. On the other hand, if the *Sig*. value is greater than $\alpha = 0.05$ level, then the hypothesis H_0 cannot be rejected thus this increases the confidence that the variances are equal and the homogeneity of variance assumption has been met. From table 6.9 it is possible to note that for the NASA-TLX instrument, the null hypothesis has to be rejected, both for group A and group B, while for the WP and the new instances MWL^{NI}_{def} and MWL_{def} of the defeasible framework , the null hypothesis cannot be rejected, confirming that the assumption of homogeneity of variances is reasonably satisfied. In the first case ($NASATLX$), a Welch F-test [5] is added to the ANOVA procedure and the Games-Howell[6] post-hoc tests are carried out instead of the Tukey post-hoc tests [7]. In the other three cases (WP, MWL^{NI}_{def} and MWL_{def}) the classical ANOVA procedure is adopted, and the Tukey post-hoc test is conducted as all the assumptions are met.

	Group A				Group B			
model	Statistic	df1	df2	Sig.	Statistic	df1	df2	Sig.
$NASATLX$	1.899	10	206	0.047	4.054	10	203	0.000
WP	0.566	10	209	0.840	1.852	10	204	0.054
MWL^{NI}_{def}	1.141	10	207	0.333	1.196	10	203	0.295
MWL_{def}	1.261	10	207	0.254	1.443	10	205	0.164

Table 6.9: Levene's tests of homogeneity of variances of the mental workload assessment instruments

Tables 6.10 and 6.11 shows the results of the analysis of variance of the four workload assessment instruments, for both the groups of volunteers. As it is possible to note, there was a statistically significant difference between tasks as determined by one-way ANOVA for any mental workload instruments and for both the groups. All the *Sig*. values are really small (< 0.000) thus the null hypothesis of equal variances has to be rejected. Despite the ANOVA test underlines an overall difference between tasks, it does not tell which specific tasks are different. For this reason, as mentioned above, post-hoc tests are employed to confirm where the differences occurred between groups. For the NASA-TLX, the Games-Howell post-hoc tests are run, for the WP and the two instances MWL^{NI}_{def}, MWL_{def} of the defeasible framework, the standard Tukey's HSD is used.

[5] The Welch statistics is based on the usual ANOVA F test, but it is applied when the variances of the groups under examination are significantly different.

[6] The Games-Howell post-hoc test is used when variances of groups under examination are unequal. It takes into account unequal group sizes as well and it is based on Welch's correction to degrees of freedom.

[7] The Tukey's HSD (honestly significant difference) test is a single-step multiple comparison procedure and statistical test. It is usually adopted along with an ANOVA to find significant differences of means.

	Group A				
NASATLX	Sum of Squares	df	Mean Square	F	Sig.
Between groups	30796.056	10	3079.606	$F = 13.467$	< 0.000
Within groups	47107.517	206	228.677		
Total	77903.573	216			
Welch	df1=10, df2=82.329			13.106*	< 0.000
WP					
Between groups	11118.296	10	1111.830	$F = 5.182$	< 0.000
Within groups	44842.045	209	214.555		
Total	55960.341	219			
MWL_{def}^{NI}					
Between groups	19503.159	10	1950.316	$F = 11.289$	< 0.000
Within groups	35762.100	207	172.764		
Total	55265.259	217			
MWL_{def}					
Between groups	23548.588	10	2354.859	$F = 12.146$	< 0.000
Within groups	40121.481	207	2354.859		
Total	63680.068	217			

* Asymptotically F distributed

Table 6.10: Analysis of variances, Welch tests and significance values of the mental workload assessment instruments - Group A

	Group B				
NASATLX	Sum of Squares	df	Mean Square	F	Sig.
Between groups	21110.294	10	2111.029	$F = 7.212$	< 0.000
Within groups	59418	203	292.700		
Total	80528.495	213			
Welch	df1=10, df2=81.065			10.316*	< 0.000
WP					
Between groups	8005.668	10	204	$F = 5.649$	< 0.000
Within groups	28909.109	204	141.711		
Total	36914.777	214			
MWL_{def}^{NI}					
Between groups	12434.883	10	1247.488	$F = 5.962$	< 0.000
Within groups	42341.084	203	208.577		
Total	54775.968	213			
MWL_{def}					
Between groups	18428.857	10	1842.886	$F = 9.895$	< 0.000
Within groups	38180.637	205	186.247		
Total	56609.494	215			

* Asymptotically F distributed

Table 6.11: Analysis of variances, Welch tests and significance values of the mental workload assessment instruments - Group B

The post-hoc results are presented in section C.5 (appendix C, page 188). Tables C.9 to C.24 (pages 204-219) are detailed explanations of each comparison of each pair of tasks, while tables C.25 to C.26 (pages 220-221) are summaries of the statistically significant differences spotted by each instrument, both with a 95% and 99% confidence interval. In general, the ratio between-groups (tasks) and within-groups (participants) is higher with the NASA-TLX (F(10, 206) = 13.467, F(10,81.065)=10.316) and the instance MWL_{def} (F(10, 207) = 12.146, F(10,205)=9.895), underlying higher variance. The Workload Profile instrument is the assessment instrument with the lowest variance (F(10,209)=5.182, F(10,204)=5.649) followed by the instance MWL^{NI}_{def}. The defeasible instance MWL^{NI}_{def} showed mixed outcomes for the groups. In group A it behaved similarly than the NASA-TLX and the defeasible instance MWL_{def}, outperforming the WP instruments. However, in group B, it significantly under-performed the NASA-TLX and the instance MWL_{def} showing how the addition of the attack relations among arguments, from MWL^{NI}_{def} (no interactions) to MWL_{def} (interaction of arguments), had an important impact in increasing the sensitivity capacity.

From the summary of detected statistically significant differences of tables C.25 and C.26 (pages 220, 221), the WP instrument was the lowest in sensitivity, followed by the instance MWL^{NI}_{def}, NASA-TLX and the instance MWL_{def}. Table 6.12 summaries the number of detected statistically significance differences spotted among tasks. WP is the lowest in sensitivity, detecting half of the statistically significant differences spotted by the other instruments. For group A, the two defeasible instances MWL^{NI}_{def} and MWL_{def} behaved very analogously, demonstrating similar sensitivity with the NASA-TLX but a higher sensitivity for group B, using a confidence interval of 95%. However, when increasing the confidence interval to 99% the instance MWL_{def} was clearly superior than the NASA-TLX and MWL^{NI}_{def} underlying a higher degree of robustness and being more stable in detecting differences among tasks in both the groups.

	Group A		Group B	
Model	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
$NASATLX$	22	13	14	10
WP	9	5	8	6
MWL^{NI}_{Def}	21	17	11	7
MWL_{Def}	21	18	18	13

Table 6.12: Statistically significant differences detected by each workload assessment instrument

In summary, according to the number of detected statistical significant differences of table 6.12, out of all the possible detectable differences (110 - 55 for each group), *the instance* MWL_{def} of the defeasible framework shows 39.9% and 36.3% of sensitivity more than the WP instrument and 5.45% and 14.5% of sensitivity more than the NASA-TLX instrument, respectively at significance levels of 0.05 and 0.01. Additionally, MWL_{def} shows 12.7% and 12.7% more than the other instance MWL^{NI}_{def} (where no interaction of arguments was considered), respectively at significance levels of 0.05 and 0.01 (for the tasks of table 6.2).

Diagnosticity

In order to test the *diagnosticity* of the new instances of the defeasible framework, *stepwise multinomial logistic regression* has been used to investigate the differences between tasks on the basis of the workload attributes of the cases, indicating which attributes contributed the most to task separation. This technique is aimed at analysing relationships between a non-metric dependent variable (task) and metric independent variables (workload attributes) and it extends logistic regression as it compares multiple groups (tasks) through a combination of binary logistic regressions. Differently than other studies that employed discriminant analysis to assess the diagnosticity of some subjective mental workload assessment instruments (Rubio et al., 2004; Tsang and Velazquez, 1996), the rationale behind adopting multinomial logistic regression lies in the fact that it does not impose all the assumptions required by discriminant analysis. These include normality, linearity and homogeneity of variance for the independent variables. Section C.7 of appendix C shows how not all the distributions of the independent variables follow a clear normal distribution (Significance of the Shapiro-Wilk test is lower than the significance level of 0.05), justifying the choice of adopting multinomial logistic regression. The goal here is to compare the diagnosticity capacity of the defeasible instances MWL_{def} and MWL^{NI}_{def}, against the one of the NASA-TLX and WP baseline instruments, by adopting stepwise multinomial logistic regression and by determining the impact of multiple independent mental workload attributes to predict the membership of one or other of the 22 tasks (11 for group A and 11 for group B). Multinomial logistic regression has a set of assumptions that should be met for inferring trustworthy results:

- *sample size*: the minimum number of cases per independent variable is suggested to be 10, according to multiple studies in the field as suggested in (Peduzzi et al., 1996);
- *multicollinearity*: the independent variables should be fairly uncorrelated between each other, as it is difficult to differentiate each's variable impact on the dependent variable if they are highly correlated;

The *sample size* assumption is met as each task has 20 cases for evaluation. The assumption of *multicollinearity* is tested by analysing the correlations of the independent variables. Pearson's correlation coefficient, one for every pair of mental workload attributes, has been computed, as presented in section C.6 of appendix C. None of the pair of attributes shows a high correlation ($\rho < 0.8$) thus all the mental workload attributes are potentially contributors for predicting the tasks.

In multinomial logistic regression, the presence of a relationship between the dependent variable and combinations of independent variables is built upon the statistical significance of the final model chi-square. This is based on the reduction in the likelihood values for a model which does not contain any independent variable and the model that contains the independent variables. In turn, this difference in likelihood follows a chi-square distribution and its significance is the statistical evidence used for assessing the presence of a relationship between the task and the combination of mental workload attributes. Stepwise multinomial logistic regression is slightly different than the standard (ordinal) procedure because one independent variable is entered in turn to the reference model (the empty model). In this study a forward entry method and at each step has been adopted: the most significant independent variable is added to the model until none of the stepwise independent variables left out of the model would have a statistically significant contribution if added to the model.

	22 tasks - 11 Group A, 11 Group B			
	Fitting criteria	Likelihood ratio tests		
Model	$-2log$ Likelihood	Chi-Square	df	Sig.
Intercept Only	2720.117			
NATATLX attributes				
Final	2258.907	461.210	105	0.000
WP attributes				
Final	1773.885	946.233	168	0.000
MWL_{def} & MWL^{NI}_{def} attributes				
Final	1188.568	1531.550	357	0.000

Table 6.13: Model fitting information for each mental workload instrument

Table 6.13 shows the model fitting information and as it is possible to note, every *Sig*. value for every set of attributes, considered in each workload instrument, is less than the level of significance (< 0.05). The null hypothesis of no difference (Chi-square value) between the model without independent variables (intercept only) and the model with the independent variable (final) is rejected in every test. This underlines the existence of a relationship between the mental workload attributes and the tasks conducted. However, it does not tell where exactly these differences occurred as well as the errors associated with the model. In order to assess the utility of a multinomial logistic regression model, its classification accuracy is computed. This compares the predicted task membership of the logistic model to the actual (the known one), which is the value for the dependent variable. In order to evaluate the usefulness of the logistic regression model, a benchmark of 25% improvement over the rate of accuracy achievable by chance alone is used. In other words, even if it is assumed that the independent mental workload attributes had no relationship to the tasks defined by the dependent variable, it is still expected to be correct in the predictions of task membership some percentage of the time.

The estimate of by-chance accuracy used is the proportional by-chance accuracy rate computed by summing the squared percentage of cases in each group ($20/440 = 4.5\%$). Thus the proportional by-chance accuracy criteria is 5.56% ($0.045^2 \times 22 \times 1.25 = 5.56\%$). Table 6.14 summaries the classification accuracy rates computed by each logistic regression model. As it is possible to note, all of these rates are above 5.56% satisfying the criteria for classification accuracy. Section C.9 of appendix C shows the detailed predicted task memberships with the actual task memberships for any instrument. Using the mental workload attributes of the NASA-TLX instrument, a prediction accuracy of 19.1% was achieved compared to the 32% achieved with the attributes of the WP instrument and a 53.2% accuracy achieved with the workload attributes considered in the two defeasible instances MWL_{def} and MWL^{NI}_{def}.

95% CI	Prediction accuracy
NASATLX attributes	19.1%
WP attributes	32.3%
MWL_{def} & MWL^{NI}_{def} attributes	53.2%

Table 6.14: Accuracy of each regression model for the workload attributes of each assessment instrument

These accuracies reflect the combination of a set of attributes for correctly classifying each task considered in each case. However, they cannot tell anything about the contribution of an individual independent mental workload attribute to the overall classification. The interpretation for an independent mental workload attribute focuses on its ability to distinguish between pairs of tasks and the contribution which it makes to changing the probability of being in one dependent task rather than the other. The significance of an independent mental workload variable's role in distinguishing between pairs of tasks should not be interpreted unless it has also an overall relationship to the dependent variable (task) in the likelihood ratio tests. These tests are listed in section C.8 of appendix C. From table C.31 it is possible to note how all the attributes considered in the NASA-TLX show a statistically significant relationship with the dependent variable (task) as all the *Sig*. values are less than the level of significance (< 0.05). The same interpretation applies for the attributes considered in the WP instrument whose results are depicted in table C.32. All the *Sig*. values are less than the level of significance (< 0.05) supporting the fact that each of them has an influential role in classifying each case's task. Regarding the instances of the defeasible framework MWL_{def} and MWL_{def}^{NI}, table C.33 shows the likelihood ratio tests. Here, the attributes all have a significance value less than 0.05, but the *mental demand* and *intention* attributes are not included, as not considered significant to classify tasks by the step-wise multinomial logistic regression procedure.

The information associated to the likelihood ratio tests tells which variable has an overall relationship to the dependent variable, considering all the tasks. However, it does not tell the individual strength of each workload attribute for classifying tasks. Section C.10 of appendix C lists the step summary tables for each multinomial logistic regression procedure of each workload assessment instruments. From these it is possible to analyse in which order and what workload attribute is entered in the empty multinomial logistic regression model (including just the intercept) and the contributions that each attribute had to the model's goodness-of-fit.

In the case of the attributes accounted in the original NASA-TLX, *temporal demand*, *effort* and *performance* were the most significant contributors as their addition, at each step, reduced the chi-square significantly. Table C.37 shows how *temporal demand* reduced the chi-square of 2720.117 to 2579.573 in turn reduced by *effort* to 2446.946 and in turn reduced by the *performance* to 2337.516. *Psychological stress* and *mental demand*, although valid contributors, they had a less powerful role in reducing the chi-square.

Regarding the attributes accounted in WP instrument, all had a significant effect in reducing the chi-square. From table C.38, the attribute *auditory resources* was the most impact full in reducing the chi-square, followed by *central processing* and *manual response*. The attribute *visual resources* was the last contributor to the model's goodnes-of-fit.

Eventually, the attributes considered in the two instances MWL_{def} and MWL_{def}^{NI} of the defeasible framework, all had a significant role in reducing the chi-square of the intercept model (empty model), except the attributes *mental demand* and *intention* that were not used. From table C.39, the most impact-full contributor was *auditory resources*, followed by *parallelism*, *temporal demand* and *effort*. The attributes with the lowest power in increasing the goodness-of-fit were *arousal* and *central processing*.

In summary, the attributes taken into account in the instances of the defeasible framework show a greater diagnosticity capacity compared to the one achieved by the attributes of the NASA-TLX and the WP instruments, in terms of ability to classify each case in the right category (one of the web-tasks). Considering the set of web-tasks listed in table 6.2 the instances of the defeasible framework had an accuracy rate 34.1% higher than the NASA-TLX instrument and 20.9% higher than the WP instrument confirming its prospective in assessing subjective mental workload.

Validity

In order to test the *validity* of the new instances of the defeasible framework, the correlations of the workload scores computed by the four mental workload instruments (NASA-TLX, WP and MWL_{def}, MWL^{NI}_{def}) and the correlations of their workload scores again objective performance measure have been computed. The former is referred to as *convergent validity* while the latter is referred to as *concurrent validity*, both assessed using Pearson's correlation coefficients and Spearman's rank correlation coefficients [8]. The objective performance measure employed for testing convergent validity is the objective *time* participants required for completing each task. Unfortunately, some cases do not have an associated time due to errors in the measurements. Table 6.15 lists the correlations for *convergent validity* and *concurrent validity* considering all the tasks used in the experiments. Similarly table 6.16 shows the correlations excluding those task with imposed time-limit (tasks 3, 6, 11 of table 6.2).

		Pearson					Spearman				
		$NASATLX$	WP	MWL^{NI}_{def}	MWL_{def}	Time	$NASATLX$	WP	MWL^{NI}_{def}	MWL_{def}	Time
$NASATLX$	Correlation	1	.584	.562	.778	.315	1	.571	.579	.780	.335
	Sig.		.000	.000	.000	.000		.000	.000	.000	.000
	Cases		440	440	440	352		440	440	440	352
WP	Correlation		1	.654	.859	.264		1	.658	.854	.259
	Sig.			.000	000	.000			.000	.000	.000
	Cases			440	440	352			440	440	352
MWL^{NI}_{def}	Correlation			1	.713	.272			1	.738	.250
	Sig.				.000	.000				.000	.000
	Cases				440	352				440	352
MWL_{def}	Correlation				1	.381				1	.346
	Sig.					.000					.000
	Cases					352					352

Table 6.15: Pearson's and Spearman correlation coefficients between mental workload scores against each other and against time

The *convergent validity* of the mental workload instruments is significant, with the instance MWL_{def} highly correlating with the NASA-TLX and WP both according to Pearson and Spearman correlations coefficients (Pearson: .778, .859 considering all the tasks, .763, .856 not considering time-limit tasks - Spearman: .780, .854 considering all the tasks. .761, .853 not considering time-limit tasks). The NASA-TLX and the WP showed a moderate positive correlation (Pearson: .584 considering all the tasks, .590, not

[8] The Spearman's rank correlation coefficient is a nonparametric measure of statistical dependence between two variables. Likewise the Pearson's correlation coefficient, it tells how the relationship between two variables can be described using a monotonic function, but upon the ranked variables.

		Pearson					Spearman				
		NASATLX	*WP*	MWL_{def}^{NI}	MWL_{def}	Time	*NASATLX*	*WP*	MWL_{def}^{NI}	MWL_{def}	Time
NASATLX	Correlation	1	.590	.597	.763	.384	1	.571	.623	.761	.369
	Sig.		.000	.000	.000	.000		.000	.000	.000	.000
	Cases		320	320	320	248		320	320	320	248
WP	Correlation		1	.679	.856	.305		1	.681	.853	.286
	Sig.			.000	.000	.000			.000	.000	.000
	Cases			320	320	248			320	320	248
MWL_{def}^{NI}	Correlation			1	.752	.344			1	.779	.333
	Sig.				.000	.000				.000	.000
	Cases				320	248				320	248
MWL_{def}	Correlation				1	.447				1	.392
	Sig.					.000					.000
	Cases					248					248

Table 6.16: Pearson's and Spearman correlation coefficients between mental workload scores against each other and against time - No time-limit tasks

considering time-limit tasks - Spearman: .571 considering all the tasks, .571 not considering time-limit tasks).The instance MWL_{def}^{NI} of the framework with no interaction of arguments only moderately correlated to the NASA-TLX and WP instruments (Pearson: .562, .654 considering all the tasks, .597, .679 not considering time-limit tasks - Spearman: .579, .658 considering all the tasks. .623, .681 not considering time-limit tasks) having less convergent validity than its counterpart with interactions of arguments (MWL_{def}). All the coefficients are statistically significant.

Regarding the *concurrent validity*, the instance MWL_{def} of the defeasible framework correlates better with time, showing a moderate positive correlation (Pearson: .381 considering all the tasks, .447, not considering time-limit tasks - Spearman: .346 considering all the tasks. .392 not considering time-limit tasks) than the NASA-TLX (Pearson: .315 considering all the tasks, .384, not considering time-limit tasks - Spearman: .335 considering all the tasks. .369 not considering time-limit tasks), the WP instrument (Pearson: .264 considering all the tasks, .305, not considering time-limit tasks - Spearman: .259 considering all the tasks. .286 not considering time-limit tasks) and the instance MWL_{def}^{NI} of the defeasible framework with no interaction of arguments (Pearson: .272 considering all the tasks, .344, not considering time-limit tasks - Spearman: .250 considering all the tasks. .333 not considering time-limit tasks). All the correlation coefficients are statistically significant.

In summary, the instance MWL_{def}, as hypothesised, shows a high convergent validity with the Nasa Task load Index and the Workload Profile instruments and has a better concurrent validity against the objective time than the other two baseline instruments and its counterpart MWL_{def}^{NI} (with no interaction of arguments). The fact that MWL_{def} can, in general, correlate with the objective time better than MWL_{def}^{NI} highlights the importance of incorporating theoretical relationships of workload attributes within a mental workload assessment instrument.

Chapter 7

Discussion

In this chapter the main findings are summarised, showing how the research question behind the thesis can be positively answered. The assumption of treating mental workload as a defeasible phenomenon, according to the outcomes of the evaluation chapter, seems to be valid. The impact of argumentation theory for mental workload representation is supported by the fact that instances of the framework were successfully built to replicate two well known state-of-the-art subjective mental workload assessment instruments, maintaining the same assessment capacity. Similarly, the impact of argumentation theory for mental workload assessment is confirmed by the comparative analysis of the assessments produced by a brand new instance, built with the framework, against the same two state-of-the-art assessment techniques, in terms of sensitivity, diagnosticity and validity of assessments. Below we present the advantages and limitations related to the use and application of this framework. Eventually, a set of illustrative case-studies is provided which aims to show how the assessments produced by the brand new instance which has been designed could be practically employed in the field of human-web interaction for supporting A/B testing of interfaces and customisation of systems.

7.1 Impact of argumentation theory for workload representation

The defeasible framework for knowledge representation, designed in chapter 4, has been proved to be appealing for mental workload representation. Sections 5.2.1 and 5.2.2 have shown how two state-of-the-art subjective mental workload assessment techniques can be successfully translated into two particular instances of the defeasible framework. These are the NASA Task Load Index (NASA-TLX), built by the Human Performance Group at NASA's Ames Research Centre (Hart and Staveland, 1988), (Hart, 2006), and the Workload Profile (WP), designed by psychologists (Tsang and Velazquez, 1996) and based on Wickens's multiple-resource theory (Wickens and Hollands, 1999), (Wickens, 2008), as described in section 2.1.2 (page 12). NASA-TLX and WP are two multi-dimensional instruments that consider multiple workload attributes for shaping mental workload. However, the former includes a form of preferentiality of attributes leading to a weighted aggregation of the values carried by each of these attributes. The latter does not take account of preferentiality and the attributes are simply summed to obtain a mental workload index. These features have been successfully maintained in the two instances built with the defeasible framework showing how their assessment positively correlates with the assessment of the respective original instruments. This high convergent validity between

the original instruments and their defeasible translations was close to 1, indicating how all assess the same theoretical concept, though this concept was shaped in different ways. Having a nearly perfect convergent validity suggests that the defeasible framework was useful to build particular instances able to reproduce the original NASA-TLX and the WP instruments, using the notion of defeasible arguments, thus modelling these with a common underlying structure that can be better compared. In addition, in example 13 (page 81), it has been shown how the defeasible framework could be also applied for translating the Rating Scale Mental Effort, a uni-dimensional workload assessment instrument (Zijlstra, 1993), into defeasible arguments. The main advantage of the defeasible framework, compared to the aforementioned multi-dimensional or uni-dimensional workload instruments, is represented by its capacity to model workload attributes as defeasible arguments. These are special structured pieces of evidence, in the form of premises-conclusion, that claim a workload degree that is open to retraction by other arguments. This retraction is possible because of the interaction of defeasible arguments, manifested as attack relations. This property represents the most important difference with respect to state-of-the-arts assessment techniques. However, interaction of arguments might generate inconsistent and contradictory points of view that can be considered for workload assessment. These inconsistencies are handled by acceptability semantics (section 4.6, page 92), computational procedures that generate one or more conflict-free sub-sets of the same initial input argument set. A strategy for selecting the most credible set represents the final step in workload assessment. The overall methodology starting from knowledge-base translation, into a form of interactive defeasible arguments and their manipulation towards a mental workload assessment, is appealing because of its modularity and explanatory capacity.

7.2 Impact of argumentation theory for workload assessment

Enhancement in the quality of workload assessment has been proved by designing two new instances of the defeasible framework (as in section 5.2.3), with and without interaction of arguments. These instances consider a wider set of workload attributes, compared to the NASA-TLX and WP, these being baselines for the evaluation of the defeasible framework. In line with the majority of the research aimed at comparing the psychometric properties of different mental workload assessment instruments, this thesis focused on the evaluation of the defeasible framework with respect to sensitivity, diagnosticity and validity. In general, most workload designers assume that in subjective techniques, the intrusiveness of an instrument is not a significant problem (Rubio et al., 2004). In fact, these techniques usually requires the completion of rating scales subsequent to task performance, and is therefore not intrusive. In the experimental studies conducted in this thesis, subjects were required to fill in the questionnaire presented in appendix A.6 which includes the original NASA-TLX and WP questions plus a few more questions concerning other attributes believed to influence mental workload (by the author of this thesis). In addition, subjects were required to perform the pair-wise comparisons of the 6 dimensions accounted for in the NASA-TLX. In terms of number of questions, in the NASA-TLX 21 overall rates were required (6 questions + 15 comparisons) while for the instance of the defeasible framework 19 questions were required (no comparisons). As a consequence it can be assumed that there is no significant difference in gathering rates for quantifying the workload attributes behind each assessment instrument. In turn intrusiveness is assumed to be the same. However, the questionnaire behind the WP can be assumed to be less intrusive as just 8 answers are required.

7.2.1 Sensitivity

Taking into account sensitivity, analysis of variance has been used to assess the capacity of each assessment instrument to detect changes in resource demands, task difficulty, user features and environmental influence, these being the different task conditions. The goal was to find out to what extent the global indices of mental workload varied as a function of objective changes and manipulation of tasks. The instance of the defeasible framework MWL_{def} (with interaction of arguments) had an $F - ratio$ similar to the NASA-TLX showing the same pattern of sensitivity. The WP and the instance of the defeasible framework MWL^{NI}_{def} (with no interaction of arguments) showed nearly half of the above sensitivity. However, taking a closer look at each comparison between tasks, the statistically significant differences spotted by each instrument demonstrated how MWL_{def} was always superior to the other three instruments. In fact, this instance of the framework was more accurate in detecting changes, this being confirmed by the stability in spotting statistically significant differences between tasks when increasing the confidence interval from 95% to 99%. Quantification of such stability demonstrated how, considering the tasks used in the experimental studies, MWL_{def} was from 5.45% to 39.9% more accurate than the other instruments. In addition, the high variance (F-ratio) of the NASA-TLX was justified by the computed workload indexes, that were more spread in the range $[0..100] \in \Re$ and not because the instrument was really able to detect all the differences between tasks.

7.2.2 Diagnosticity

In order to measure the diagnosticity of the designed instance of the defeasible framework MWL_{def}, which is still compared to the NASA-TLX and the WP, step-wise multinomial logistic regression has been employed. A few previous research studies have assessed diagnosticity by adopting discriminant function analysis (Tsang and Velazquez, 1996) (Rubio et al., 2004). The rationale behind adopting such a different statistical technique was driven by the fact that multinomial logistic regression does not require all the statistical assumptions of the underlying data required by discriminant function analysis. The dataset gathered during the experimental studies justified the use of this tool showing how these statistical assumptions were not always met. Diagnosticity indicated the capacity of a workload assessment instrument to quantify the contributions to mental workload. The goal was to determine to what extent the mental workload attributes taken into account by each assessment instrument allowed discrimination between tasks as well as how each single attribute contributed to task separation. The former was assessed by investigating the difference between the multinomial empty logistic model (just with the intercept) and the same model with the workload attributes. The latter was assessed by using the step-wise method and by analysing what workload attributes are entered in the empty model and in which order, as well as the contribution that each of these attributes had to the model's goodness-of-fit.

Findings underlined the existence of a relationship between the mental workload attributes of each instrument and the experimental tasks conducted. However, in order to assess the utility of each multinomial logistic regression model, its classification accuracy was computed. The aim was to compare the predicted task membership of the logistic model to the actual (known) one. Results showed how the attributes accounted for in the instances MWL_{def} (the same as the instance MWL^{NI}_{def}) were clearly superior in discriminating tasks,

compared to the NASA-TLX and the WP instruments, having 53.2%, 19.1% and 32.3% of accuracy respectively. These figures are in line with expectations. In fact, the attributes taken into account in the instance MWL_{def} are a combination of the attributes accounted for in the original NASA-TLX and WP. Thus, it was reasonable to expect that this hybrid model demonstrated higher diagnosticity capacity. However, in MWL_{def}, six additional workload attributes were included, so in order to test whether each of these additional attributes contributed to the overall classification, further analyses were performed. These included the investigation of the likely ratio-tests behind the multinomial logistic regression procedure. Analysis showed how each of the attributes accounted for in the NASA-TLX and in the WP showed a statistically significant relationship with the dependent variable (task). However, this was not the case with all the attributes accounted for in the instance MWL_{def}: the attributes 'mental demand' and 'intention' were not considered significant to classify tasks by the step-wise multinomial logistic regression procedure. However, the new added attributes 'parallelism', 'context bias', 'past knowledge', 'skills' and 'arousal' all had a role for task separation. This suggests how additional workload attributes can increase the diagnosticity of mental workload assessment. In turn, it suggests that the flexibility of the framework in allowing a MWL designer to incorporate new MWL attributes is a positive and appealing property.

7.2.3 Validity

Regarding the validity of the instances built with the defeasible framework, comparisons with the NASA-TLX and WP were drawn to assess the capacity to measure mental workload. These two subjective state-of-the-art instruments were used again as a baseline to measure their convergent validity with the instance MWL_{def} and the instance MWL^{NI}_{def} (where no interaction of arguments was accounted). Results were encouraging, showing how both the instances of the defeasible framework positively correlated with the NASA-TLX and WP, demonstrating that all of them are able to assess the same theoretical concept. However, the MWL_{def} (where interaction of arguments was taken into account) showed a greater validity than its counterpart MWL^{NI}_{def} (where no interaction was accounted). This confirms the importance of the consideration of relationships among those pieces of evidence and knowledge believed to influence overall mental workload, as hypothesised in this thesis. Convergent validity, however, did not say anything about the superiority of the instance MWL_{def} to explain objective performance measure. In order to assess whether an assessment technique is capable of justifying the objective performance related to the execution of tasks, the concurrent validity of each assessment instrument with the objective time for task completion was investigated. This was assessed by analysing the correlation of the global workload scores computed by each instrument, and the objective time of task completion. Results showed how the instance MWL_{def} was always the most accurate instrument for justifying objective time, followed by the NASA-TLX, the instance MWL^{NI}_{def} and the WP, both when all the tasks were considered and when just the task with no-time limit taken into account. This achievement is important because it underlines how objective user's behaviour on tasks can be better explained by building particular instances of the defeasible framework. In turn, assessments produced by these instances provide an alternative and accurate way of designing interactive systems and technologies that satisfies end-users and optimises their experience.

7.2.4 Summary of findings

The defeasible framework seems to deliver encouraging outcomes according to the preliminary applications conducted on selected web-based tasks. The instance MWL_{def} built with this defeasible framework is superior in terms of sensitivity, diagnosticity and validity compared to the NASA-TLX and WP, these being state-of-the-art subjective assessment techniques used in the field, as was noted in our review of the literature on mental workload. However, assessments of MWL with two instruments delivered mixed outcomes according to the experimental studies conducted in this thesis, as summarised in figure 7.1. The NASA-TLX that demonstrated a higher sensitivity, is the weakest in term of diagnosticity. Similarly, the WP that showed the lowest sensitivity, was much better in terms of diagnosticity. The MWL_{def} out-performed these two instruments both in sensitivity and diagnosticity. Regarding validity, the three instruments showed a positive convergent validity, indicating that all assess the same underlying theoretical concept (mental workload). In term of concurrent validity with objective performance (time), the MWL_{def} was the best in justifying the objective time spent on tasks, although this was very similar to the NASA-TLX. The WP demonstrated the lowest concurrent validity with time. Eventually, the difference between the instances MWL_{def} and MWL^{NI}_{def} became clear, confirming that taking into account relationships of those pieces of knowledge believed to influence mental workload (built as defeasible arguments), has an important impact on workload assessment. Thus, we can arrive at an answer to the research question behind this thesis: argumentation theory can enhance the representation of the construct of mental workload and improve the quality of its assessment in the field of human-computer interaction. Certainly this statement is confined to this thesis and it cannot be claimed yet having a general applicability, but it is promising and encouraging, suggesting that tackling mental workload as a defeasible phenomenon is worthy of further research.

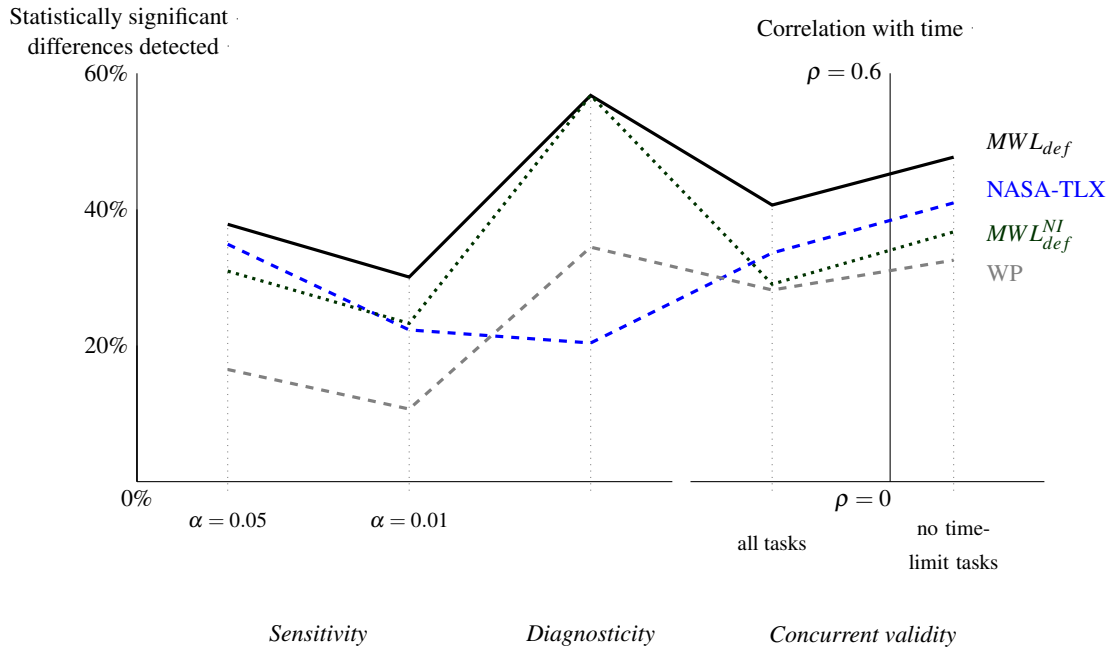


Fig. 7.1: Comparison of sensitivity, diagnosticity and validity of state-of-the-art subjective mental workload assessment instruments, and the defeasible instances built with the defeasible framework

7.3 Advantages and limitations of the defeasible framework

The main advantage of the defeasible framework, designed in Chapter 4, is represented by its flexibility and extensibility. It allows a workload designer to incorporate those attributes believed to be useful for shaping mental workload and it provides a methodology for reasoning upon them. This methodology includes the translation a designer's knowledge-base and beliefs into interacting defeasible arguments and the resolution of the inconsistencies and contradictions that might arise from their interaction (by applying formal acceptability semantics). This translation process is modular and based upon natural language terms familiar to a workload designer, aimed at enhancing intuitiveness. In addition, the outcomes of the application of an acceptability semantics are conflict-free sets of the same input arguments, thus easily interpretable. Another advantage is that a designer's knowledge-base, once translated into interactive defeasible arguments, has to be elicited and activated with objective inputs, related to the user, the task and the context of application. This elicitation process defines a sub-set of the same input arguments that can be evaluated by applying an acceptability semantic. In the event of input being corrupted or missing in a given scenario, the process towards mental workload assessment is not abandoned as a whole, but simply performed upon those arguments that are actually activated with available inputs. The defeasible framework does not make any assumption about the quantification of each workload attribute included in a given knowledge-base. In other words, inputs can be either quantified through subjective ratings, performance measures such as task completion time, objective measures such as physiological measurement or combined measures. The defeasible framework allows the translation of different knowledge-bases, each defining a particular instance and built with the same methodology. Each instance represents a proper test aimed at defining and controlling mental workload. These tests can be compared and they can be seen as attempts to falsify each other. In turn, something that is falsifiable does not mean it is negative; rather it is a positive quality because it can be tested with empirical experimentation in different contexts, thus increasing the understanding of mental workload, its representation, assessment and application.

7.3.1 Differences with machine learning and fuzzy logic

One of the main disadvantages of using such a defeasible framework is that it requires an initial effort for the translation of a knowledge-base into interactive defeasible arguments. This effort is more acute when several pieces of evidence need to be taken into account. Argumentation theory AT, has emerged in the AI community, as a paradigm that contrasts classic Machine Learning (ML) techniques (Longo and Hederman, 2013). The former is a knowledge-based paradigm while the latter techniques are learning-based paradigms. The main difference is that AT is not capable of learning from examples, a property that instead characterises ML. AT is an approach for supporting the translation of a knowledge-base and making inferences from it while ML is instead an approach to automatically learn relationships between pieces of evidence and/or inferring/predicting something from them according to previous examples and cases. Another limitation of the defeasible framework is that it does not consider mental workload over the time line. In other words, the assessment of an index of workload refers to a particular instance of time and it does not account for how it evolved over the execution of the task. The instance MWL_{def} of the framework designed in chapter 4 is based upon subjective ratings, inputs explicitly gathered from end-users through the use of a questionnaire

submitted at the end of a task, just once. In turn, the workload experienced by an end-user is quantified by this instance just once, with this unique set of subjective ratings. In the event that the questionnaire is submitted multiple times within the execution of a task, multiple indexes of mental workload can be assessed and combined to consider their evolution over time. For example, this could be done by using the approach proposed in (Xie and Salvendy, 2000b) and described in section 2.4.4 (page 36). However, requiring subjects to provide explicit ratings multiple times while executing a task can be very intrusive, with evident deterioration in the quality of workload assessments. Objective performance measures can be used to reduce intrusiveness and the assessment of mental workload can be done at multiple occasions within the time interval required for task completion. Multiple assessment is possible because performance measures are usually implicitly gathered, without requiring the explicit intervention of raters. This property is extremely appealing, above all in modern human-computer interactive settings, where inputs cannot always be gathered using subjective ratings. However, the evidence implicitly obtainable is usually smaller than the evidence that can be subjectively provided. Thus the issue is how to assess a meaningful mental workload score with this limited amount of evidence. The defeasible framework designed in chapter 4 is appropriate to tackle this issue because it can be applied with any kind and amount of evidence, with no assumptions about how it has been gathered. In order to clarify how the defeasible framework might be employed only considering measures implicitly gathered (not subjective ratings), an illustrative design problem is introduced (example 17). In this example, only a sub-set of the attributes accounted for in the instance MWL_{def} of the defeasible framework, as proposed in chapter 6 (page 121), is used.

Example 17

A web-designer wants to investigate how two different interfaces of a web-search system impose mental workload on end-users. The goal is to choose the interface that minimises imposed workload given a set of typical search-based tasks. The designer is not willing to use subjective ratings; instead a plug-in for a web-browser would be developed aimed at automatically gathering end-users actions such as clicking, scrolling, use of keyboard and timestamp of each of these actions. Information more strictly related to end-users such as skills, past-knowledge, arousal, frustration and motivation is not available in this context, nor is data related to environmental factors. However, skills and past knowledge might be subjectively provided by end-users just once, when registering with the web-search system. The attributes 'selection of response', 'auditory resources', 'speech response', as considered in the original WP instrument, are assumed to be low and they can be set to a low degree. The attribute 'visual resources', on the other hand, is assumed to be relatively high, while 'task/space' is not taken into account as it is not possible to objectively assess how much spatial attention an end-user will pay while executing a task.

Other workload attributes that the designer is willing to consider are:

- effort (quantified through an analysis of the usage of keyboard and mouse over time)
- mental demand (quantified according to the degree of difficulty of different designed tasks)
- temporal demand (quantified through an analysis of the objective time spent on task)
- performance (quantified through an analysis of the degree of accomplishment of the task)
- parallelism (quantified by analysing the keyboard/mouse usage on different unrelated parallel tasks)

- solving/deciding (quantified according to the degree of decision-making/problem-solving/remembering required by each designed task)
- verbal material (quantified according to the verbal material required by each task to process by end-users)
- manual response (quantified by independently analysing the usage of mouse and keyboard)

This knowledge-base can be translated into defeasible interactive arguments by using the framework designed in chapter 4. In turn these interacting arguments are evaluated with objective evidence gathered by the web-plugin and quantified appropriately according to some criteria or ad-hoc formula decided in advance by the designer. The activated argumentation framework is abstractly evaluated by running acceptability semantics for the extraction of consistent and conflict-free extensions of arguments. From these, the most credible is extracted and the assessment of workload can be performed according to the procedure summarised in figure 5.1 (page 97).

Example 17 shows how the designed defeasible framework could also be applied in settings where available evidence is partial and sometimes incomplete. This is an important property that differentiates the solution proposed in the thesis from state-of-the-art subjective assessment techniques. The latter require full quantification of the workload attributes accounted for by the instrument, while the former can work even if some attribute cannot be quantified in a given scenario. The main advantages behind the defeasible framework are multiple and they are summarised as follows.

- *inconsistency and incompleteness*: the defeasible framework provides a methodology for reasoning on available evidence, related to mental workload, even if such evidence is partial and inconsistent. Missing inputs are simply discarded and even if an argument, built upon some workload attribute, cannot be elicited, the argumentative process can still be executed with the remaining evidence;
- *extensibility and updatability*: the defeasible framework is an open and extensible solution that allows the retraction of a workload assessment in the light of new evidence. An argumentation framework that emerges from the translation of a knowledge-base can be updated with new arguments and evidence, when available;
- *expertise and uncertainty*: the defeasible framework captures knowledge in an organised fashion; it is able to handle the uncertainty and the vagueness associated with the mental workload evidence, usually expressed with natural language propositions and statements;
- *intuitiveness*: the defeasible framework is not based on statistics or probability, and it allows to reason on mental workload similarly to the way humans reason. If a workload designer is anyway inclined to use statistical evidence, for instance associated with some workload attribute, this can be modelled as an argument that can be embedded in an argumentation framework. In addition, vague knowledge-bases can be structured as arguments, built upon the familiar linguistic terms of the workload designer;
- *explainability*: the defeasible framework leads to explanatory reasoning thanks to the modular and incremental way of reasoning with available knowledge related to workload;

- *knowledge-bases comparability*: the defeasible framework allows comparisons of different subjective knowledge-bases. Two workload designers can build their own argumentation framework and identify differences in the definition of their interactive arguments. This property is appealing because it supports and increases our understanding of the construct of mental workload.
- *dataset independency*: the defeasible framework does not require a complete dataset to run, and it might be useful for emerging and not fully structured knowledge, such as in the field of mental workload, where evidence is fragmented or has not yet been gathered;

The main limitations are summarised as follows:

- *knowledge-base translation*: the initial translation of the knowledge-base of a workload designer into interactive defeasible arguments may require considerable effort, particularly when several pieces of evidence are considered
- *lack of learning*: the defeasible framework is not a learning-based paradigm. The inference rules underlying each argument as well as the relations between arguments cannot be automatically detected as in machine learning. However, machine learning often relies on big datasets of evidence to extract meaningful rules and this process might require valuable time for learning.
- *acceptability*: the framework has not been tested on other designers, thus its acceptability has not been proved yet. This is part of the future work and it represents a very important task. The goal is to maximise operator acceptance without being onerous.

Behind the advantages and limitations of the solution proposed in this thesis, the ultimate goal of assessing mental workload is to have a new metric that can be used for supporting and enhancing the design of human-computer interactive systems. In the following section, two case studies are presented, based on the experiments and assessments produced by the brand new instance designed in chapter 6.

7.4 Case-studies - A/B testings

Indexes of mental workload can be applied for many different reasons, as reviewed in 2.5 (page 40). This section concludes the discussion illustrating few case studies (A/B testings[1]), based on the experiments conducted in chapter 6, in the field of human-web interaction. The comparison made between two versions of the same system, in each case study, might be argued to be not to be fully appropriate or meaningful. However, the goal here is just to illustrate how these different versionss employing mental workload indexes may be evaluated.

7.4.1 Enhancing web-search

In web-based systems, such as Wikipedia, where a huge amount of content is delivered, the lack of a simple small search-box might be vital for the success of the system itself. Similar elements that help users to navigate within the web-site, such as a side-menu, might affect a user's experience as well. Consider task 1

[1] A/B testing is a methodology of using randomised experiments with two variants, A and B. The former is the control while the latter is the treatment in the controlled experiment. Two versions (A and B) of a system are compared, which are identical except for one variation that might impact on users' behaviour.

on list 6.2 'T_1: Find out how many people live in Sidney', performed by two groups of subjects (A and B) on the original Wikipedia.com and on a slightly modified version, as per screenshots of figures B.1 (Appendix B). The altered version does not have a search box in the top-right corner and does not have the navigation left-side menu, with a different background. An independent sample T-test can be run over the indexes computed by the instance MWL_{def} of the defeasible framework to determine whether the original Wikipedia interface (used by subjects in group A) imposed more or less statistically significant mental workload on end-users than the altered version (used by subjects in group B). The null hypothesis is the equality of workload indexes computed for the two groups.

$$H_0 : MWL^A_{def} = MWL^B_{def}$$

$$H_1 : MWL^A_{def} \neq MWL^B_{def}$$

Table 7.1 shows the group statistics, while table 7.2 shows the T-test: the assumptions behind this test are met[2] except the homogeneity of variance (the sig. value of .028 is less than the $\alpha = 0.05$ level, thus the null hypothesis of equal variability of the two groups is rejected). As a consequence the T-test of the second row (equal variance not assumed) is interpreted. The t-value of -1.612 has a significance ≤ 0.0001 which is less than the $\alpha = 0.05$, thus the null hypothesis H_0 is rejected.

Group	N	Mean	Std. dev.	Std. Error mean
A	19	19.68	9.34	2.14
B	2	37.45	15.64	3.49

Table 7.1: Group statistics for the Wikipedia illustrative example

	Leven's Test for equality of variances		T-test for equality of means						
	F	Sig.	t	df	Sig (2-tailed)	Mean diff.	Std. err diff.	95% confidence interval of the diff	
MWL_{def}								Lower	Upper
Equal variances	5.240	.028	-4.278	37	.000	-17.770	4.154	-26.187	-9.353
Not equal variances			-4.332	31.3	.000	-17.770	4.102	-26.143	-9.406

Table 7.2: Independent-sample t-test for the Wikipedia illustrative example

The T-test revealed a statistically significant difference between the mental workload scores computed for Group A of subjects (mean: 19.68, std.dev: 9.34) and the scores for group B (mean: 37.45, std.dev: 15.64), $t(31.209) = 4.332$, $p \leq 0.001$, $\alpha = 0.05$. Therefore, the structural changes performed over the Wikipedia interface imposed a higher mental workload on end-users. Although this outcome was expected in this scenario, a metric for evaluating the impact of the removal of a simple search box, or other navigation aids, is likely to be extremely useful for evaluating end-user experience and for maximising interface design.

[2] The assumption of a Ttest are as follows. The dependent variable (index of mental workload) is continuous. The independent variable (group A and B) consists of two categorical independent groups. Observations of the two groups are independent. The are no outliers in the distributions of the workload scores of the two groups. The dependent variable is normally distributed for both the groups. There is homogeneity of variance.

7.4.2 Supporting customisation

Personalisation is an important property of modern web-technologies that try to accommodate the differences between individuals to improve user experience during human-computer interaction. Personalisation is also related to the notion of customisation that includes any manipulation of the information provided to users like images, text or adaptation of content that is already available to them. Implicit personalisation refers to the automatic adaptation of the content by the system while explicit personalisation indicates that the interface, used by humans to interact with a system, is subjectively altered by the humans themselves by using some features provided by the system itself. Regardless of the typology of personalisation, the adaptation process produces an altered version of the interface that might impose different levels of mental workload on end-users, given the same task, compared to the original version. In turn, if these levels fall to within optimal range the personalisation process is likely to fail, negatively affecting user experience Consider tasks 4 and 5 of list 6.2 (page 124) performed over Google.com

- T_4 : Find out the difference (in years) between the year of the foundation of the Apple Computer Inc. and the year of the 14th FIFA world cup
- T_5: Find out the difference (in years) between the foundation of the Microsoft Corporation and the year of the 23rd Olympic games

These tasks represent typical searches (fact-finding) that can be performed over the Google web-site. They are slightly more complicated than standard searches because they embed two sub-searches that had to be executed in order to find out the solution to each task. For this reason they required more time for completion. Let's suppose that the designers of the classic results page of Google, as in figures B.4 (b) and B.5 (b) (appendix B.1) are willing to investigate the impact of the following changes in relation to imposed mental workload on end-users:

- each result surrounded by a box and a different font (figure B.4 (a),appendix B.1)
- a dark background with no left-side menu (figure B.5 (a),appendix B.1)

As in the previous case-study, independent sample T-tests are run over the mental workload indexes computed by the instance MWL_{def} of the defeasible framework to determine whether the original Google interface imposed statistically significant mental workload on end-users than the altered versions. The null hypotheses are the equality of workload indexes computed for the two groups of people for both the tasks.

$$T_4 - H_0 : MWL^A_{def} = MWL^B_{def} \qquad H_1 : MWL^A_{def} \neq MWL^B_{def}$$

$$T_5 - H_0 : MWL^A_{def} = MWL^B_{def} \qquad H_1 : MWL^A_{def} \neq MWL^B_{def}$$

Table 7.3 shows the group statistics, while table 7.4 shows the T-tests for both tasks T_4 and T_5: the assumptions behind each test are met and there is homogeneity of variance as revealed by the Leven's test (the sig. values are all greater than the alpha level $\alpha = 0.05$). The T-value of .932 for task T_4 has a significance of $0.357 > 0.05$ thus there is no evidence to reject the null hypothesis H_0 which is therefore accepted. Similarly the t-value of .498 for task T_5 has a significance of $0.621 > 0.05$ thus there is no evidence to reject the null hypothesis H_0.

Task	Group	N	Mean	Std. dev.	Std. Error mean
T_4	A	20	34.69	12.06	2.69
T_4	B	20	30.90	13.61	3.04
T_5	A	19	31.60	11.17	2.56
T_5	B	20	29.33	16.58	3.70

Table 7.3: Group statistics for the Google illustrative example

	Leven's Test for equality of variances		T-test for equality of means						
	F	Sig.	t	df	Sig (2-tailed)	Mean diff.	Std. err diff.	95% confidence interval of the diff	
MWL_{def} - T_4								Lower	Upper
Equal variances	0.119	.732	.932	38	.357	3.793	4.067	-4.441	12.027

	Leven's Test for equality of variances		T-test for equality of means						
	F	Sig.	t	df	Sig (2-tailed)	Mean diff.	Std. err diff.	95% confidence interval of the diff	
MWL_{def} - T_5								Lower	Upper
Equal variances	3.433	.072	.498	37	.621	2.269	4.553	-6.957	11.495

Table 7.4: Independent-sample t-test for the Google illustrative example

The t-tests failed to reveal a statistically significant difference between the mental workload scores computed for group A of subjects (T_4 - mean: 34.69, std.dev: 12.06; T_5 - mean:31.60, std.dev: 11.17) and the scores for group B (T_4 - mean: 30.90, std.dev: 13.61; T_5 - mean: 29.33, std.dev: 16.58), $t(38) = .932$, $p = .357$, $\alpha = 0.05$ for task 4 and $t(37) = .498$, $p = .621$, $\alpha = 0.05$ for task 5. Therefore, the structural changes performed over the Google results page imposed the same mental workload on end-users as the original interface. This suggests that the use of a new font and the introduction of a box that wrapped each result did not really influence user behaviour, nor did the removal of the left-side menu or the addition of a new dark background. In turn, explicit personalisation might be proposed as a feature in the classic Google interface and the left-menu might be removed because it was not really an influence. However, the latter decision should be taken analysing the mental workload of users on more difficult tasks, where the search-solution is hard to be found.

To conclude, any structural change over a web-page might influence user experience. However, it is very difficult to assume a-priori end-users' reactions and behaviour. As a consequence, the use of mental workload as a metric for tackling such an issue and investigating user experience can be extremely useful. In addition, not only can components of the interface be considered, but also manipulations of the algorithms for content delivering can be viewed as structural changes, these being influential elements of the user experience.

Chapter 8

Conclusion

This chapter summarises the thesis, highlighting its main contributions and achievements, as well as its strengths and limitations. It describes the work that remains to be done and the future directions in the research of mental workload formalisation and assessment.

8.1 Thesis summary

This thesis described a novel methodology for mental workload representation and assessment based on defeasible reasoning (DR). DR is a form of non-monotonic reasoning built upon reasons that can be defeated; a conclusion, derived from the application of previous knowledge, can be revised in the light of new evidence. This type of reasoning was formally implemented by designing a modular framework, built upon argumentation theory, that not only supports the representation of mental workload, but is also capable of workload assessments. The framework represents a proof-of-concept and has been empirically evaluated with a user study involving humans. The following sections summarise the path which has been taken in accomplishing the aim of the study.

8.1.1 Introduction

The introductory chapter outlined the motivations for this work by stressing how the concept of mental workload has progressively acquired importance due to the increasing use of computer- and web-based technologies that have led the activities of humans to become more cognitively focused. As a consequence, the assessment of mental workload may exert an important influence in supporting the development of digital interfaces and in understanding human performance within emerging complex human-computer interactive systems. Although the construct has been mainly applied, so far, in the automobile and aviation industries, and principally investigated by psychologists, ergonomists and neuroscientists, it is argued that it can be applied in the much broader domain of human-computer interaction, where its future impact is likely to be highly significant. Despite 40 years of research efforts, no clear definition of mental workload has yet emerged. Several operational approaches have been proposed, but there appears to be a lack of agreement among them about its sources, measurement methods and consequences. The literature has suggested that mental workload is

a multi-faceted construct that involves two main components: a task and a person. This interaction might be mediated by several other elements such as available cognitive resources, the abilities, skills and state of a person, the effort exerted and external factors of influence. This supported the reasonable assumption that mental workload is a complex construct, built upon a network of pieces of evidence. Furthermore, it was assumed that the interaction of these pieces of evidence is an important aspect for defining and assessing mental workload. These assumptions are the key components of a defeasible concept, that is a concept built upon a set of reasons that can be defeated in the light of new evidence. State-of-the-art theoretical models of defeasible reasoning, in the field of AI, are formally implemented using argumentation theory (AT) and argument-based computations. This is a multi-disciplinary paradigm that systematically studies how arguments can be formally built, maintained or discarded in a reasoning process and it investigates the validity of the conclusions reached. These considerations led to the formalisation of the research question which investigates the application of defeasible argumentation theory as a novel technique to support the representation of the construct of mental workload, and to improve the quality of its assessment.

8.1.2 Literature review: mental workload

The second chapter was focused on a literature review of current state-of-the-art techniques in representing, defining and measuring mental workload. This review highlighted the complexity in modelling the construct but also the importance of its application in many domains. The core tenets which were found underlined the multi-dimensional nature of mental workload: a context-aware, task-specific and user-specific hypothetical construct. This is thought not to be detectable directly but rather through the measurement and aggregation of some other factors believed to correlate strongly with it and which interact with each other. These characteristics supported the need for a framework for reasoning upon mental workload that ideally is flexible, replicable and inconsistency-aware. In other words, this framework should be open and adaptable, enabling a workload practitioner to incorporate those attributes believed to be useful in shaping mental workload and in aggregating them towards a meaningful and justifiable assessment. This aggregation should take stock of inconsistencies and contradictions that might arise from the interaction of evidence which has been noted. Flexibility should enable falsifiability: each proper use of the framework, aimed at defining and controlling mental workload, is an attempt to falsify it, thus increasing our understanding of the construct itself. It should also be replicable in different contexts, showing reliability in the assessments. The implementation of such an ideal framework was assumed to be achievable by employing defeasible reasoning (DR). However, the validity of this assumption was actually tested by reviewing the literature on DR.

8.1.3 Literature review: defeasible reasoning

The third chapter was devoted to the introduction of defeasible reasoning and non-monotonic logic - notions that form the core of this thesis. Subsequently, argumentation theory, based upon these notions, was reviewed, describing its basic building blocks and elements for knowledge representation. This included monological models aimed at internally represent arguments, and dialogical models aimed at investigating their conflicts and resolving possible contradictions arising from their interaction. In addition, various ways of formalising conflicts between arguments and different methods for evaluating their credibility were described. In turn, it

was shown how these techniques supported the computation of consistent and conflict-free sets of arguments, sets of evidence that can be used for making more rational decisions, or for better justifying claims. This review showed how defeasible reasoning and argumentation theory could be considered appropriate candidates for modelling the construct of mental workload as a defeasible computational concept.

8.1.4 Design

According to the difficulties and issues related to the representation and assessment of mental workload, (which emerged from the literature review of chapter 2) a formal computational modular framework based on defeasible reasoning and argument-based computations was designed. This framework was implemented according to the state-of-the-art techniques in argumentation theory that were presented in chapter 3. The framework proposes a multi-layer methodology for mental workload representation and assessment. This methodology starts with a technique for building formal defeasible arguments from vague natural language propositions in a monological structure, adopting the notion of 'degree of truth', borrowed from fuzzy logic. These defeasible arguments can be connected in a dialogical structure employing the notion of attack. Arguments and attack relations form an instance of the framework, a directed graph that formally represents the translated knowledge-base of a workload designer into interactive pieces of knowledge. This instance can be fully or partially elicited through the quantification of the degree of truth of the premises of arguments with quantitative inputs. The argumentation graph elicited can be evaluated by applying state-of-the-art acceptability semantics for the resolution of the potential inconsistencies that might emerge from the interaction of activated arguments. The output of an acceptability semantics is a single or multiple sub-set of activated arguments, coherent and conflict-free points of view. Eventually, the sub-set of arguments with the highest combined degree of truth, that means the most credible, can be used for mental workload assessment.

8.1.5 Implementation and instantiation

The actual implementation of the framework was described in chapter 5 summarising the algorithm for mental workload assessment and some technical details about practical deployment. Subsequently it was demonstrated how the framework can be employed in practice for translating two of the state-of-the-art subjective mental workload instruments, namely the NASA Task Load Index Hart (2006) and the Workload Profile Tsang and Velazquez (1996), into two particular instances as well as for creating two brand new instances from scratch, with and without interaction of arguments.

8.1.6 Evaluation

The evaluation of the framework included an empirical experimental study involving human participants, who were required to perform a set of web-based tasks and provided with subjective ratings, inputs for the instances built in chapter 5. The evaluation strategy included a comparison of the degree of sensitivity, diagnosticity and validity of the assessments produced by the brand new designed instances against the two original state-of-the-art subjective mental workload assessment instruments.

8.1.7 Discussion and applications

The discussion chapter was devoted to a critical interpretation of the results of the evaluation of the defeasible framework. The findings support the main aspect of the research question, that is a demonstration of how defeasible argumentation theory can be successfully adopted to support the representation of mental workload and to enhance the quality of its assessments. Specifically, it was demonstrated how the two selected state-of-the-art subjective assessment techniques (The NASA Task Load Index and the Workload Profile) could be successfully translated into two instances of the defeasible framework while still showing a strong convergent validity. In other words, the indexes of mental workload inferred by the original two instruments, and those generated by their corresponding translations (instances of the framework), showed a positive and nearly perfect statistical correlation. Additionally, one of the new designed instances of the defeasible framework (with interaction of arguments) showed a better sensitivity and a higher diagnosticity capacity than the two selected state-of-the art techniques. The former technique had a higher convergent validity with the latter technique, but a better concurrent validity with task completion time, this being a performance objective measure. The defeasible instance, with interaction of arguments, generated indexes of mental workload that better correlated with the objective time for task completion compared to the two selected state-of-the-art instruments. The findings suggest that accounting for interactions between arguments, that means relationships of pieces of knowledge, significantly enhanced the quality of mental workload assessments. This suggests that the application of defeasible reasoning in the domain of mental workload is promising and worthy of further investigation.

A summary of the limitations related to the use of the defeasible framework was presented. The main weakness of the proposed solution, an area where improvement is needed, is its acceptability. The framework was employed by just one workload designer, thus its acceptability could not be actually tested. Another limitation concerns the reliability of the results emerging from the user-study. The experiment conducted involved 11 tasks and 40 users, clearly not enough to claim a strong reliability for the designed framework. Improvement of acceptability and reliability will indeed be an objective of future research.

Finally, an example was provided to illustrate how the framework could be also applied in other settings where the amount of available evidence is limited, partial or not easily acquirable, as with subjective ratings. The chapter concluded with practical uses of the mental workload indexes and output of the experimental user-study, demonstrating, through two illustrative A/B testing examples, how to enhance and support web-design and customisation of web-interfaces.

8.2 Contributions to the body of knowledge

The proposed defeasible framework has been showed to be actually employable to improve mental workload representation and to enhance the quality of its assessment. This suggests that this novel approach, based on defeasible reasoning and formally implemented using argumentation theory, is a suitable alternative to existing techniques. A major contribution was the presentation of a methodology, developed as a formal

framework, to represent mental workload as a defeasible computational concept. This methodology is based upon defeasible reasoning, a form of non-monotonic reasoning built upon reasons that can be defeated and a conclusion, derived from the application of previous knowledge, can be retracted in the light of new evidence. In addition to this major contribution, this thesis also makes several other minor contributions. Firstly, it provided further monological models of argumentation by employing the notion of degree of truth, borrowed from fuzzy logic. Secondly, it linked this new monological structure to dialogical models of arguments, very often investigated in the field of argumentation theory individually and not conjointly. Thirdly, it showed how the quality of mental workload assessment can be enhanced by employing the proposed framework. Eventually, the thesis demonstrated how computed indexes of mental workload could be practically applied to enhance web-design. These contributions are summarised as follows:

- *Major contribution*: considering mental workload as a defeasible phenomenon and formally modelling it adopting defeasible argumentation theory.
- *Minor contributions*:
  - extension of the landscape of formal monological structure of arguments with the notion of degree of truth, borrowed from fuzzy logic;
  - linkage of such a new monological structure to state-of-the-art dialogical models of argumentation;
  - enhancement of the quality of mental workload assessment produced by instances built employing the defeasible framework;
  - application of the assessment of mental workload in the field of human-computer interaction for supporting and enhancing web-design and customisation.

The publications related to this thesis are as follows:

- (Longo and Barrett, 2010a) and (Longo and Barrett, 2010b) were aimed at understanding the set of those attributes useful for shaping mental workload and how they could be computationally modelled. These studies confirmed that reasoning on mental workload and formalising it as a computational concept are very subjective tasks and that a unique way to address these two problems does not exist. They led the author to focus on a common and more generally applicable structure that could be employed and adopted by different mental workload designers.
- (Longo and Kane, 2011) and (Longo et al., 2012b) were focused on ad-hoc applications of the construct of mental workload in the field of health informatics and web-design, leading the author to focus instead on a common structure and methodology that can be employed in different fields.
- (Longo et al., 2012a), (Longo and Hederman, 2013) and (Longo and Dondio, 2014) were devoted to reviewing defeasible reasoning and to applying argumentation theory in more than one field of healthcare. These studies led the author of this thesis to adopt the same paradigms for tackling the construct of human mental workload.
- (Longo, 2011) and (Longo, 2012) were aimed at proposing and refining the idea behind this thesis at two top-tier doctoral consortia in the field of human-computer interaction and user-modelling.

8.3 Future work

The solution proposed in this thesis can be improved in different ways. Firstly, as was already anticipated, in order to assess the reliability of the defeasible framework which has been designed, further tests have to be carried out. These include the creation of other instances of such a framework considering the knowledge-base of other designers, belonging to different domains. These instances have to be applied in different settings where evidence can be both acquired using subjective ratings of users and via implicit objective indicators of effort, such as mouse and keyboard usage and other workload influencers. Secondly, these instances have to be systematically evaluated through comparisons of their assessment capacity in terms of sensitivity, diagnosticity and validity against current state-of-the-art mental workload assessment techniques. In addition, the outcomes of these instances have to be systematically correlated with objective performance measures, such as time or error rates in order to state their superiority and potential usefulness in justifying and explaining users' behaviour and performance on a system. Yet, regarding the evaluation of the capacity of the framework to assess mental workload, a more effective sensitivity analysis can be performed, with a more focused manipulation of task loads. This will help to achieve a better understanding of the relationship between the changes in the task loads (input) and the assessed mental workload (output). Similarly, future work will include a more detailed investigation of the diagnosticity capacity of different instances of the framework to detect the pool of mental resources being taxed. Eventually, the theoretical methodology built as a modular framework, will be coupled with a graphical interface in order to assist the operator in instantiating his/her own expertise and knowledge-base and probably enhancing the acceptability of the framework itself.

On the theoretical side, the designed methodology can be also enhanced and extended in various ways. Firstly, an in-depth investigation of the usefulness of the two reluctancy thresholds for argument and attack activation will be performed. This is aimed at deciding whether they are actually important for the elicitation of a knowledge-base, with a consequent impact on the quality of the assessments, or whether they can be removed, simplifying the methodology. Secondly, there can be an extension of the accrual of arguments, in the final layer of the argumentative schema, towards the generation of an index of workload. An alternative to the currently proposed average of the claims of the arguments, in the most credible acceptable extensions, may be the adoption of classical fuzzy logic operators (and, or). Different acceptability semantics will be investigated for the resolution of the inconsistencies that might arise from the interaction of arguments. Moreover, the experimental study conducted in this thesis suggests that expressing a clear preference over all the accounted workload attributes might be not-trivial and non-intuitive. In turn, the framework might be simplified, removing the explicit function that returns the importance of workload attributes and in turn preferentiality could be implemented using meta-level preference arguments, as emerged in the literature of argumentation theory.

8.4 Final remark

As the first of its kind, the solution described in this thesis showed how the complex and multi-faceted problem of mental workload representation and assessment can be tackled from a different perspective. This novel perspective assumes mental workload to be a defeasible phenomenon that could be formally modelled using argumentation theory, a paradigm that computationally implements defeasible reasoning, a form of non-monotonic reasoning. This thesis demonstrated how the proposed solution is a suitable alternative to conventional instruments for mental workload representation and for enhancements of the quality of mental workload assessments. The solution proposed was designed for those scholars interested in engaging in the multi-disciplinary domain of mental workload and it is aimed at increasing its understanding and use, especially in emerging and modern human-computer interactive systems. The aim behind this thesis is to offer a new perspective on the formalisation of mental workload, encouraging further research on its representation, assessment and application in the more general field of human-computer interaction.

Bibliography

Addie, J. and Widyanti, A. (2011). Cultural influences on the measurement of subjective mental workload. *Ergonomics*, 54(6):509–518.

Albers, M. J. (2011). Tapping as a measure of cognitive load and website usability. In *29th ACM international conference on Design of communication*, pages 25–32.

Amgoud, L. and Cayrol, C. (2002). A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215.

Amgoud, L. and Kaci, S. (2005). An argumentation framework for merging conflicting knowledge bases: The prioritized case. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, LNCS - 3571:527–538.

Annett, J. (2002a). Subjective rating scales in ergonomics: a reply. *Ergonomics*, 45(14):1042–1046.

Annett, J. (2002b). Subjective rating scales: science or art? *Ergonomics*, 45(14):966–987.

Atkinson, K., Bench-Capon, T., and McBurney, P. (2006). Computational representation of practical argument. *Knowledge, Rationality & Action. Special section Synthese*, 152(2):157–206.

Bailey, B. P. and Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate and affective state. *Computers in Human Behaviour*, 22:685–708.

Baldauf, D., Burgard, E., and Wittman, M. (2009). Time perception as a workload measure in simulated car driving. *Applied ergonomics*, 40(5):929–935.

Baroni, P., Caminada, M., and Giacomin, M. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410.

Baroni, P. and Giacomin, M. (2009). Semantics of abstract argument systems. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 25–44. Springer US.

Baroni, P., Guida, G., and Mussi, S. (1997). Full nonmonotonicity: a new perspective in defeasible reasoning,. In *ESIT 97, European Symposium on Intelligent Techniques*, pages 58–62.

Bench-Capon, T. J. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Bench-Capon, T. J. and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641.

Bentahar, J., Moulin, B., and Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., and Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space and Environment Medicine*, 78(5):B231–244.

Bertram, D. A., Hershey, C. O., Opila, D. A., and Quirin, O. (1990). A measure of physician mental work load in internal medicine ambulatory care clinics. *Medical care*, 28(5):458–467.

Bertram, D. A., Opila, D. A., Brown, J. L., Gallagher, S. J., Schifeling, R. W., Snow, I. S., and Hershey, C. O. (1992). Measuring physician mental workload: reliability and validity assessment of a brief instrument. *Medical Care*, 30(2):95–104.

Bondarenko, A., Dung, P. M., Kowalski, R., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1–2):63–101.

Brookings, J. B., Wilson, G. F., and Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377.

Byrne, A. (2011). Measurement of mental workload in clinical medicine: a review study. *Anesthesiology and pain medicine*, 1(2):90–94.

Byrne, A., Oliver, M., Bodger, O., Barnett, W., Williams, D., Jones, H., and Murphy, A. (2010). Novel method of measuring the mental workload of anaesthetists during clinical practice. *British journal of anaesthesia*, 105(6):767–771.

Cain, B. (2007). A review ot the mental workload literature. Technical report, Defence Research and Development Canada toronto, Human System Integration Section, Human System Integration Section, 1133 Sheppard Avenue West, Toronto, Ontario M3M 3B9, Canada.

Caminada, M. W. A., Carnielli, W. A., and Dunne, P. E. (2012). Semi-stable semantics. *Journal of Logic and Computation*, 22(5):1207–1254.

Carswell, M. C., Clarke, D., and Seales, B. W. (2005). Assessing mental workload during laparoscopic surgery. *Surgical innovation*, 12(1):80–90.

Castor, M. C. (2003). Garteur handbook of mental workload measurement. Flight Mechanism Action Group FM-AG13, GARTEUR, Group for Aeronautical Research and Technology in Europe.

Chaouachi, M., Jraidi, I., and Frasson, C. (2011). Modeling mental workload using eeg features for intelligent systems. In *User modeling, adaptation and personalisation*, volume LNCS 6787, pages 50–61.

Cinaz, B., Arnrich, B., La Marca, R., and Tröster, G. (2011). Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing Journal*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Inc., 2nd edition.

Cook, J. R. and Salvendy, G. (1997). Job enrichment and mental workload in computer-based work: Implications for adaptive job design. *International Journal of Industrial Ergonomics*, 24(1):13–23.

Cooper, G. E. and Harper, R. P. (1969). The use of pilot ratings in the evaluation of aircraft handling qualities. Technical Report. AD689722, report 567, AGARD. Advisory Group for Aerospace Research & Development, 7 rue Ancelle, Neuilly Sur Seine, France.

Damos, D. L. (1988). Individual differences in subjective estimates of workload. In Hancock, P. A. and Meshkati, N., editors, *Human mental Workload*, volume 52 of *Advances in Psychology*, pages 231–237. North-Holland.

Davis, D., Oliver, M., and Byrne, A. (2009). A novel method of measuring the mental workload of anaesthetists during simulated practice. *British journal of anaesthesia*, 103(5):665–669.

De Greef, T., Lafeber, H., Van Oostendorp, H., and Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. In *Foundations of Augmented Cognition*, pages 219–228.

De Waard, D. (1996). *The measurement of drivers' mental workload*. The Traffic Research Centre VSC, University of Groningen, P.O. Box 69, 9750 AB HAREN, The Netherlands.

Dey, A. and Mann, D. D. (2010). Sensitivity and diagnosticity of nasa-tlx and simplified swat to assess the mental workload associated with operating an agricultural sprayer. *Ergonomics*, 53(7):848–857.

Diane Kuhl, M. (2000). Mental workload and arl workload modeling tools. Technical Report 35, U.S. Research Laboratory, Human Research & Engineering Directorate, Aberdeen Proving Ground MD.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–358.

Dung, P. M., Mancarellab, P., and Toni, F. (2007). Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15):642–674.

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., and Wooldridge, M. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486.

Eggemeier, T. F. and O'Donnell, R. D. (1998). A conceptual framework for development of a workload assessment methodology. Technical report, Defense Technical Information Center OAI-PMH Repository (United States).

Eggemeier, T. F. and Wilson, G. F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In Damos, D., editor, *Multiple-task performance*, pages 217–278. Taylor & Francis.

Eggemeier, T. F., Wilson, G. F., Kramer, A. F., and Damos, D. L. (1991). Workload assessment in multi-task environments. In Damos, D., editor, *Multiple-task performance*, pages 208–216. Taylor & Francis.

Fairclough, S. H. (1993). Psychophysiological measures of workload and stress. In *Driving Future Vehicles*, pages 377–390. Taylor & Francis, in a.m. parkes & s. franzèn edition.

Felton, E. A., Williams, J. C., Vanderheiden, G. C., and Radwin, R. G. (2012). Mental workload during brain-computer interface training. *Ergonomics*, 55(5):526–537.

Fox, J., Kraus, P., and Elvan-Goransson, M. (1993). Argumentation as a general framework for uncertain reasoning. In *Proceedings of the 9th conference on uncertainty in AI*, pages 428–434. Morgan-Kaufmann.

France, D. J., Levin, S., Hemphill, R., Chen, K., Rickard, D., Makowski, R., Jones, I., and Aronsky, D. (2005). Emergency physicians' behaviors and workload in the presence of an electronic whiteboard. *Medical Informatics*, 74(10):827–837.

Gaba, D. M. and Lee, T. (1990). Measuring the workload of the anesthesiologist. *Anesthesia & Analgesia*, 71(4):354–361.

Gopher, D. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*, 26(5):519–532.

Gopher, D. and Donchin, E. (1986). Workload - an examination of the concept. In Boff, K. R., Kaufman, L., and Thomas, J. P., editors, *Handbook of perception and human performance*, volume 2, pages 41/1–41/49. John Wiley & Sons.

Grasso, F. (2002). Towards a framework for rhetorical argumentation. In *Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, pages 53–60.

Gwizdka, J. (2009). Assessing cognitive load on web search tasks. *The ergonomic open journal*, 2(1):114–123.

Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the american society & information science & technology*, 61(11):2167–2187.

Hancock, P. A. (1988). The effect of gender and time of day upon the subjective estimate of mental workload during the performance of a simple task. In Hancock, P. A. and Meshkati, N., editors, *Human mental Workload*, volume 52 of *Advances in Psychology*, pages 239–250. North-Holland.

Hancock, P. A. (1989). The effect of performance failure and task demand on the perception of mental workload. *Applied ergonomics*, 20(3):197–205.

Hancock, P. A. and Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transaction on systems, man and cybernetics*, 18(4):647–658.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland.

Her, C.-C. and Hwang, S.-L. (1980). Application of queueing theory to quantify information workload in supervisory control systems. *Industrial ergonomics*, 4(1):51–60.

Huey, B. M. and Wickens, C. D. (1993). *Workload transition: implication for individual and team performance*. National Academy Press, Washington, DC.

Hwang, S.-L., Yau, Y.-J., Lin, Y.-T., Chen, J. H., Huang, T.-H., Yenn, T.-C., and Hsu, C.-C. (2007). A mental workload predicator model for the design of pre alarm systems. In *Engineering Psychology and Cognitive Ergonomics*, volume LNAI 4562, pages 316–323.

Janssen, J., De Cock, M., and Vermeir, D. (2008). Fuzzy argumentation frameworks. In *Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 513–520.

Jou, T.-Y., Yenn, T.-C., Lin, C. J., Yang, C.-W., Lin, C., and Chih-Cheng, Y. (2009a). Evaluation of operators' mental workload of human-system interface automation in the advanced nuclear power plants. *Nuclear engineering and design*, 239(11):2537–2542.

Jou, T.-Y., Yenn, T.-C., Lin, C. J., Yang, C.-W., and Lin, S.-F. (2009b). Evaluation of mental workload in automation design for a main control room task. In *Networking, Sensin and Control*, pages 313–317.

Kaci, S. and Labreuche, C. (2010). Argumentation framework with fuzzy preference relations. In *13th international conference on Information processing and management of uncertainty*, pages 554–563.

Kahneman, D. (1973). *Attention and effort*. Prentice Hall, New Jersey, U.S.A.

Kataoka, J., Sasaki, M., and Kanda, K. (2011). Effects of mental workload on nurses' visual behaviors during infusion pump operation. *Japan journal of nursing science*, 8(1):47–56.

Kellar, M., Watters, C., and Shepherd, M. (2006). A Goal-based Classification of Web Information Tasks. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22.

Kokini, C. M., Lee, S., Koubek, R. J., and Moon, S. K. (2012). Considering context: The role of mental workload and operator control in users' perceptions of usability. *International Journal of Human-Computer Interaction*, 28(9):543–559.

Kramer, A. F. (1991). Physiological metrics of mental workload: a review of recent progress. In Damos, D., editor, *Multiple-task performance*, pages 279–238. Taylor & Francis.

Kramer, A. F., Sirevaag, E. J., and Braune, R. (1987). A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors*, 29:145–160.

Krause, P., Ambler, S., Elvang-Gøransson, M., and Fox, J. (1995). A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11(1):113–131.

Leedal, J. and Smith, A. (2005). Methodological approaches to anaesthetists' workload in the operating theatre. *British journal of anaesthesia*, 94(6):702–709.

Li, H., Oren, N., and Norman, T. J. (2011). Probabilistic argumentation frameworks. In *First international conference on Theory and Applications of Formal Argumentation*, pages 1–16.

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *INTERACT (4)*, pages 402–405.

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *UMAP*, pages 369–373.

Longo, L. and Barrett, S. (2010a). Cognitive effort for multi-agent systems. In *Brain Informatics, International Conference, BI 2010, Toronto, ON, Canada, August 28-30, 2010. Proceedings*, pages 55–66.

Longo, L. and Barrett, S. (2010b). A computational analysis of cognitive effort. In *Intelligent Information and Database Systems, Second International Conference, ACIIDS, Hue City, Vietnam, March 24-26, 2010. Proceedings, Part II*, pages 65–74.

Longo, L. and Dondio, P. (2014). Defeasible reasoning and argument-based systems in medical fields: An informal overview. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, USA, May 27-29, 2014*, pages 376–381.

Longo, L. and Hederman, L. (2013). Argumentation theory for decision support in health-care: A comparison with machine learning. In *Brain and Health Informatics - International Conference, BHI 2013, Maebashi, Japan, October 29-31, 2013. Proceedings*, pages 168–180.

Longo, L. and Kane, B. (2011). A novel methodology for evaluating user interfaces in health care. In *Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems, 27-30 June, 2011, Bristol, United Kingdom*, pages 1–6.

Longo, L., Kane, B., and Hederman, L. (2012a). Argumentation theory in health care. In *Proceedings of CBMS 2012, The 25th IEEE International Symposium on Computer-Based Medical Systems, June 20-22, 2012, Rome, Italy*, pages 1–6.

Longo, L., Rusconi, F., Noce, L., and Barrett, S. (2012b). The importance of human mental workload in web design. In *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, 18 - 21 April, 2012*, pages 403–409.

Luximon, A. and Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, 44(3):229–243.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., and Linton, P. M. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies. Technical Report ADA212879, U.S. Army Research Institute for the Behavioural and Social Sciences, Fort Bliss, Texax. Interim report.

Macdonald, W. (1999). Train controller interface design: factors influencing mental workload. pages 31–36.

Martínez, D. C., García, A. J., and Simari, G. R. (2008). An abstract argumentation framework with varied-strength attacks. In *Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, pages 135–143.

Matt, P.-A., Morgem, M., and Toni, F. (2010). Combining statistics and arguments to compute trust. In *9th International Conference on Autonomous Agents and Multiagent Systems*, volume 1.

Matt, P.-A. and Toni, F. (2008). A game-theoretic measure of argument strength for abstract argumentation. In Hölldobler, S., Lutz, C., and Wansing, H., editors, *Logics in Artificial Intelligence*, volume 5293 of *Lecture Notes in Computer Science*, pages 285–297. Springer Berlin Heidelberg.

Meshkati, N. and Loewenthal, A. (1988). The effects of individual differences in information processing behavior on experiencing mental workload and perceived task difficulty: A preliminary experimental investigation. In Hancock, P. A. and Meshkati, N., editors, *Human mental Workload*, volume 52 of *Advances in Psychology*, pages 269–288. North-Holland.

Modgil, S. (2007). An abstract theory of argumentation that accommodates defeasible reasoning about preferences. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume LNCS 4724, pages 648–659.

Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934.

Muckler, F. A. and Seven, S. A. (1992). Selecting performance measures: 'objective' versus 'subjective' measurement. *Human Factors - Special issue: measurement in human factors*, 34(4):441–455.

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics*, 26(4):259–263.

Neville, M., Eisend, P., Monet, L., and Turksen, I. B. (1988). Fuzzy analysis of skill and rule-based mental workload. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 289–304. North-Holland.

North, R. A. and Riley, V. A. (1989). W/index: A predictive model of operator workload. In McMillan, G., Beevis, D., Salas, E., Strub, M.H.and Sutton, R., and Van Breda, L., editors, *Application of human performance models to system design*, volume 2 of *Defence Research Series*, pages 81–89. Plenum, New York.

Noyes, J. M. and Bruneau, D. P. J. (2007). A self-analysis of the nasa-tlx workload measure. *Ergonomics*, 50(4):514–519.

O' Donnel, R. D. and Eggemeier, T. F. (1986). Workload assessment methodology. In *Handbook of perception and human performance*, volume 2, pages 42:1–42:49. New York, Wiley-Interscience.

Parsons, S., McBurney, P., and Sklar, E. (2010). Reasoning about trust using argumentation: a position paper. In *7th international conference on Argumentation in Multi-Agent Systems*, pages 159–170.

Pasquier, P., Rahwanm, I., Dignum, F., and Sonenberg, L. (2006). Argumentation and persuasion in the cognitive coherence theory. In *Proceedings of the 1st international conference on computational models of argument (COMMA)*, pages 223–234. IOS Press.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Pickup, L., Wilson, J. R., Sharpies, S., Norris, B., Clarke, T., and Young, M. S. (2005). Fundamental examination of mental workload in the rail industry. *Theoretical issues in ergonomics science*, 6(6):463–482.

Pollock, J. L. (1974). *Knowledge and justification*. Princeton University press.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, 11(4):481–518.

Pollock, J. L. (1994). Justification and defeat. *Artificial Intelligence*, 67(2):377–407.

Popper, K. (1967). *The logic of scientific discovery*. Hutchinson, London.

Popper, K. (1969). *Conjectures and refutations*. Routledge and Kegan, London.

Prade, H. (2007). A qualitative bipolar argumentative view of trust. *Scalable Uncertainty Management*, LNCS - 4772:268–276.

Prakken, H. (2005). Ai & law, logic and argument schemes. *Argumentation (Special issue on The Toulmin model today)*, 19:303–320.

Prakken, H. (2011). An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124.

Prakken, H. and Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75.

Prakken, H. and Vreeswijk, G. (2002). Logics for defeasible argumentation. In abby DM, G. F., editor, *Handbook of philosophical logic*, volume 4, pages 219–318. Kluwer, Dordrecht, 2n edition.

Pretorius, A. and Cilliers, P. (2012). Development of a mental workload index: a systems approach. *Ergonomics*, 50(9):1503–1515.

Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliott, L. R., and Coovert, M. D. (2010). Managing workload in human-robot interaction: a review of empirical studies. *Computers in human behavior*, 26(5):840–856.

Rahwan, I. and McBurney, P. (2007). Argumentation technology (guest editors). *IEEE Intelligent Systems*, 22(6):21–23.

Reed, C. and Walton, D. (2003). Argumentation schemes in argument-as-process and argument-as-product. In *Proceedings of the conference celebrating informal Logic*, volume 25.

Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, chapter 8, pages 185–218. North-Holland.

Reimer, B. and Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, 54(10):932–942.

Roscoe, A. H. and Ellis, G. A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. Technical report TR 90019, Royal Aerospace Establishment, Farnborough (UK).

Rouse, W. B., Edwards, S. L., and Hammer, J. M. (1993). Modeling the dynamics of mental workload and human performance in complex systems. *Systems, Man and Cybernetics, IEEE Transactions on*, 23(6):1662–1671.

Rubio, S., Diaz, E., Martin, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86.

Schmutz, P., Heinz, S., Métrailler, Y., and Opwis, K. (2009). Cognitive load in ecommerce applications-measurement and effects on user satisfaction. *Advances in Human-Computer Interaction*, 1(1):9.

Schultheis, H. and Jameson, A. (2004). Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *Adaptive hypermedia and adaptive web-based systems*, volume LNCS 3137, pages 225–234.

Shingledecker, C. A. (1983). Behavioral and subjective workload metrics for operational environments. Technical report ADP002983, Air Force Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio 45433.

Stefanidis, D., Haluck, R., Pham, T., Dunne, B. J., Reinke, T., Markley, S., Korndorffer, J. R., Arellano, P., Jones, D. B., and Scott, D. J. (2006). Construct and face validity and task workload for laparoscopic camera navigation: virtual reality versus videotrainer systems at the sages learning center. *Surgical Endoscopy*, 21(7):1158–1164.

Steichen, B., O'Connor, A., and Wade, V. (2011). Personalisation in the wild: providing personalisation across semantic, social and open-web resources. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 73–82, New York, NY, USA. ACM.

Svensson, E., Angelborg-Thanderz, M., Sjöberg, L., and Olsson, S. (1997). Information complexity mental workload and performance in combat aircraft. *Ergonomics*, 40(3):362–380.

Thomas E., N. (1991). Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. *Human factors*, 33:17–33.

Toni, F. (2008). Assumption-based argumentation for closed and consistent defeasible reasoning. In *Proceedings of the 2007 conference on New frontiers in artificial intelligence*, JSAI'07, pages 390–402, Berlin, Heidelberg. Springer-Verlag.

Toni, F. (2010). Argumentative agents. In *Proceedings of the Multiconference on Computer Science and Information Technology*, pages 223–229.

Toulmin, S. (1958). *The use of argument*. Cambridge University Press.

Tracy, J. P. and Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual Conference for Technical Communication*, pages 256–260.

Tremoulet, P. D., Craven, P. L., Regli, S. H., Wilcox, S., Barton, J., Stibler, K., Gifford, A., and Clark, M. (2009). Workload-based assessment of a user interface design. In *Proceedings of the 2nd International Conference on Digital Human Modeling*, volume LNCS 5620, pages 333–342.

Tsang, P. S. (2006). Mental workload. In *International Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 166. Taylor & Francis.

Tsang, P. S. and Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381.

Tsang, P. S. and Vidulich, M. A. (2006). Mental workload and situation awareness. In *Handbook of Human Factors and Ergonomics*, pages 243–268. John Wiley & Sons, Inc.

Veltman, J. and Gaillard, A. (1993). Indices of mental workload in complex task environment. *Neuropsychobiology*, 28(1-2):72–75.

Verwey, W. B. and Veltman, H. A. (1996). Detecting short periods of elevated workload. a comparison of nine workload assessment techniques. *Journal of Experimental Psychology: applied*, 2(3):270–285.

Vidulich, M. A. and Tsang, P. S. (1986). Techniques of subjective workload assessment: a comparison of swat and the nasa-bipolar methods. *Human Factors Society*, 29(11):1385–1398.

Vidulich, M. A. and Ward Frederic G., S. J. (1991). Using the subjective workload dominance (sword) technique for projective workload assessment. *Human Factors Society*, 33(6):677–691.

Vitorio, D. M., Masculo, F. S., and Melo, M. O. (2012). Analysis of mental workload of electrical power plant operators of control and operation centers. *Work*, 41(1):2831–2839.

Vreeswijk, G. (1993). Defeasible dialectics: A controversy-oriented approach towards defeasible argumentation. *Journal of Logic and Computation*, 3:3–27.

Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning (Studies in Argumentation Theory)*. Lawrence Erlbaum Associates, Inc.

Wickens, C. D. (1984). Processing resources in attention. In Parasuraman, R. and Davies, D., editors, *Varieties of Attention*, pages 63–102. Academic Press, London.

Wickens, C. D. (1991). Processing resources and attention. In Damos, D., editor, *Multiple-task performance*, pages 3–34. Taylor & Francis.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(2):449–454.

Wickens, C. D. and Hollands, J. G. (1999). *Engineering Psychology and Human Performance*. Prentice Hall, 3rd edition.

Wiebe, E. N., Roberts, E., and Behrend, T. S. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, 26(3):474–481.

Wierwille, W. W. (1988). Important remaining issues in mental workload estimation. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, chapter 16, pages 315–328. North-Holland.

Wierwille, W. W. and Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceeding of the 27th Annual Meeting of the Human Factors Society*, pages 129–133. Human Factors Society.

Wierwille, W. W. and Eggemeier, T. F. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2):263–281.

Wierwille, W. W., Rahimi, M., and Casali, J. G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human factors*, 27(5):489–502.

Wilson, G. F. and Eggemeier, T. F. (1991). Psychophysiological assessment of workload in multi-task environments. In Damos, D., editor, *Multiple-task performance*, pages 329–360. Taylor & Francis.

Wilson, G. F. and Eggemeier, T. F. (2006). Mental workload measurement. In *International Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 167. Taylor & Francis.

Wilson, G. F. and Schlegel, R. E. (2004). Operator functional state assessment. Technical Report RTO-TR-HFM-104, NATO Research and Technology Organization., Neuilly sur Seine, France.

Wu, Y., Caminada, M., and Podlaszewski, M. (2010). A labelling based justification status of arguments. *13th International Workshop on Non-Monotonic Reasoning, Studies in Logic*, 3(4):12–29.

Xie, B. and Salvendy, G. (2000a). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics*, 4(3):213–242.

Xie, B. and Salvendy, G. (2000b). Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. *Work and Stress*, 14(1):74–99.

Yerkes, R. M. and Dodson, J. D. (1908). The relation of the strength of stimulus to rapidity of habit-formation. *journal of Comparative Neurology and Psychology*, 18:459–482.

Young, G., Zavelina, L., and Hooper, V. (2008). Assessment of workload using nasa task load index in perianesthesia nursing. *Perianesthesia nursing*, 23(2):102–110.

Young, L. R. and Sheena, D. (1975). Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7(5):397–429.

Young, M. S. and Stanton, N. A. (2002a). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, 3(2):178–194.

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

Young, M. S. and Stanton, N. A. (2004). Mental workload. In *Handbook of Human Factors and Ergonomics Methods*, pages 39–1–39–9. CRC Press.

Young, M. S. and Stanton, N. A. (2006). Mental workload: theory, measurement, and application. In Karwowski, W., editor, *International encyclopedia of ergonomics and human factors*, volume 1, pages 818–821. Taylor & Francis, 2nd edition.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zadeh, L. A. (1966). *Fuzzy Sets, Fuzzy Logic, Fuzzy Systems*. World Scientific Press.

Zhang, Y. and Luximon, A. (2005). Subjective mental workload measures. In *Ergonomia*, volume 27, pages 199–206.

Zhang, Y., Owechko, Y., and Zhang, J. (2004). Driver cognitive workload estimation: a data-driven perspective. In *Intelligent Transportation Systems*, pages 642–647.

Zijlstra, F. R. H. (1993). Efficiency in work behaviour. Doctoral thesis, Delft University, The Netherlands.

Appendix A

A.1 The Nasa Task Load Index

Dimension	Question
Mental demand	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical demand	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Performance	How successful do you think you were in accomplishing the goals, of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Frustration	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table A.1: Nasa Task Load Index (NASA-TLX) sub-scales

A.1.1 The Nasa Task Load Index pair-wise comparison

For each pair of the 6 dimensions, select the dimension that represents the more important contributor to workload for the task.

Examples:

• Effort or Performance • Mental demand or Temporal Demand • Effort or Psychological stress • ....

The combination of the 6 questions generates 15 comparisons.

A.2 Subjective Workload Assessment Technique

Dimension	Possibilities	Value
Time load	Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.	1
	Occasionally have spare time. Interruptions or overlap among activities occur infrequently.	2
	Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.	3
Mental Effort Load	Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention	1
	Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required	2
	Extensive mental effort and concentration are necessary. Very complex activity required total attention	3
Psychological stress load	Little Confusion, risk, frustration, or anxiety exists and can be easily accommodated	1
	Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance	2
	High to very intense stress due to confusion, frustration, or anxiety. High extreme determination and self-control required	3

Table A.2: Subjective Workload Assessment Technique (SWAT) dimensions

A.2.1 Hypothetical weighting schemes for the SWAT procedure

Rank order	TES	TSE	ETS	EST	STE	SET
1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1	1 1 1
2	1 1 2	1 2 1	1 1 2	2 1 1	1 2 1	2 1 1
3	1 1 3	1 3 1	1 1 3	3 1 1	1 3 1	3 1 1
4	1 2 1	1 1 2	2 1 1	1 1 2	2 1 1	1 2 1
5	1 2 2	1 2 2	2 1 2	2 1 2	2 2 1	2 2 1
6	1 2 3	1 3 2	2 1 3	3 1 2	2 3 1	3 2 1
7	1 3 1	1 1 3	3 1 1	1 1 3	3 1 1	1 3 1
8	1 3 2	1 2 3	3 1 2	2 1 3	3 2 1	2 3 1
9	1 3 3	1 3 3	3 1 3	3 1 3	3 3 1	3 3 1
10	2 1 1	2 1 1	1 2 1	1 2 1	1 1 2	1 1 2
11	2 1 2	2 2 1	1 2 2	2 2 1	1 2 2	2 1 2
12	2 1 3	2 3 1	1 2 3	3 2 1	1 3 2	3 1 2
13	2 2 1	2 1 2	2 2 1	1 2 2	2 1 2	1 2 2
14	2 2 2	2 2 2	2 2 2	2 2 2	2 2 2	2 2 2
15	2 2 3	2 3 2	2 2 3	3 2 2	2 3 2	3 2 2
16	2 3 1	2 1 3	3 2 1	1 2 3	3 1 2	1 3 2
17	2 3 2	2 2 3	3 2 2	2 2 3	3 2 2	2 3 2
18	2 3 3	2 3 3	3 2 3	3 2 3	3 3 2	3 3 2
19	3 1 1	3 1 1	1 3 1	1 3 1	1 1 3	1 1 3
20	3 1 2	3 2 1	1 3 2	2 3 1	1 2 3	2 1 3
21	3 1 3	3 3 1	1 3 3	3 3 1	1 3 3	3 1 3
22	3 2 1	3 1 2	2 3 1	1 3 2	2 1 3	1 2 3
23	3 2 2	3 2 2	2 3 2	2 3 2	2 2 3	2 2 3
24	3 2 3	3 3 2	2 3 3	3 3 2	2 3 3	3 2 3
25	3 3 1	3 1 3	3 3 1	1 3 3	3 1 3	1 3 3
26	3 3 2	3 2 3	3 3 2	2 3 3	3 2 3	2 3 3
27	3 3 3	3 3 3	3 3 3	3 3 3	3 3 3	3 3 3

Table A.3: Six hypothetical weighting schemes of the original SWAT procedure

A.3 Workload Profile

Dimension	Question	MRT Area
Perceptual/central processing	How much attention was required for activities like remembering, problem-solving decision-making, perceiving (detecting, recognising and identifying objects)?	Processing stages
Response processing	How much attention was required for selecting the proper response channel (manual - keyboard/mouse, or speech - voice) and its execution?	Processing stages
Spatial processing	How much attention was required for spatial processing (spatially pay attention around you)?	Processing codes
Verbal processing	How much attention was required for verbal material (eg. reading, processing linguistic material, listening to verbal conversations)?	Processing codes
Visual processing	How much attention was required for executing the task based on the information visually received (eyes)?	Input modality
Auditory processing	How much attention was required for executing the task based on the information auditorily received (ears)?	Input modality
Manual Responses	How much attention was required for manually respond to the task (eg. keyboard/mouse usage)?	Output modality
Speech responses	How much attention was required for producing the speech response (eg. engaging in a conversation, talk, answering questions)?	Output modality

Table A.4: Workload Profile (WP) questionnaire

A.4 Cooper-Harper rating scale

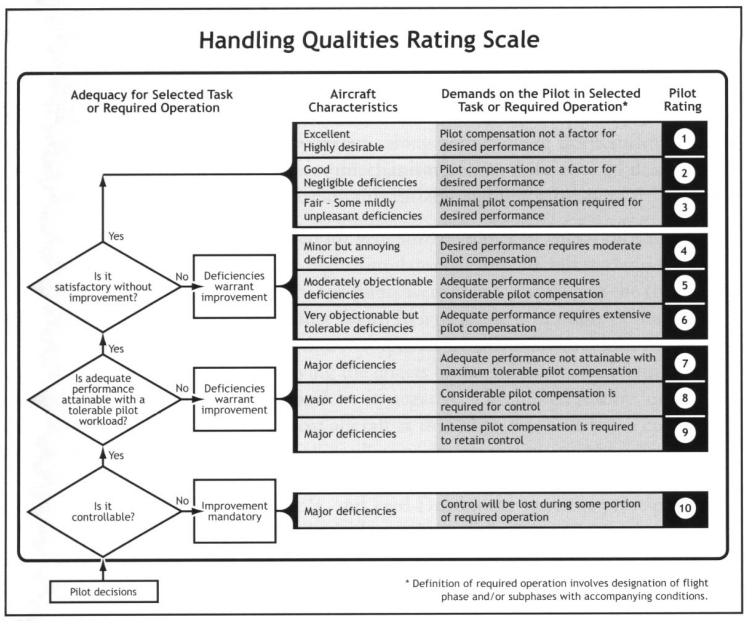


Fig. A.1: Cooper-Harper rating scale

A.5 Bedford Rating Scale

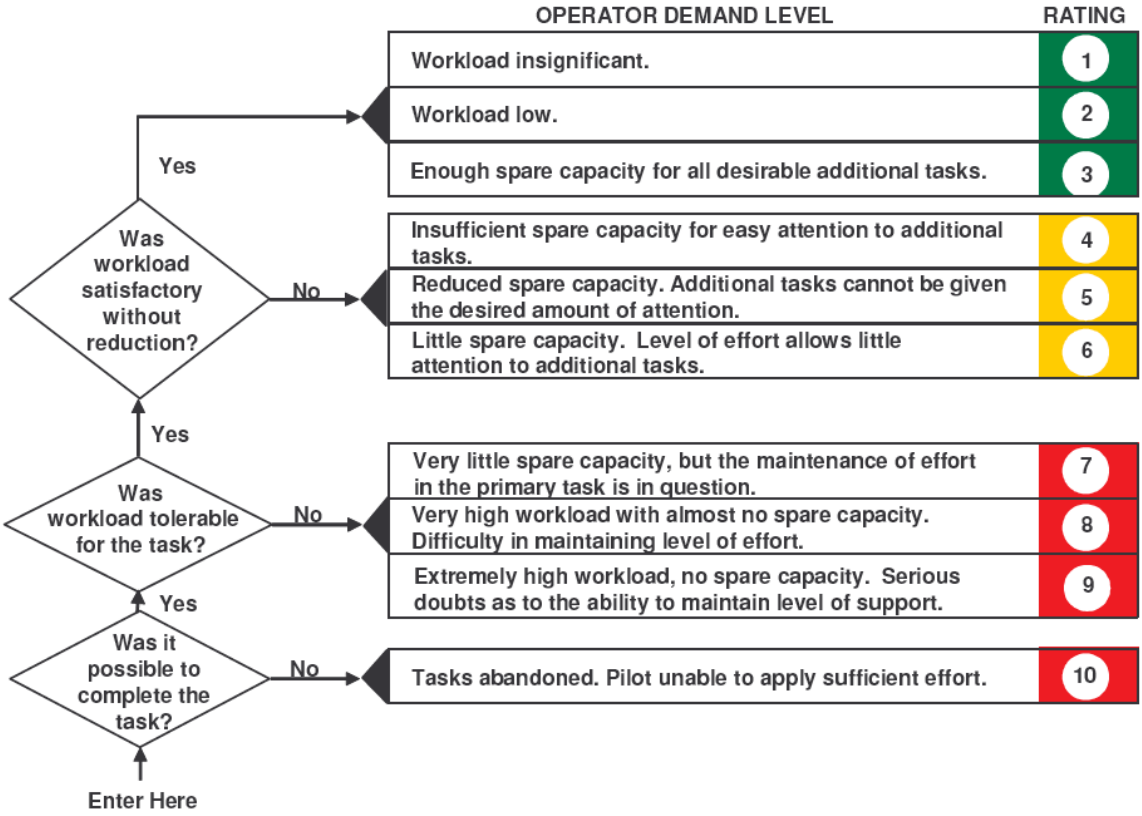


Fig. A.2: Bedford rating scale

A.6 Questionnaire used for experimental studies

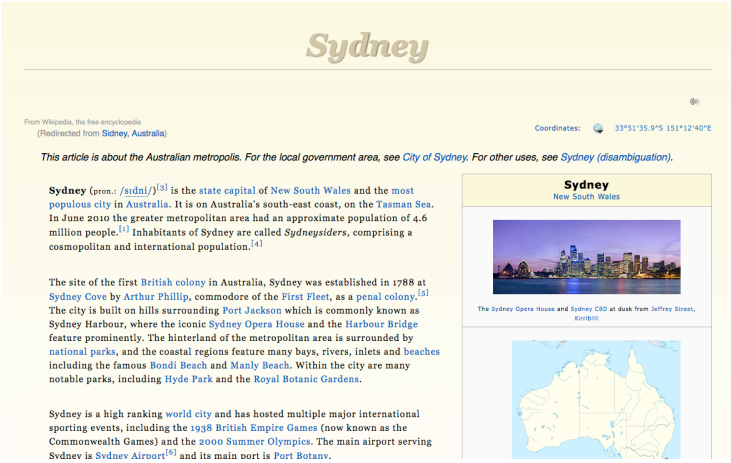
Dimension	Question
Mental demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy (low mental demand) or complex (high mental demand)?
Temporal demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely (low temporal demand) or rapid and frantic (high temporal demand)?
Effort	How much conscious mental effort or concentration was required? Was the task almost automatic (low effort) or it required total attention (high effort)?
Performance	How successful do you think you were in accomplishing the goal of the task? How satisfied were you with your performance in accomplishing the goal?
Frustration	How secure, gratified, content, relaxed and complacent (low psychological stress) versus insecure, discouraged, irritated, stressed and annoyed (high psychological stress) did you feel during the task?
Solving and deciding	How much attention was required for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognizing and identifying objects)?
Selection of response	How much attention was required for selecting the proper response channel and its execution?(manual - keyboard/mouse, or speech - voice)
Task and space	How much attention was required for spatial processing (spatially pay attention around you)?
Verbal material	How much attention was required for verbal material (eg. reading or processing linguistic material or listening to verbal conversations)?
Visual resources	How much attention was required for executing the task based on the information visually received (through eyes)?
Auditory resources	How much attention was required for executing the task based on the information auditorily received (ears)?
Manual Response	How much attention was required for manually respond to the task (eg. keyboard/mouse usage)?
Speech response	How much attention was required for producing the speech response(eg. engaging in a conversation or talk or answering questions)?
Context bias	How often interruptions on the task occurred? Were distractions (mobile, questions, noise, etc.) not important (low context bias) or did they influence your task (high context bias)?
Past knowledge	How much experience do you have in performing the task or similar tasks on the same website?
Skill	Did your skills have no influence (low) or did they help to execute the task (high)?
Motivation	Were you motivated to complete the task?
Parallelism	Did you perform just this task (low parallelism) or were you doing other parallel tasks (high parallelism) (eg. multiple tabs/windows/programs)?
Arousal	Were you aroused during the task? Were you sleepy, tired (low arousal) or fully awake and activated (high arousal)?

Table A.5: Experimental study questionnaire

Appendix B

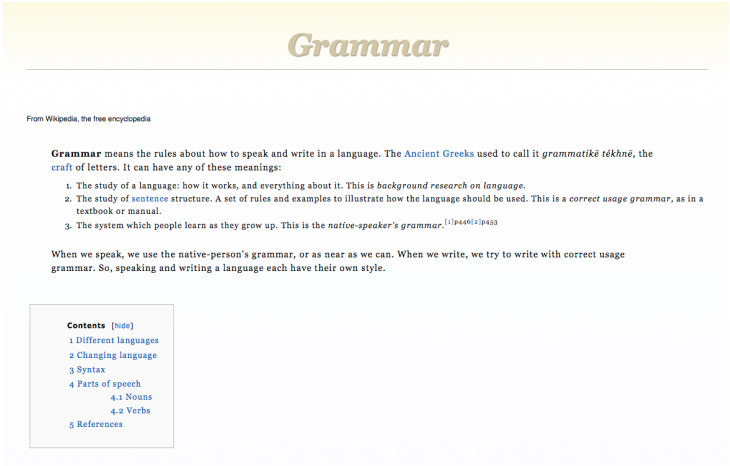
B.1 Screenshots of web-interfaces used in experimental studies

(a) Task 1 a

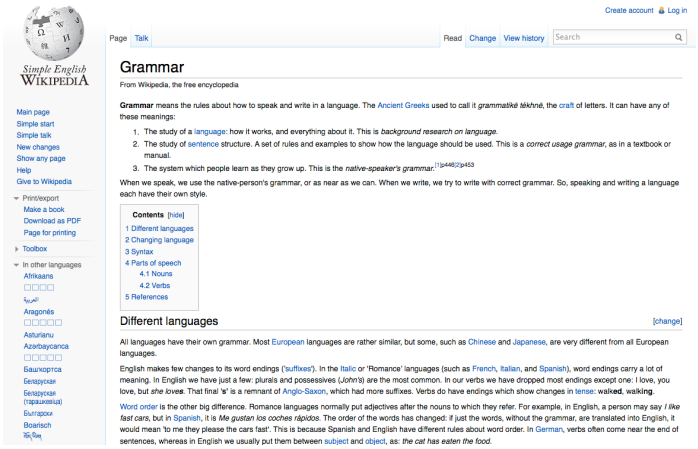


(b) Task 1 b

Fig. B.1: Web-interfaces used for task 1 of experimental study

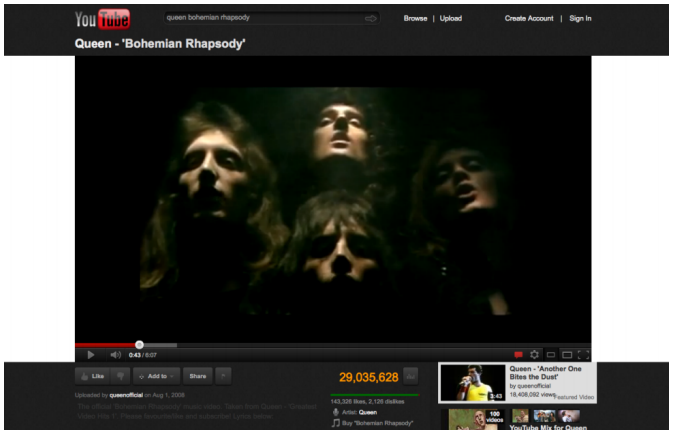


(a) Task 2 a

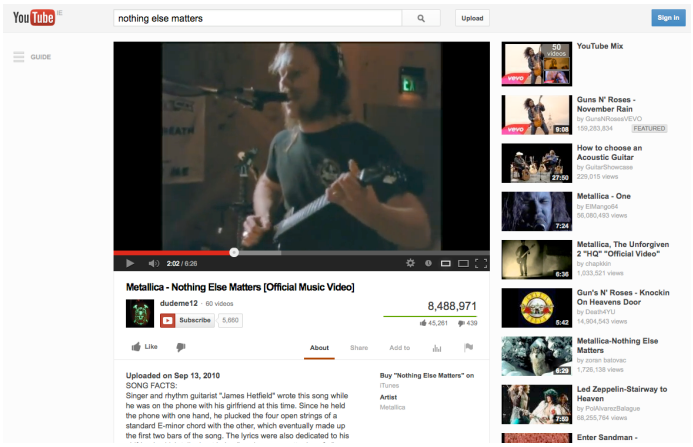


(b) Task 2 b

Fig. B.2: Web-interfaces used for task 2 of experimental study

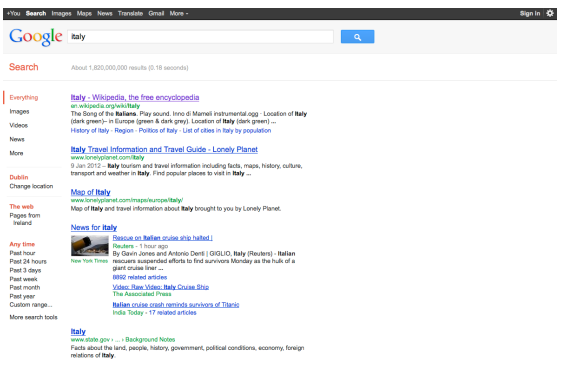


(a) Task 3 a

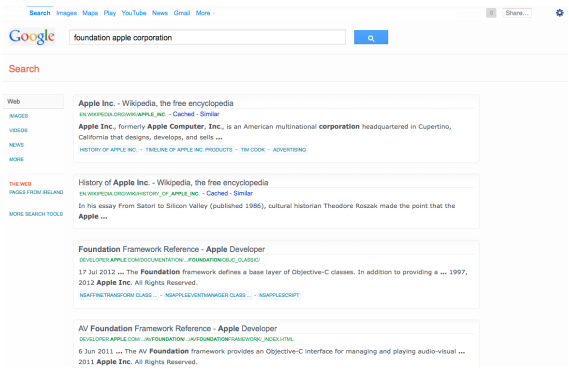


(b) Task 3 b

Fig. B.3: Web-interfaces used for task 3 of experimental study



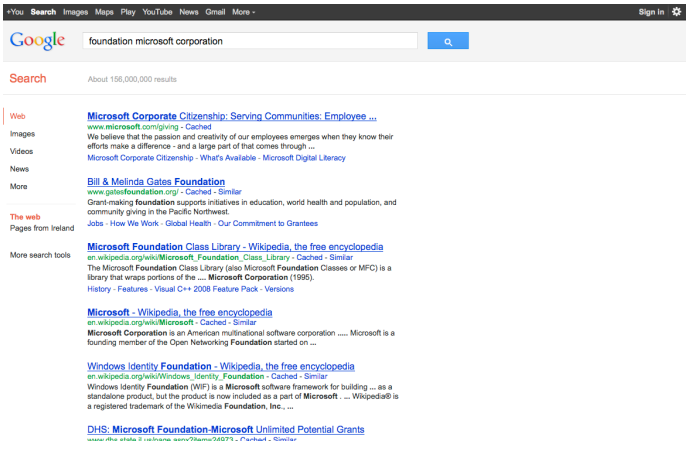
(a) Task 4 a



(b) Task 4 b

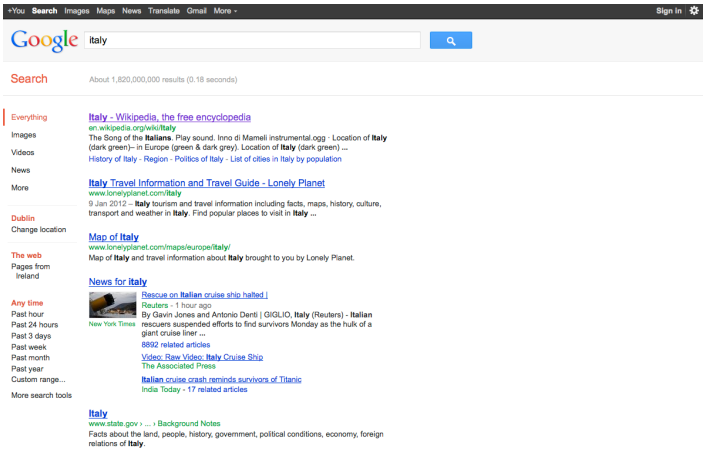
Fig. B.4: Web-interfaces used for task 4 of experimental study

(a) Task 5 a

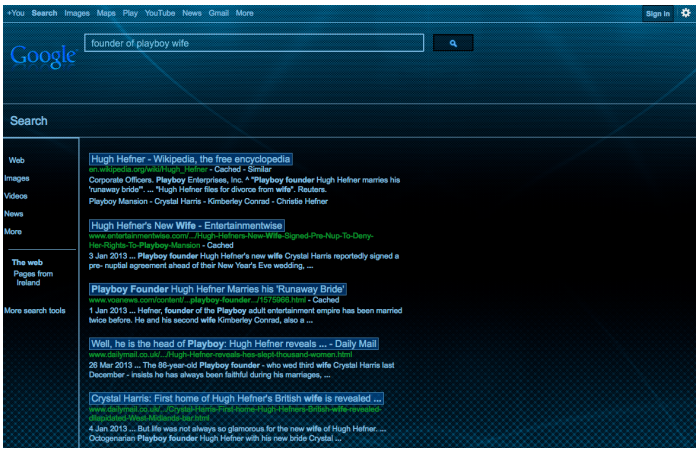


(b) Task 5 b

Fig. B.5: Web-interfaces used for task 5 of experimental study

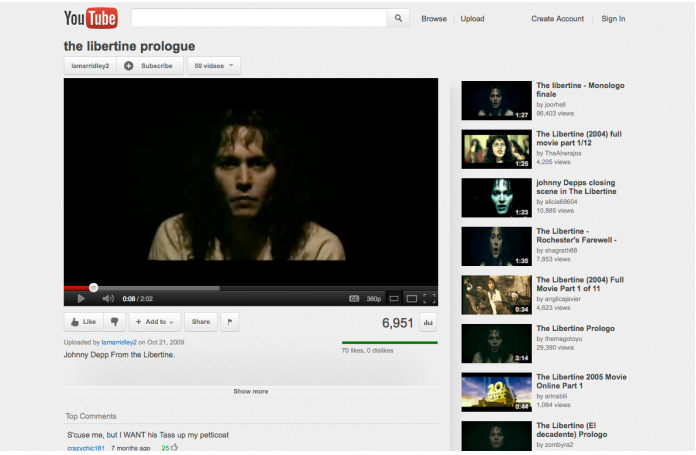


(a) Task 6 a

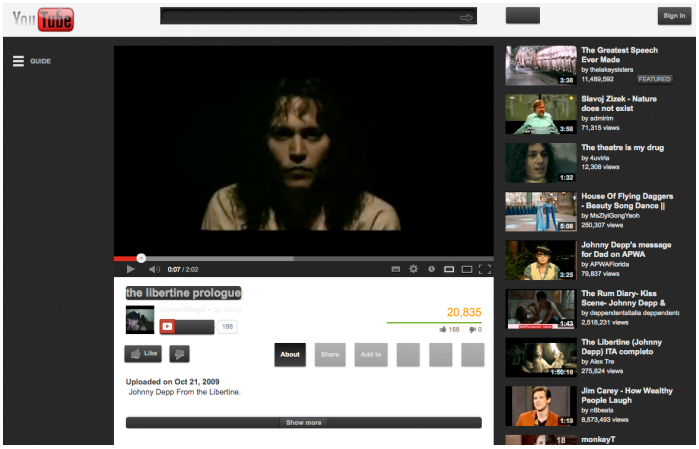


(b) Task 6 b

Fig. B.6: Web-interfaces used for task 6 of experimental study

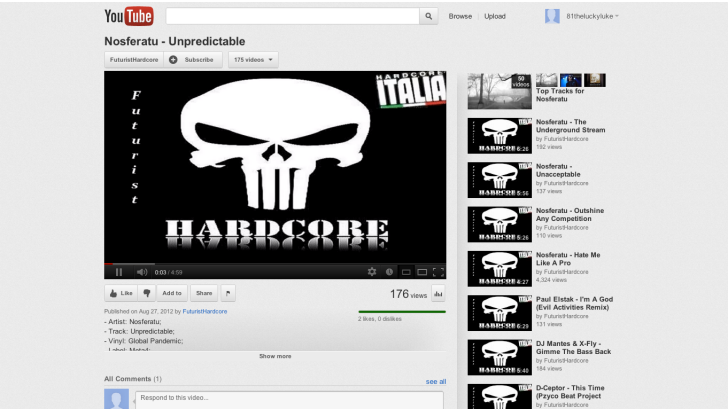


(a) Task 7 a

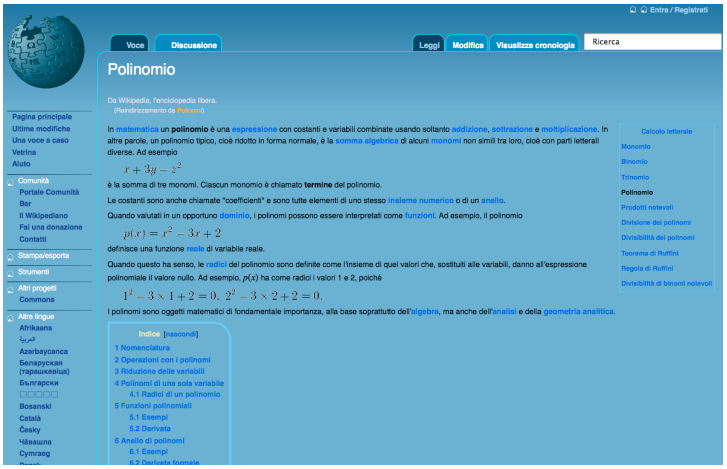


(b) Task 7 b

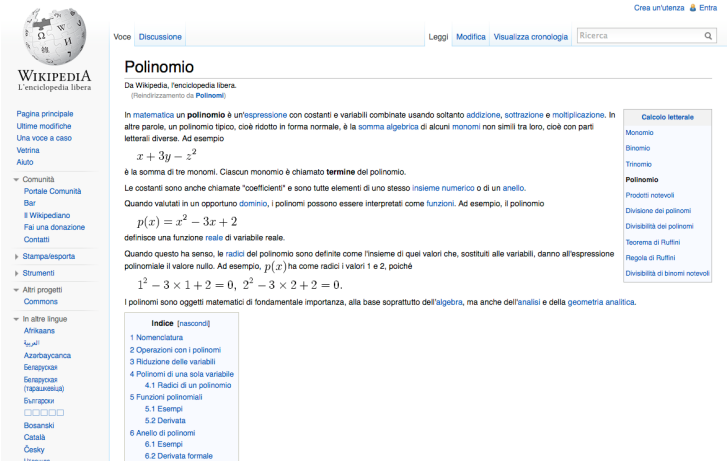
Fig. B.7: Web-interfaces used for task 7 of experimental study



(a) Task 8 a/b - subtask a



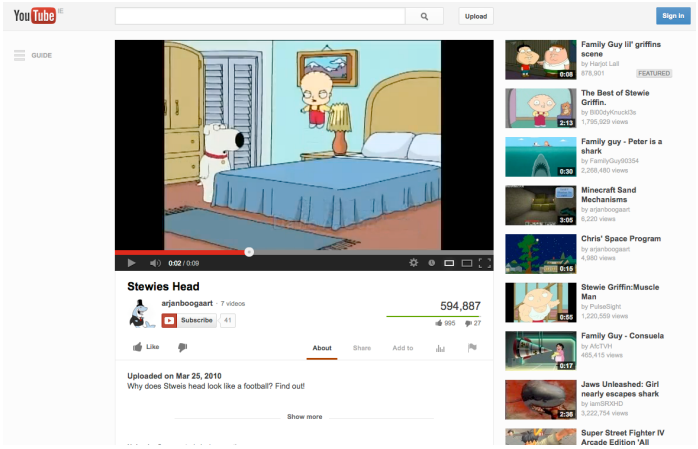
(b) Task 8 a - subtask b



(c) Task 8 b - subtask b

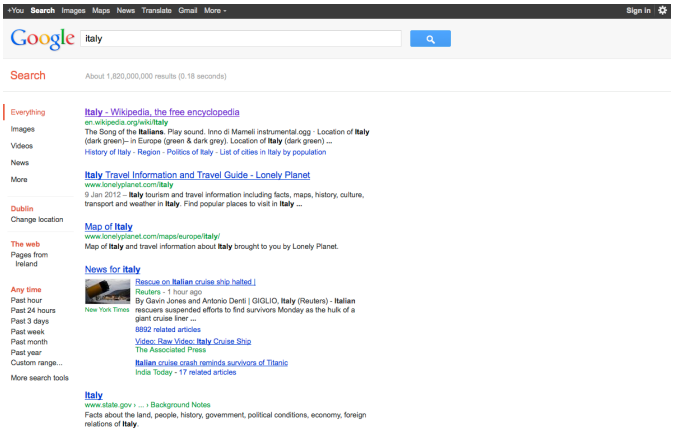
Fig. B.8: Web-interfaces used for task 8 of experimental study

(a) Task 9 a



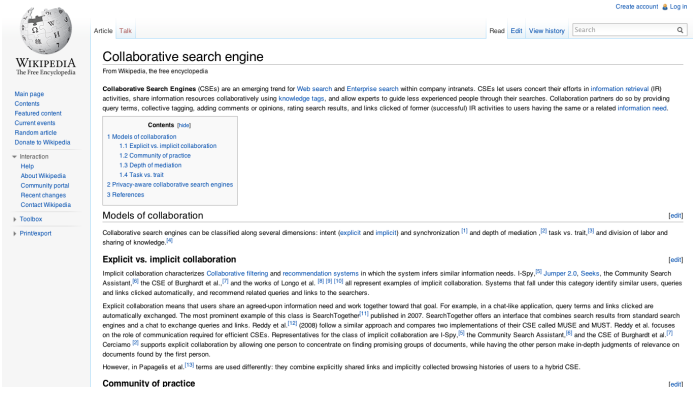
(b) Task 9 b

Fig. B.9: Web-interfaces used for task 9 of experimental study



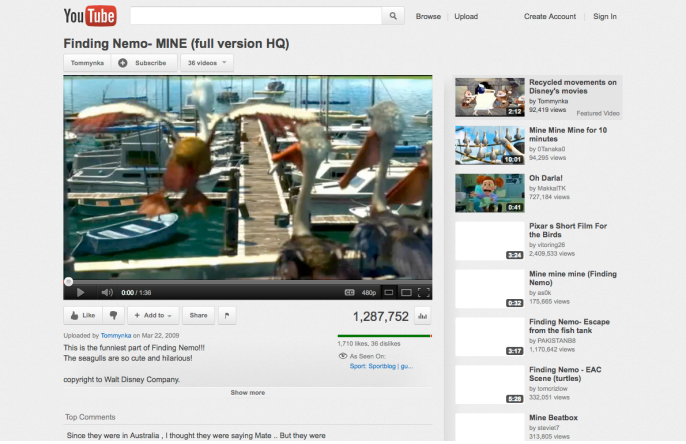
(a) Task 10 a/b - Sub-task a

(b) Task 10 a - Sub-task b/c

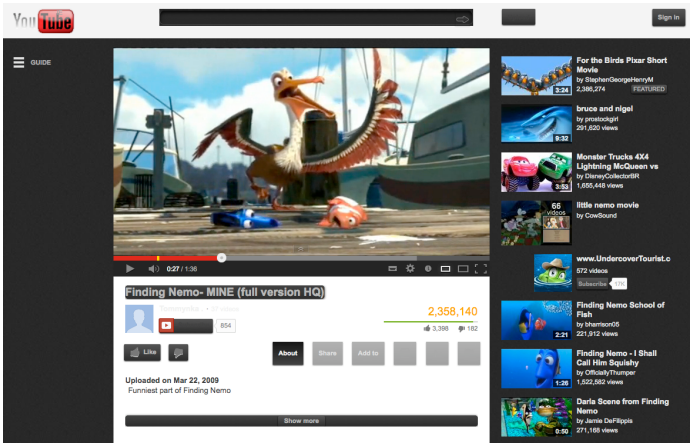


(c) Task 10 b - Sub-task b/c

Fig. B.10: Web-interfaces used for task 10 of experimental study



(a) Task 11 a



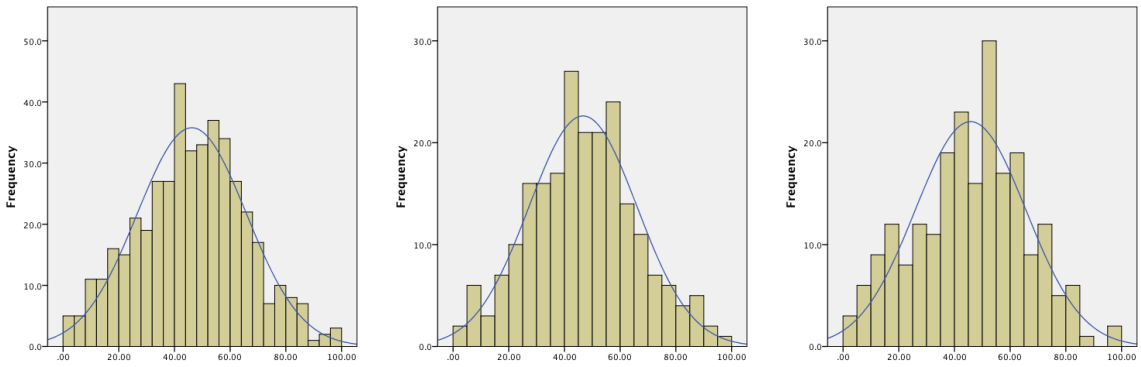
(b) Task 11 b

Fig. B.11: Web-interfaces used for task 11 of experimental study

Appendix C

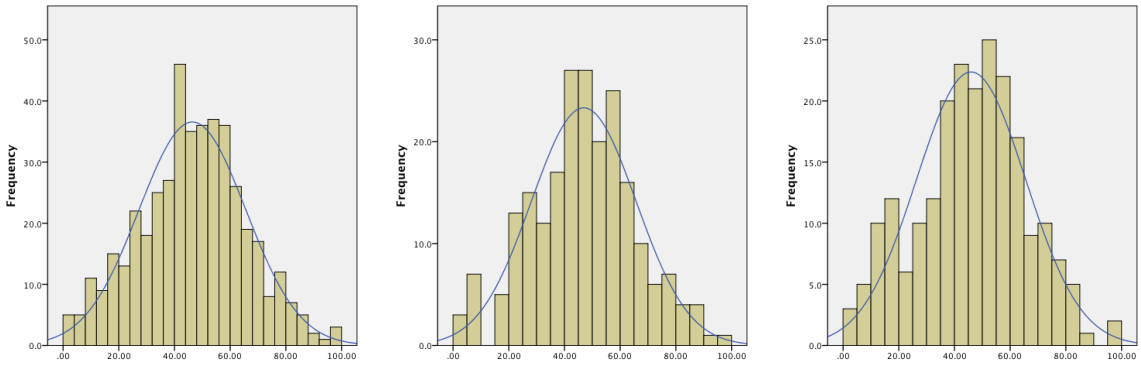
C.1 Results of experiments

C.1.1 Distributions of workload scores



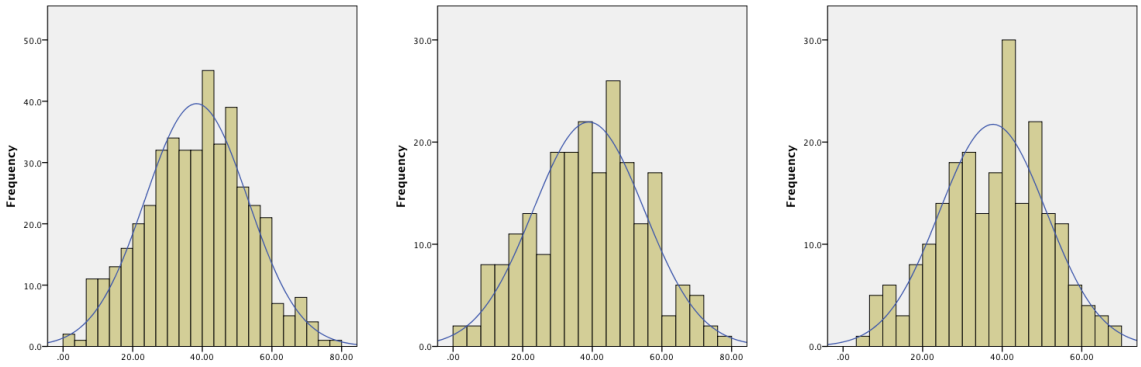
(a) NASATLX -Total (b) NASATLX - Group A (c) NASATLX - Group B

Fig. C.1: Mental workload scores computed by the NASA Task Load Index



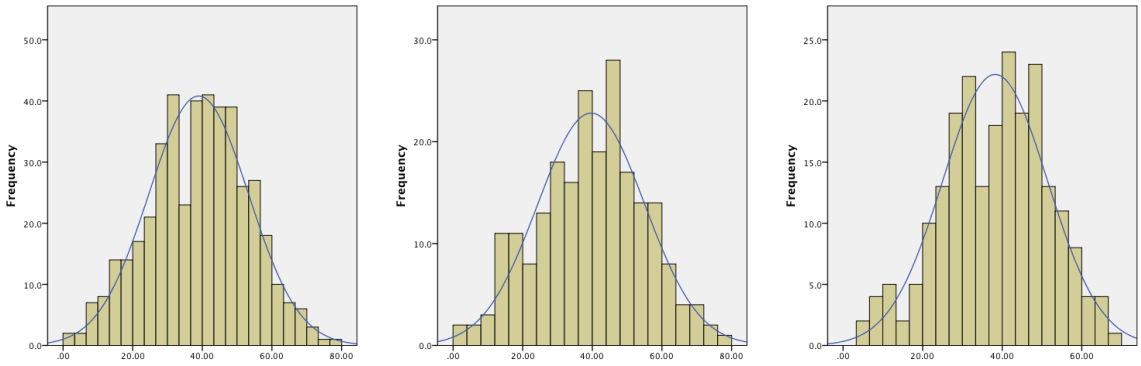
(a) $NASATLX_{def}$ - Total (b) $NASATLX_{def}$ - Group A (c) $NASATLX_{def}$ - Group B

Fig. C.2: Mental workload scores computed by the defeasible translation of the NASA Task Load Index



(a) WP - Total (b) WP - Group A (c) WP - Group B

Fig. C.3: Mental workload scores computed by the Workload Profile instrument



(a) WP_{def} - Total (b) WP_{def} - Group A (c) WP_{def} - Group B

Fig. C.4: Mental workload scores computed by the defeasible translation of the Workload Profile Instrument

C.1.2 Scatterplots of NASA-TLX, WP and their defeasible translations

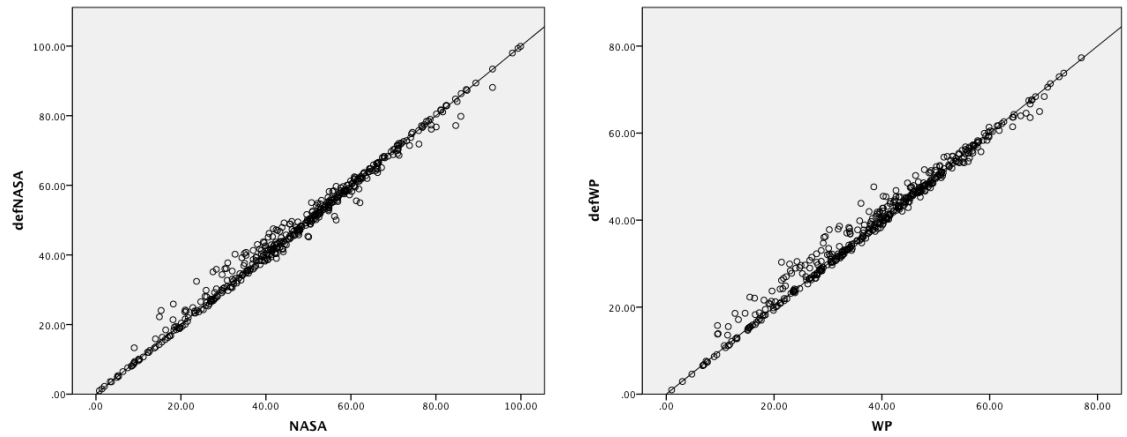


Fig. C.5: Scatterplots of $NASATLX_{def}$ vs NASATLX, WP_{def} vs WP for all the 440 cases

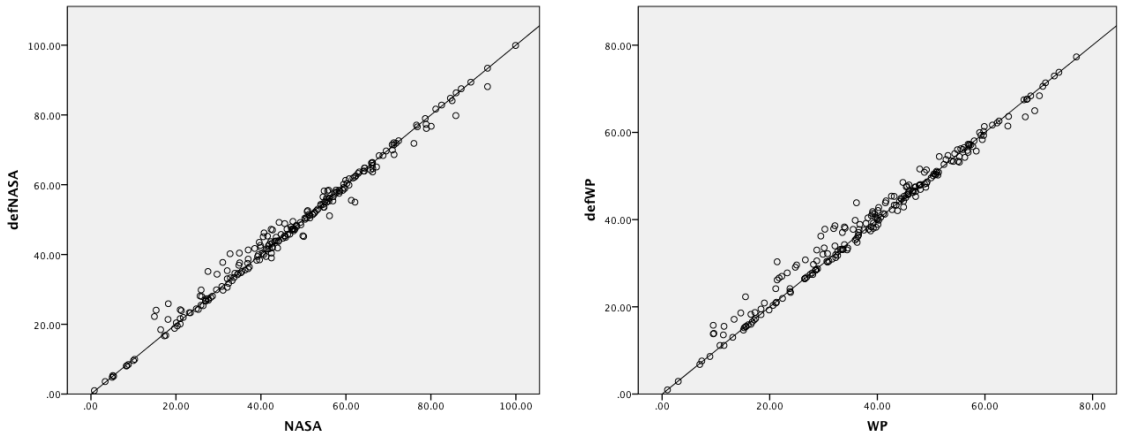


Fig. C.6: Scatterplots of $NASATLX_{def}$ vs NASATLX, WP_{def} vs WP for group A (220 cases)

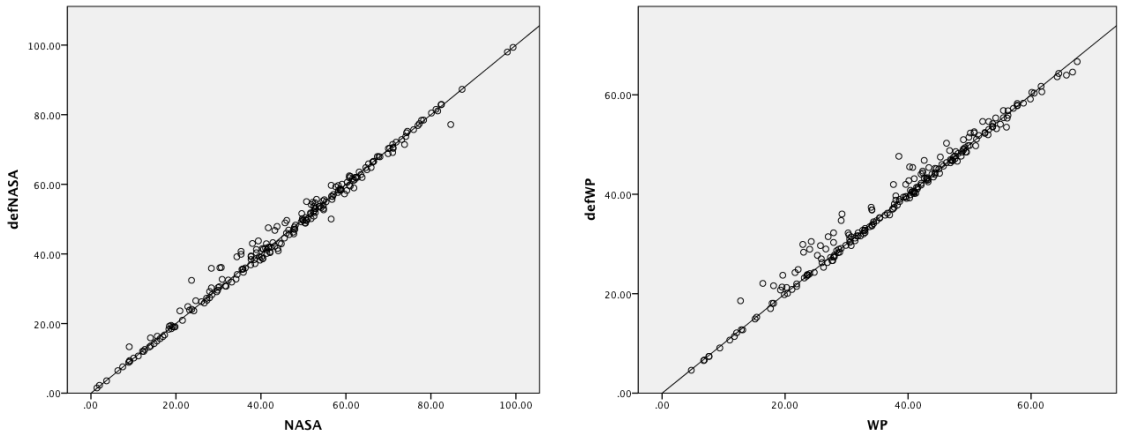
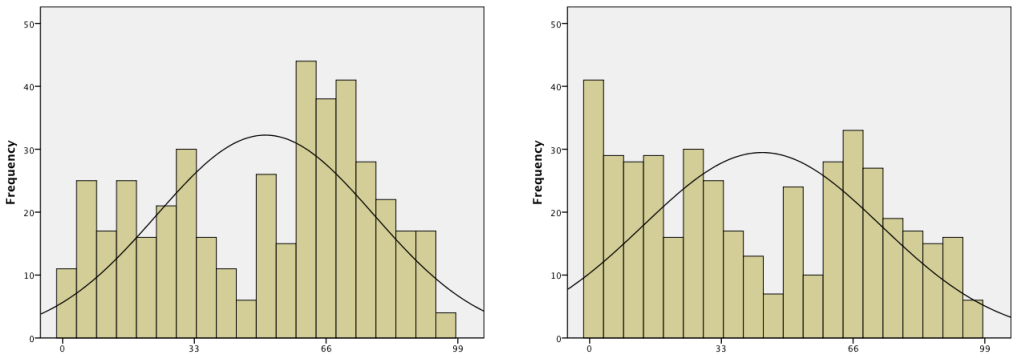
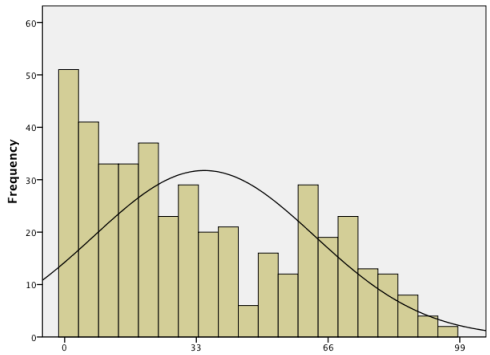


Fig. C.7: Scatterplots of $NASATLX_{def}$ vs NASATLX, WP_{def} vs WP for group B (220 cases)

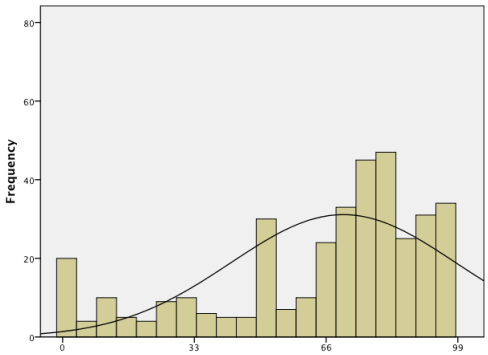
C.1.3 Distributions of mental workload attributes



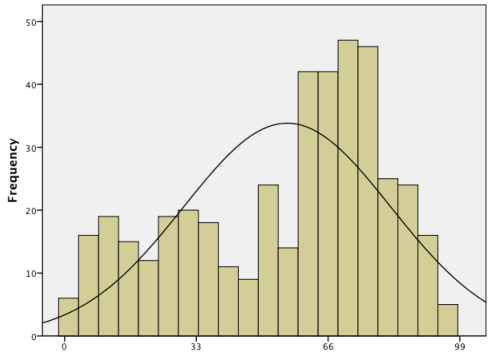
(a) Mental (b) Temporal



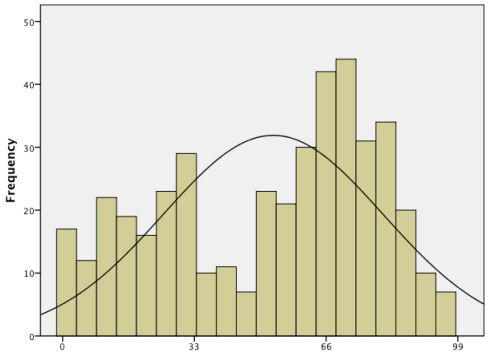
(c) Frustration



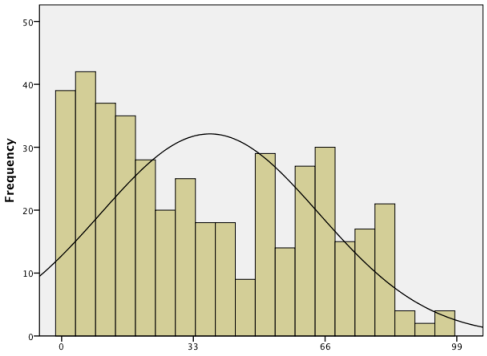
(d) Performance



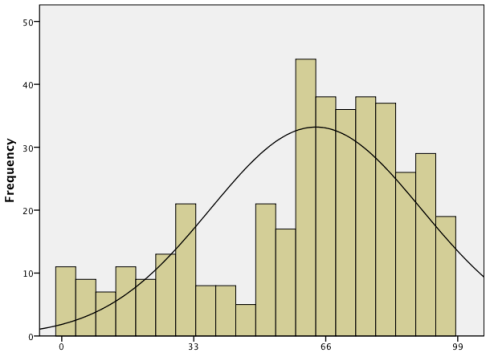
(e) Effort



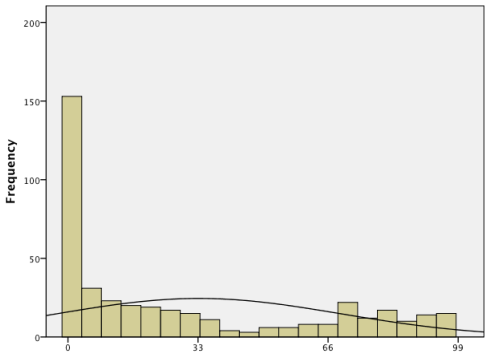
(f) Solving/deciding



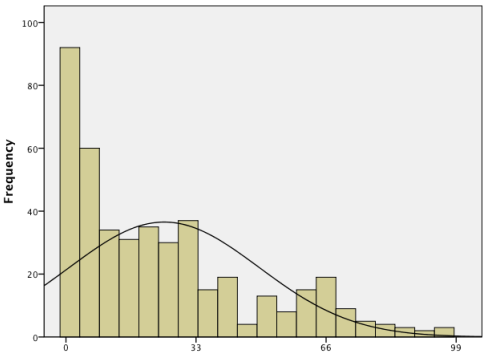
(g) Selection of response



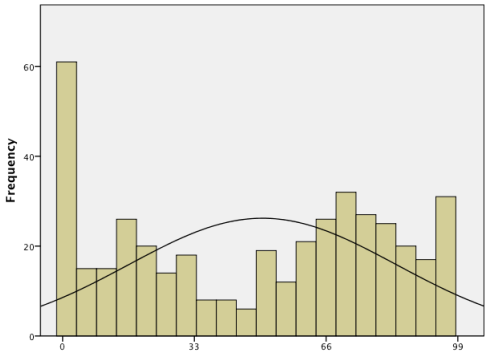
(h) Task and space



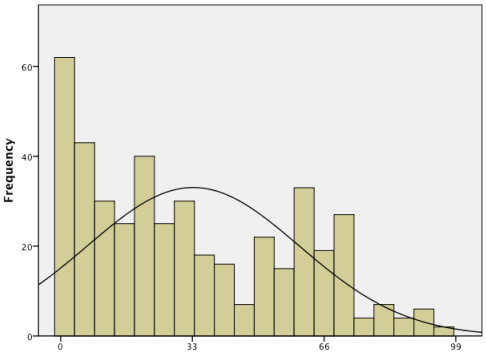
(i) Verbal material



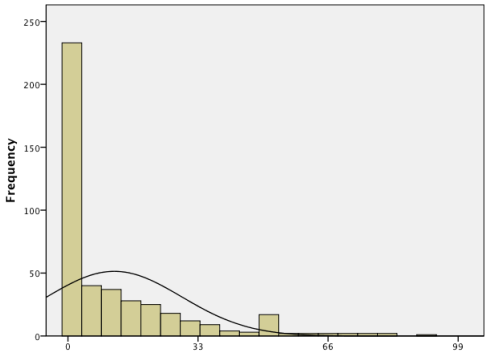
(j) Visual resources



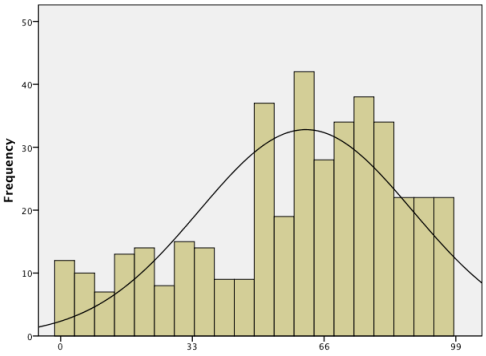
(k) Auditory resources



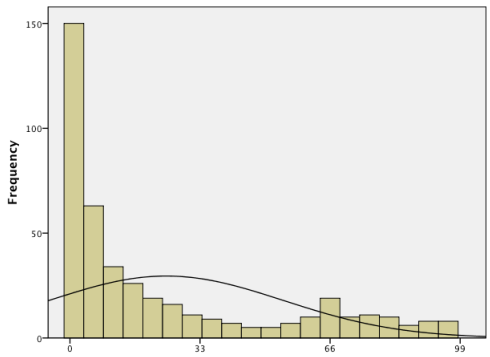
(l) Manual response



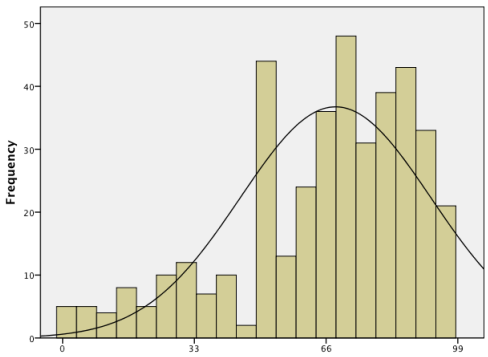
(m) Speech response



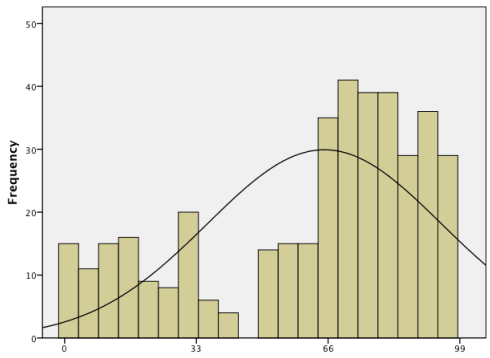
(n) Context bias



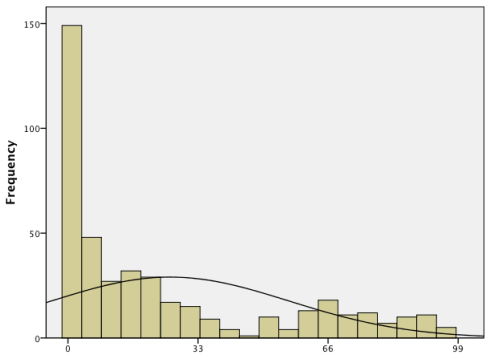
(o) Past knowledge



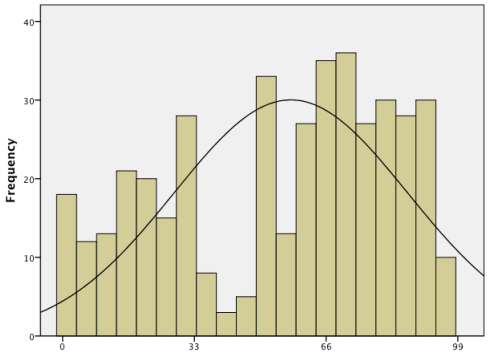
(p) Skill



(q) Motivation



(r) Parallelism



(s) Arousal

Fig. C.8: Distributions of subjective ratings provided by users for mental workload attributes

C.2 Descriptive statistics of mental workload scores

GROUP A					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	22.3835	14.23515	3.18308	15.7212	29.0458	.80	47.67
T_2	20	41.9630	17.81143	3.98276	33.6270	50.2990	8.47	89.40
T_3	20	35.5200	18.57778	4.15412	26.8253	44.2147	5.33	78.67
T_4	20	44.2465	14.15502	3.16516	37.6217	50.8713	19.67	66.00
T_5	20	44.5535	13.02741	2.91302	38.4565	50.6505	20.33	71.80
T_6	20	51.0930	14.36992	3.21321	44.3677	57.8183	25.27	82.53
T_7	20	42.0540	12.74580	2.85005	36.0888	48.0192	21.00	62.13
T_8	20	49.0345	19.54758	4.37097	39.8859	58.1831	14.93	78.87
T_9	20	54.8530	13.70973	3.06559	48.4366	61.2694	25.93	93.33
T_{10}	20	57.4260	17.02675	3.80730	49.4572	65.3948	21.73	93.33
T_{11}	20	70.1705	16.19270	3.62080	62.5921	77.7489	33.67	99.93

GROUP B					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	45.5730	25.31065	5.65963	33.7272	57.4188	9.00	99.33
T_2	20	41.3795	15.70743	3.51229	34.0282	48.7308	11.13	60.80
T_3	20	25.4370	13.32720	2.98005	19.1997	31.6743	1.40	58.87
T_4	20	41.0825	14.47343	3.23636	34.3087	47.8563	9.13	66.53
T_5	20	35.3625	17.92141	4.00735	26.9750	43.7500	6.33	69.93
T_6	20	45.5575	16.45863	3.68026	37.8546	53.2604	10.07	71.80
T_7	20	46.3535	14.13385	3.16042	39.7387	52.9683	12.13	65.33
T_8	20	54.1705	24.12045	5.39350	42.8818	65.4592	2.00	98.00
T_9	20	49.7290	20.81553	4.65450	39.9870	59.4710	3.67	82.40
T_{10}	20	55.2400	14.05001	3.14168	48.6644	61.8156	35.33	81.20
T_{11}	20	63.7870	12.60606	2.81880	57.8872	69.6868	39.60	87.33
Total	440	46.2259	19.62278	.93548	44.3873	48.0645	.80	99.93

Table C.1: Descriptive statistics for workload scores computed by the Nasa Task Load Index

GROUP A					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	26.7935	14.12071	3.15749	20.1848	33.4022	3.00	55.25
T_2	20	30.8945	12.90877	2.88649	24.8530	36.9360	7.38	50.38
T_3	20	33.1820	17.42789	3.89700	25.0255	41.3385	1.00	67.50
T_4	20	39.2440	12.73285	2.84715	33.2848	45.2032	15.50	59.88
T_5	20	37.9245	12.90701	2.88610	31.8838	43.9652	15.62	55.88
T_6	20	34.7510	13.99764	3.12997	28.1999	41.3021	9.62	62.62
T_7	20	46.4370	13.19099	2.94959	40.2634	52.6106	21.12	67.88
T_8	20	37.6435	17.72304	3.96299	29.3489	45.9381	9.50	70.75
T_9	20	43.1815	16.98279	3.79747	35.2333	51.1297	15.25	72.88
T_{10}	20	49.0570	14.10191	3.15328	42.4571	55.6569	18.38	77.00
T_{11}	20	49.5985	13.79522	3.08471	43.1421	56.0549	28.62	71.25

GROUP B					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	37.4550	9.78412	2.18780	32.8759	42.0341	20.25	54.25
T_2	20	27.4190	9.39676	2.10118	23.0212	31.8168	6.75	48.75
T_3	20	29.6370	13.97368	3.12461	23.0971	36.1769	4.75	56.00
T_4	20	36.6115	13.01407	2.91003	30.5207	42.7023	11.00	61.75
T_5	20	34.5610	13.54298	3.02830	28.2227	40.8993	7.62	53.75
T_6	20	35.7325	12.53252	2.80236	29.8671	41.5979	6.88	59.88
T_7	20	43.2125	12.18143	2.72385	37.5114	48.9136	18.12	64.25
T_8	20	36.8250	15.42298	3.44868	29.6068	44.0432	7.62	66.75
T_9	20	40.3575	13.50702	3.02026	34.0360	46.6790	15.38	57.75
T_{10}	20	47.3745	12.14539	2.71579	41.6903	53.0587	16.38	67.50
T_{11}	20	45.0740	9.54769	2.13493	40.6055	49.5425	22.12	61.62
Total	440	38.3167	14.77397	.70432	36.9324	39.7009	1.00	77.00

Table C.2: Descriptive statistics for workload scores computed by the Workload Profile instrument

GROUP A					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	21.3485	11.74392	2.62602	15.8522	26.8448	3.13	52.92
T_2	20	30.8630	13.82903	3.09227	24.3908	37.3352	5.84	57.85
T_3	20	29.5770	16.19947	3.62231	21.9954	37.1586	1.27	57.05
T_4	20	34.6975	12.06062	2.69684	29.0530	40.3420	9.27	53.90
T_5	20	33.3425	13.37591	2.99094	27.0824	39.6026	10.01	66.43
T_6	20	35.5295	13.89124	3.10618	29.0282	42.0308	19.30	70.63
T_7	20	41.9060	12.91155	2.88711	35.8632	47.9488	21.18	67.38
T_8	20	40.6505	18.04984	4.03607	32.2029	49.0981	15.72	70.83
T_9	20	49.3355	15.68105	3.50639	41.9965	56.6745	24.64	81.44
T_{10}	20	53.8735	14.53974	3.25118	47.0687	60.6783	25.93	80.19
T_{11}	20	54.7720	13.01265	2.90972	48.6819	60.8621	28.39	74.21

GROUP B					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	37.4575	15.64352	3.49800	30.1361	44.7789	14.68	73.41
T_2	20	23.5040	10.03791	2.24455	18.8061	28.2019	1.74	44.93
T_3	20	23.2470	11.42698	2.55515	17.8990	28.5950	6.02	44.03
T_4	20	30.9045	13.61943	3.04540	24.5304	37.2786	4.99	53.33
T_5	20	29.3320	16.58838	3.70927	21.5684	37.0956	1.60	67.43
T_6	20	34.1395	15.57406	3.48247	26.8506	41.4284	4.04	71.81
T_7	20	38.0055	12.72641	2.84571	32.0494	43.9616	9.31	56.36
T_8	20	39.5910	19.34804	4.32635	30.5358	48.6462	3.08	83.04
T_9	20	44.5445	15.28606	3.41807	37.3904	51.6986	20.36	71.81
T_{10}	20	52.3905	14.36866	3.21293	45.6658	59.1152	24.79	78.87
T_{11}	20	48.3145	11.05157	2.47121	43.1422	53.4868	25.26	67.11
Total	440	37.6057	16.97698	0.80935	36.0151	39.1964	1.27	83.04

Table C.3: Descriptive statistics for workload scores computed by the new instances of the defeasible framework (MWL_{def})

GROUP A					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	21.4905	11.63171	2.60093	16.0467	26.9343	2.13	49.34
T_2	20	32.5410	11.47619	2.56616	27.1700	37.9120	11.79	52.37
T_3	20	30.8270	14.67742	3.28197	23.9578	37.6962	7.50	59.29
T_4	20	35.5680	12.39429	2.77145	29.7673	41.3687	8.50	55.97
T_5	20	34.9070	13.27832	2.96912	28.6926	41.1214	9.52	66.30
T_6	20	35.7520	13.06679	2.92182	29.6366	41.8674	18.92	66.24
T_7	20	40.3765	12.00162	2.68364	34.7596	45.9934	19.32	63.24
T_8	20	40.4600	17.29922	3.86822	32.3637	48.5563	16.66	64.67
T_9	20	48.1935	14.46117	3.23362	41.4255	54.9615	25.22	76.47
T_{10}	20	53.1445	13.13388	2.93682	46.9977	59.2913	25.52	78.46
T_{11}	20	54.2205	12.65252	2.82919	48.2989	60.1421	28.20	74.63

GROUP B					95% Conf. Int. for mean			
Task Id	N	Mean	Std	Std. Error	Lower bound	Upper bound	Min	Max
T_1	20	36.4205	15.35523	3.43353	29.2340	43.6070	13.18	70.27
T_2	20	25.5215	9.38281	2.09806	21.1302	29.9128	8.30	47.68
T_3	20	22.8010	11.87626	2.65561	17.2427	28.3593	5.31	46.95
T_4	20	31.8315	13.80714	3.08737	25.3696	38.2934	4.44	53.16
T_5	20	29.2900	16.64239	3.72135	21.5011	37.0789	1.05	63.95
T_6	20	34.0240	15.10134	3.37676	26.9564	41.0916	3.33	66.48
T_7	20	38.7985	12.98923	2.90448	32.7194	44.8776	9.02	57.38
T_8	20	40.0460	19.31031	4.31792	31.0085	49.0835	2.54	81.60
T_9	20	43.4750	15.12126	3.38122	36.3980	50.5520	19.90	68.49
T_10	20	52.3685	14.33612	3.20565	45.6590	59.0780	22.25	78.87
T_11	20	48.1560	10.14175	2.26777	43.4095	52.9025	24.07	66.88
Total	440	37.7370	16.31292	0.77769	36.2085	39.2654	1.05	81.60

Table C.4: Descriptive statistics for workload scores computed by the new instances of the defeasible framework (MWL^{NI}_{def})

C.3 Tests of normality of distributions of computed workload scores

	Group A			Group A - no outliers			Group B			Group B - no outliers		
Task	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.
T_1	0.953	20	0.416	0.953	20	0.416	0.953	20	0.418	0.953	20	0.418
T_2	0.956	20	0.465	0.959	19	0.557	0.887	20	0.024	0.956	16	0.590
T_3	0.966	20	0.660	0.966	20	0.660	0.951	20	0.382	0.951	20	0.382
T_4	0.960	20	0.549	0.960	20	0.549	0.979	20	0.923	0.979	20	0.923
T_5	0.982	20	0.959	0.982	20	0.959	0.966	20	0.663	0.966	20	0.663
T_6	0.983	20	0.969	0.983	20	0.969	0.911	20	0.066	0.911	20	0.066
T_7	0.927	20	0.132	0.927	20	0.132	0.901	20	0.042	0.963	18	0.654
T_8	0.940	20	0.245	0.940	20	0.245	0.976	20	0.874	0.976	20	0.874
T_9	0.936	20	0.198	0.971	18	0.812	0.970	20	0.746	0.970	20	0.746
T_{10}	0.938	20	0.936	0.938	20	0.936	0.947	20	0.325	0.947	20	0.325
T_{11}	0.977	20	0.894	0.977	20	0.894	0.987	20	0.990	0.987	20	0.990

Table C.5: Shapiro-Wilk normality tests of the workload scores computed by the Nasa Task Load Index

	Group A			Group A - no outliers			Group B			Group B - no outliers		
Task	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig
T_1	0.968	20	0.703	0.968	20	0.703	0.976	20	0.864	0.976	20	0.864
T_2	0.950	20	0.365	0.950	20	0.365	0.986	20	0.986	0.983	18	0.972
T_3	0.983	20	0.966	0.983	20	0.966	0.968	20	0.712	0.968	20	0.712
T_4	0.948	20	0.343	0.948	20	0.343	0.973	20	0.823	0.973	20	0.823
T_5	0.928	20	0.138	0.928	20	0.138	0.936	20	0.198	0.936	20	0.198
T_6	0.979	20	0.926	0.979	20	0.926	0.956	20	0.475	0.973	19	0.838
T_7	0.967	20	0.696	0.967	20	0.696	0.957	20	0.490	0.957	20	0.490
T_8	0.964	20	0.617	0.964	20	0.617	0.970	20	0.746	0.970	20	0.746
T_9	0.973	20	0.818	0.973	20	0.818	0.930	20	0.157	0.930	20	0.157
T_{10}	0.971	20	0.779	0.971	20	0.779	0.969	20	0.727	0.983	19	0.971
T_{11}	0.921	20	0.105	0.921	20	0.105	0.974	20	0.835	0.977	19	0.905

Table C.6: Shapiro-Wilk normality tests of the workload scores computed by the Workload Profile instrument

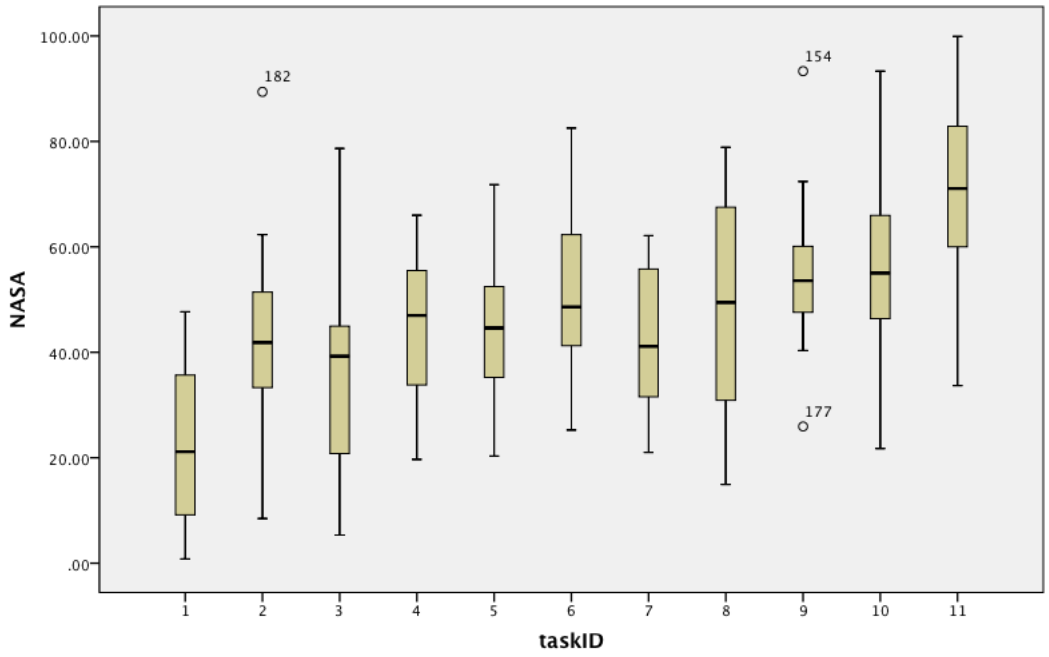
	Group A			Group A - no outliers			Group B			Group B - no outliers		
Task	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig
T_1	0.942	20	0.262	0.940	19	0.259	0.958	20	0.506	0.958	20	0.506
T_2	0.984	20	0.972	0.984	20	0.972	0.985	20	0.980	0.985	20	0.980
T_3	0.971	20	0.767	0.971	20	0.767	0.929	20	0.145	0.929	20	0.145
T_4	0.971	20	0.780	0.971	20	0.780	0.968	20	0.706	0.968	20	0.706
T_5	0.970	20	0.746	0.981	19	0.954	0.973	20	0.815	0.973	20	0.815
T_6	0.907	20	0.057	0.907	20	0.057	0.979	20	0.915	0.975	19	0.867
T_7	0.952	20	0.393	0.952	20	0.393	0.930	20	0.155	0.942	18	0.319
T_8	0.931	20	0.161	0.931	20	0.161	0.960	20	0.534	0.938	19	0.238
T_9	0.963	20	0.610	0.963	20	0.610	0.954	20	0.437	0.954	20	0.437
T_{10}	0.976	20	0.879	0.976	20	0.879	0.986	20	0.989	0.986	20	0.989
T_{11}	0.967	20	0.688	0.967	20	0.688	0.970	20	0.761	0.970	20	0.761

Table C.7: Shapiro-Wilk normality tests of the workload scores computed by the instance MWL_{def} of the defeasible framework

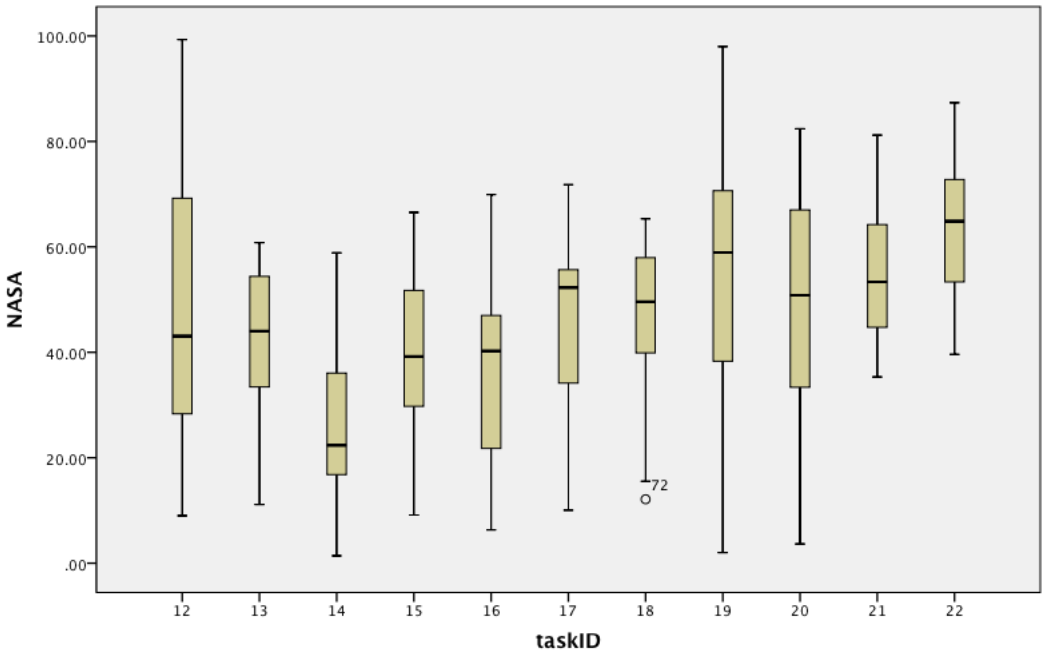
	Group A			Group A - no outliers			Group B			Group B - no outliers		
Task	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig.	Statistic	df	Sig
T_1	0.944	20	0.287	0.944	20	0.287	0.960	20	0.537	0.960	20	0.537
T_2	0.966	20	0.675	0.966	20	0.675	0.948	20	0.343	0.948	20	0.343
T_3	0.966	20	0.679	0.966	20	0.679	0.942	20	0.258	0.942	20	0.258
T_4	0.975	20	0.848	0.975	20	0.848	0.966	20	0.659	0.959	18	0.577
T_5	0.976	20	0.876	0.981	19	0.956	0.976	20	0.872	0.976	19	0.880
T_6	0.933	20	0.177	0.933	20	0.177	0.979	20	0.921	0.979	20	0.921
T_7	0.971	20	0.766	0.971	20	0.766	0.920	20	0.099	0.929	18	0.186
T_8	0.910	20	0.065	0.910	20	0.065	0.953	20	0.410	0.955	19	0.487
T_9	0.961	20	0.562	0.961	20	0.562	0.953	20	0.420	0.953	20	0.420
T_{10}	0.964	20	0.636	0.964	20	0.636	0.990	20	0.998	0.990	20	0.998
T_{11}	0.981	20	0.942	0.981	20	0.942	0.968	20	0.711	0.968	20	0.711

Table C.8: Shapiro-Wilk normality tests of the workload scores computed by the instance MWL^{NI}_{def} of the defeasible framework

C.4 Boxplots of the computed mental workload scores

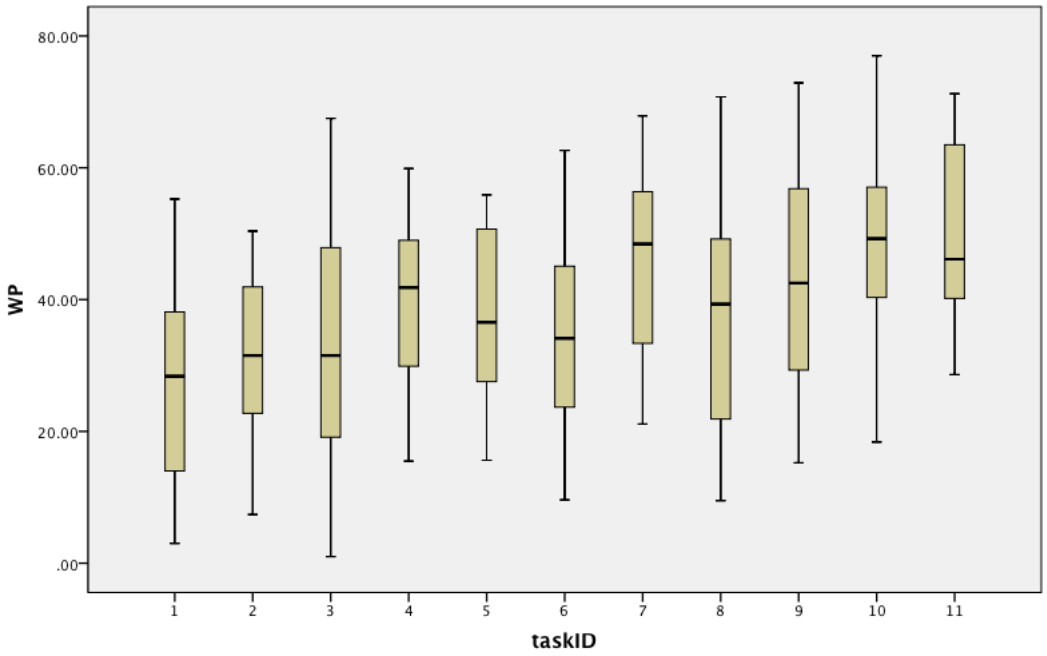


(a) Group A

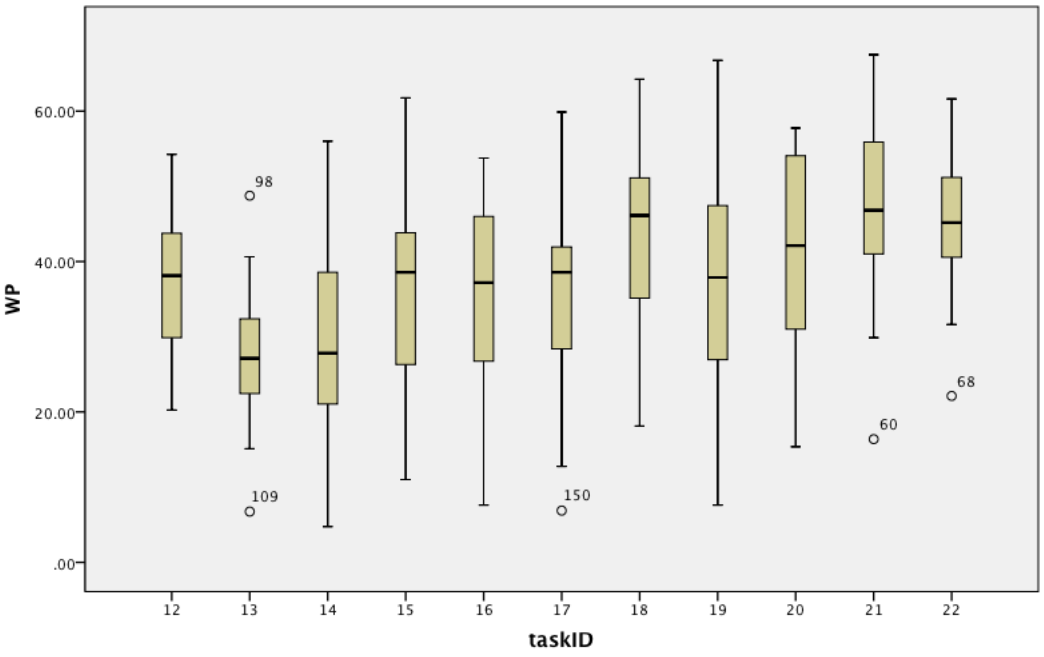


(b) Group B

Fig. C.9: Boxplots of the mental workload scores computed by the Nasa Task Load Index

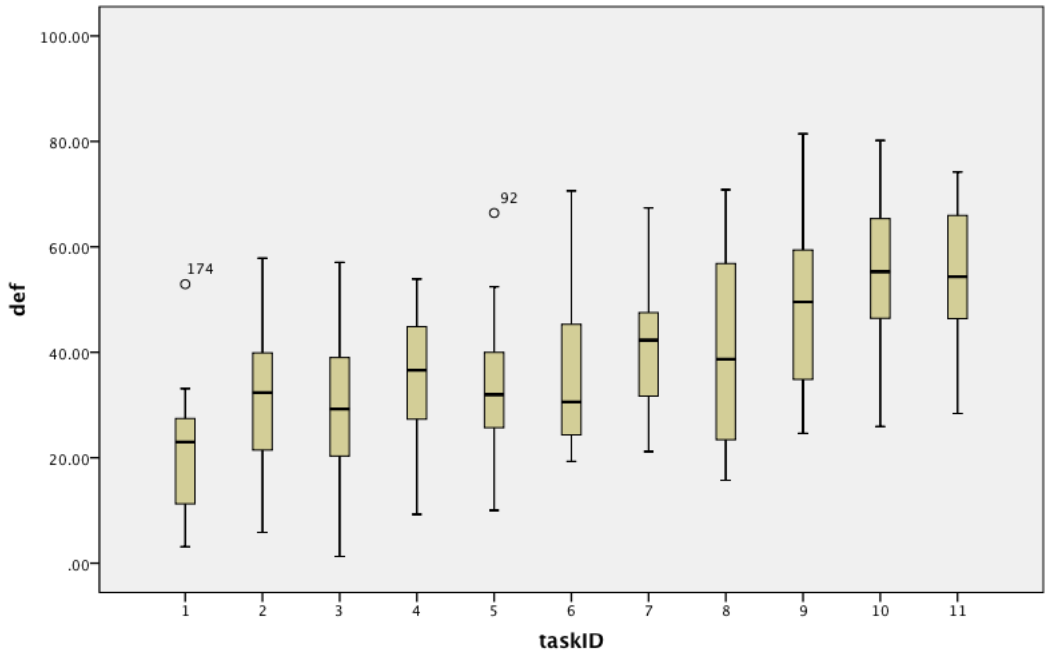


(a) Group A

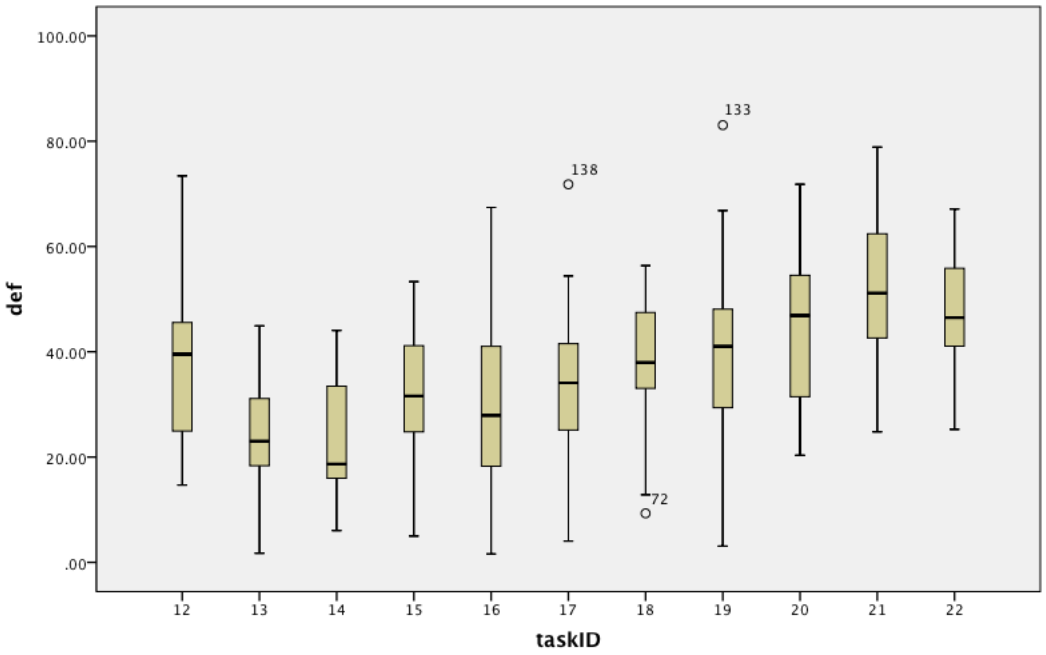


(b) Group B

Fig. C.10: Boxplots of the mental workload scores computed by the Workload Profile isntrument

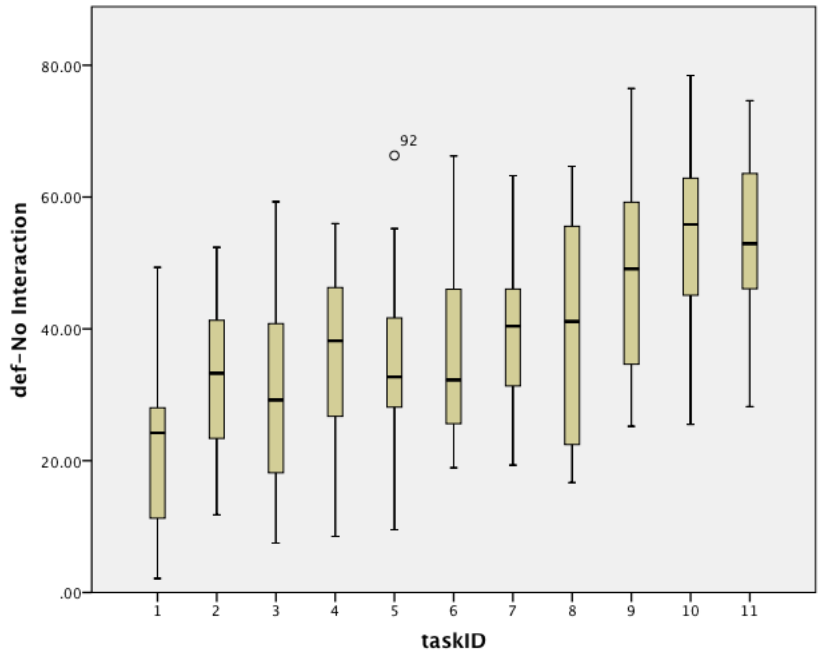


(a) Group A

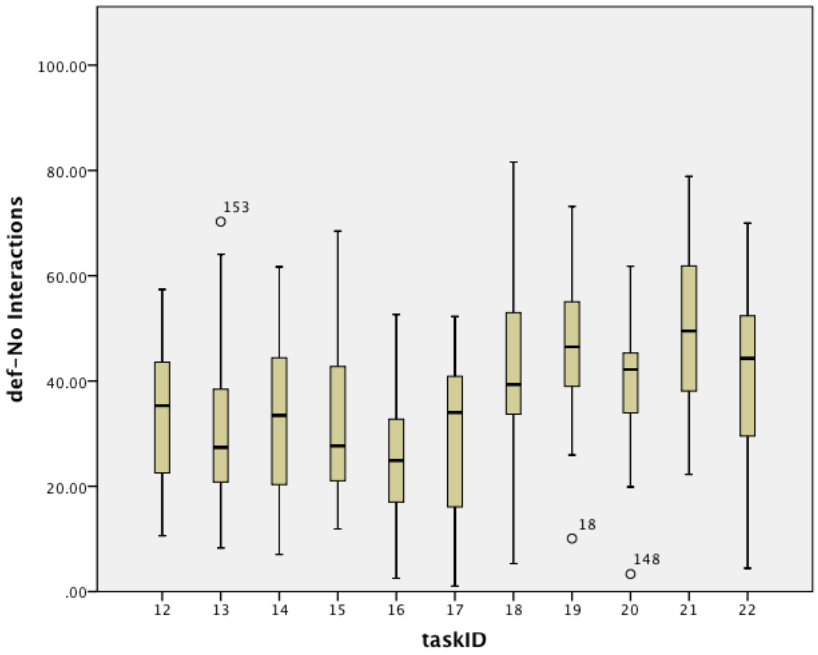


(b) Group B

Fig. C.11: Boxplots of the mental workload scores computed by the instance MWL_{def} of the defeasible framework



(a) Group A



(b) Group B

Fig. C.12: Boxplots of the mental workload scores computed by the instance MWL_{def}^{NI} of the defeasible framework

C.5 Post-hoc Anova results

NASATLX - Games-Howell					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-17.08282 *	4.56407	.023	-32.7258	-1.4398
	3	-13.13650	5.23342	.333	-31.1146	4.8416
	4	-21.86300 *	4.48890	.001	-37.2204	-6.5056
	5	-22.17000 *	4.31482	.000	-36.9388	-7.4012
	6	-28.70950 *	4.52291	.000	-44.1833	-13.2357
	7	-19.67050 *	4.27256	.002	-34.2985	-5.0425
	8	-26.65100 *	5.40716	.001	-45.2558	-8.0462
	9	-31.93872 *	3.77014	.000	-44.9947	-18.8827
	10	-35.04250 *	4.96261	.000	-52.0529	-18.0321
	11	-47.78700 *	4.82101	.000	-64.2970	-31.2770
2	3	3.94632	5.28730	1.000	-14.2209	22.1135
	4	-4.78018	4.55160	.992	-20.3811	10.8207
	5	-5.08718	4.38001	.983	-20.1156	9.9412
	6	-11.62668	4.58514	.318	-27.3410	4.0876
	7	-2.58768	4.33838	1.000	-17.4793	12.3040
	8	-9.56818	5.45932	.798	-28.3521	9.2157
	9	-14.85591 *	3.84458	.020	-28.2310	-1.4808
	10	-17.95968 *	5.01939	.035	-35.1765	-.7428
	11	-30.70418 *	4.87944	.000	-47.4308	-13.9776
3	4	-8.72650	5.22254	.840	-26.6702	9.2172
	5	-9.03350	5.07369	.783	-26.5141	8.4471
	6	-15.57300	5.25180	.143	-33.6093	2.4633
	7	-6.53400	5.03780	.963	-23.9051	10.8371
	8	-13.51450	6.03010	.492	-34.1478	7.1188
	9	-18.80222 *	4.61938	.013	-34.9832	-2.6213
	10	-21.90600 *	5.63491	.015	-41.1929	-2.6191
	11	-34.65050 *	5.51061	.000	-53.5247	-15.7763
4	5	-.30700	4.30162	1.000	-15.0298	14.4158
	6	-6.84650	4.51032	.905	-22.2774	8.5844
	7	2.19250	4.25923	1.000	-12.3888	16.7738
	8	-4.78800	5.39663	.998	-23.3601	13.7841
	9	-10.07572	3.75503	.252	-23.0765	2.9251
	10	-13.17950	4.95114	.256	-30.1526	3.7936
	11	-25.92400 *	4.80920	.000	-42.3950	-9.4530
5	6	-6.53950	4.33710	.908	-21.3861	8.3071
	7	2.49950	4.07535	1.000	-11.4434	16.4424
	8	-4.48100	5.25272	.998	-22.6136	13.6516
	9	-9.76872	3.54509	.219	-22.0061	2.4686
	10	-12.87250	4.79387	.246	-29.3414	3.5964
	11	-25.61700 *	4.64713	.000	-41.5600	-9.6740
6	7	9.03900	4.29506	.581	-5.6679	23.7459
	8	2.05850	5.42495	1.000	-16.6017	20.7187
	9	-3.22922	3.79562	.998	-16.3783	9.9199
	10	-6.33300	4.98199	.968	-23.4065	10.7405
	11	-19.07750 *	4.84096	.013	-35.6535	-2.5015
7	8	-6.98050	5.21806	.955	-25.0093	11.0483
	9	-12.26822 *	3.49353	.043	-24.3191	-.2173
	10	-15.37200	4.75587	.081	-31.7212	.9772
	11	-28.11650 *	4.60792	.000	-43.9339	-12.2991
8	9	-5.28772	4.81533	.988	-22.1930	11.6176
	10	-8.39150	5.79663	.927	-28.2455	11.4625
	11	-21.13600 *	5.67588	.024	-40.5951	-1.6769
9	10	-3.10378	4.31016	1.000	-18.1431	11.9356
	11	-15.84828 *	4.14634	.022	-30.2841	-1.4125
10	11	-12.74450	5.25411	.378	-30.7225	5.2335

* The mean difference is significant at the 0.05 level

Table C.9: ANOVA Post-hoc tests for the Nasa Task Load Index - Group A - 95% Confidence Interval

NASATLX - Games-Howell					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-17.08282	4.56407	.023	-35.5629	1.3973
	3	-13.13650	5.23342	.333	-34.3980	8.1250
	4	-21.86300 **	4.48890	.001	-39.9901	-3.7359
	5	-22.17000 **	4.31482	.000	-39.6062	-4.7338
	6	-28.70950 **	4.52291	.000	-46.9740	-10.4450
	7	-19.67050 **	4.27256	.002	-36.9426	-2.3984
	8	-26.65100 **	5.40716	.001	-48.6704	-4.6316
	9	-31.93872 **	3.77014	.000	-47.4383	-16.4391
	10	-35.04250 **	4.96261	.000	-55.1387	-14.9463
	11	-47.78700 **	4.82101	.000	-67.2837	-28.2903
2	3	3.94632	5.28730	1.000	-17.5410	25.4337
	4	-4.78018	4.55160	.992	-23.2108	13.6505
	5	-5.08718	4.38001	.983	-22.8502	12.6758
	6	-11.62668	4.58514	.318	-30.1905	6.9371
	7	-2.58768	4.33838	1.000	-20.1925	15.0171
	8	-9.56818	5.45932	.798	-31.7994	12.6630
	9	-14.85591	3.84458	.020	-30.7693	1.0574
	10	-17.95968	5.01939	.035	-38.3064	2.3870
	11	-30.70418 **	4.87944	.000	-50.4659	-10.9425
3	4	-8.72650	5.22254	.840	-29.9490	12.4960
	5	-9.03350	5.07369	.783	-29.7354	11.6684
	6	-15.57300	5.25180	.143	-36.9006	5.7546
	7	-6.53400	5.03780	.963	-27.1141	14.0461
	8	-13.51450	6.03010	.492	-37.8708	10.8418
	9	-18.80222	4.61938	.013	-38.1174	.5129
	10	-21.90600	5.63491	.015	-44.6762	.8642
	11	-34.65050 **	5.51061	.000	-56.9405	-12.3605
4	5	-.30700	4.30162	1.000	-17.6884	17.0744
	6	-6.84650	4.51032	.905	-25.0604	11.3674
	7	2.19250	4.25923	1.000	-15.0238	19.4088
	8	-4.78800	5.39663	.998	-26.7707	17.1947
	9	-10.07572	3.75503	.252	-25.5082	5.3567
	10	-13.17950	4.95114	.256	-33.2327	6.8737
	11	-25.92400 **	4.80920	.000	-45.3755	-6.4725
5	6	-6.53950	4.33710	.908	-24.0685	10.9895
	7	2.49950	4.07535	1.000	-13.9582	18.9572
	8	-4.48100	5.25272	.998	-25.9748	17.0128
	9	-9.76872	3.54509	.219	-24.2743	4.7368
	10	-12.87250	4.79387	.246	-32.3495	6.6045
	11	-25.61700 **	4.64713	.000	-44.4599	-6.7741
6	7	9.03900	4.29506	.581	-8.3273	26.4053
	8	2.05850	5.42495	1.000	-20.0232	24.1402
	9	-3.22922	3.79562	.998	-18.8420	12.3836
	10	-6.33300	4.98199	.968	-26.5020	13.8360
	11	-19.07750	4.84096	.013	-38.6509	.4959
7	8	-6.98050	5.21806	.955	-28.3602	14.3992
	9	-12.26822	3.49353	.043	-26.5479	2.0115
	10	-15.37200	4.75587	.081	-34.7136	3.9696
	11	-28.11650 **	4.60792	.000	-46.8160	-9.4170
8	9	-5.28772	4.81533	.988	-25.4895	14.9141
	10	-8.39150	5.79663	.927	-31.8387	15.0557
	11	-21.13600	5.67588	.024	-44.1272	1.8552
9	10	-3.10378	4.31016	1.000	-21.0226	14.8151
	11	-15.84828	4.14634	.022	-33.0298	1.3333
10	11	-12.74450	5.25411	.378	-33.9664	8.4774

** The mean difference is significant at the 0.01 level

Table C.10: ANOVA Post-hoc tests for the Nasa Task Load Index - Group A - 99% Confidence Interval

NASATLX - Games-Howell					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-2.38075	6.06702	1.000	-23.8482	19.0867
	3	20.13600	6.39626	.105	-2.1751	42.4471
	4	4.49050	6.51962	1.000	-18.1645	27.1455
	5	10.21050	6.93472	.919	-13.6736	34.0946
	6	.01550	6.75098	1.000	-23.3123	23.3433
	7	-4.39422	6.06090	1.000	-25.8402	17.0517
	8	-8.59750	7.81801	.989	-35.3481	18.1531
	9	-4.15600	7.32774	1.000	-29.2823	20.9703
	10	-9.66700	6.47314	.911	-32.1911	12.8571
	11	-18.21400	6.32274	.180	-40.3257	3.8977
2	3	22.51675 *	3.69566	.000	9.7535	35.2800
	4	6.87125	3.90528	.794	-6.6456	20.3881
	5	12.59125	4.56465	.224	-3.3319	28.5144
	6	2.39625	4.28037	1.000	-12.4848	17.2773
	7	-2.01347	3.07902	1.000	-12.6715	8.6445
	8	-6.21675	5.81954	.990	-26.7698	14.3363
	9	-1.77525	5.14214	1.000	-19.8259	16.2754
	10	-7.28625	3.82719	.710	-20.5214	5.9489
	11	-15.83325 *	3.56691	.004	-28.1382	-3.5283
3	4	-15.64550 *	4.39940	.036	-30.7029	-.5881
	5	-9.92550	4.99395	.658	-27.0967	7.2457
	6	-20.12050 *	4.73551	.006	-36.3641	-3.8769
	7	-24.53022 *	3.68562	.000	-37.2336	-11.8268
	8	-28.73350 *	6.16202	.003	-50.1796	-7.2874
	9	-24.29200 *	5.52676	.005	-43.4018	-5.1822
	10	-29.80300 *	4.33022	.000	-44.6200	-14.9860
	11	-38.35000 *	4.10199	.000	-52.3863	-24.3137
4	5	5.72000	5.15100	.988	-11.9499	23.3899
	6	-4.47500	4.90085	.997	-21.2583	12.3083
	7	-8.88472	3.89578	.469	-22.3486	4.5791
	8	-13.08800	6.28998	.598	-34.8977	8.7217
	9	-8.64650	5.66907	.901	-28.1845	10.8915
	10	-14.15750	4.51045	.097	-29.5894	1.2744
	11	-22.70450 *	4.29181	.000	-37.4044	-8.0046
5	6	-10.19500	5.44088	.729	-28.8174	8.4274
	7	-14.60472	4.55653	.093	-30.4899	1.2805
	8	-18.80800	6.71927	.200	-41.9123	4.2963
	9	-14.36650	6.14192	.430	-35.4073	6.6743
	10	-19.87750 *	5.09205	.015	-37.3586	-2.3964
	11	-28.42450 *	4.89944	.000	-45.3028	-11.5462
6	7	-4.40972	4.27170	.993	-19.2473	10.4278
	8	-8.61300	6.52948	.959	-31.1323	13.9063
	9	-4.17150	5.93369	1.000	-24.5373	16.1943
	10	-9.68250	4.83885	.649	-26.2618	6.8968
	11	-18.22950 *	4.63573	.014	-34.1546	-2.3044
7	8	-4.20328	5.81317	1.000	-24.7330	16.3265
	9	.23822	5.13492	1.000	-17.7828	18.2592
	10	-5.27278	3.81749	.945	-18.4525	7.9070
	11	-13.81978 *	3.55651	.016	-26.0599	-1.5796
8	9	4.44150	7.12419	1.000	-19.9634	28.8464
	10	-1.06950	6.24179	1.000	-22.7409	20.6019
	11	-9.61650	6.08568	.878	-30.8516	11.6186
9	10	-5.51100	5.61556	.995	-24.8864	13.8644
	11	-14.05800	5.44150	.299	-32.9180	4.8020
10	11	-8.54700	4.22087	.634	-22.9976	5.9036

* The mean difference is significant at the 0.05 level

Table C.11: ANOVA Post-hoc tests for the Nasa Task Load Index - Group B - 95% Confidence Interval

NASATLX - Games-Howell					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-2.38075	6.06702	1.000	-28.1325	23.3710
	3	20.13600	6.39626	.105	-6.4428	46.7148
	4	4.49050	6.51962	1.000	-22.4487	31.4297
	5	10.21050	6.93472	.919	-18.0702	38.4912
	6	.01550	6.75098	1.000	-27.6496	27.6806
	7	-4.39422	6.06090	1.000	-30.1202	21.3317
	8	-8.59750	7.81801	.989	-40.1747	22.9797
	9	-4.15600	7.32774	1.000	-33.8454	25.5334
	10	-9.66700	6.47314	.911	-36.4680	17.1340
	11	-18.21400	6.32274	.180	-44.5881	8.1601
2	3	22.51675 **	3.69566	.000	7.3844	37.6491
	4	6.87125	3.90528	.794	-9.1714	22.9139
	5	12.59125	4.56465	.224	-6.3782	31.5607
	6	2.39625	4.28037	1.000	-15.3030	20.0955
	7	-2.01347	3.07902	1.000	-14.6636	10.6367
	8	-6.21675	5.81954	.990	-30.8489	18.4154
	9	-1.77525	5.14214	1.000	-23.3447	19.7942
	10	-7.28625	3.82719	.710	-22.9881	8.4156
	11	-15.83325 **	3.56691	.004	-30.4146	-1.2519
3	4	-15.64550	4.39940	.036	-33.4219	2.1309
	5	-9.92550	4.99395	.658	-30.2416	10.3906
	6	-20.12050 **	4.73551	.006	-39.3173	-.9237
	7	-24.53022 **	3.68562	.000	-39.5774	-9.4830
	8	-28.73350 **	6.16202	.003	-54.2544	-3.2126
	9	-24.29200 **	5.52676	.005	-46.9617	-1.6223
	10	-29.80300 **	4.33022	.000	-47.2936	-12.3124
	11	-38.35000 **	4.10199	.000	-54.9192	-21.7808
4	5	5.72000	5.15100	.988	-15.1631	26.6031
	6	-4.47500	4.90085	.997	-24.2944	15.3444
	7	-8.88472	3.89578	.469	-24.8531	7.0837
	8	-13.08800	6.28998	.598	-38.9952	12.8192
	9	-8.64650	5.66907	.901	-31.7883	14.4953
	10	-14.15750	4.51045	.097	-32.3730	4.0580
	11	-22.70450 **	4.29181	.000	-40.0648	-5.3442
5	6	-10.19500	5.44088	.729	-32.1805	11.7905
	7	-14.60472	4.55653	.093	-33.5235	4.3141
	8	-18.80800	6.71927	.200	-46.1441	8.5281
	9	-14.36650	6.14192	.430	-39.2176	10.4846
	10	-19.87750	5.09205	.015	-40.5449	.7899
	11	-28.42450 **	4.89944	.000	-48.4120	-8.4370
6	7	-4.40972	4.27170	.993	-22.0497	13.2302
	8	-8.61300	6.52948	.959	-35.2950	18.0690
	9	-4.17150	5.93369	1.000	-28.2468	19.9038
	10	-9.68250	4.83885	.649	-29.2657	9.9007
	11	-18.22950	4.63573	.014	-37.0631	.6041
7	8	-4.20328	5.81317	1.000	-28.8071	20.4005
	9	.23822	5.13492	1.000	-21.2932	21.7696
	10	-5.27278	3.81749	.945	-20.8966	10.3510
	11	-13.81978	3.55651	.016	-28.3081	.6685
8	9	4.44150	7.12419	1.000	-24.3823	33.2653
	10	-1.06950	6.24179	1.000	-26.8287	24.6897
	11	-9.61650	6.08568	.878	-34.9177	15.6847
9	10	-5.51100	5.61556	.995	-28.4725	17.4505
	11	-14.05800	5.44150	.299	-36.4570	8.3410
10	11	-8.54700	4.22087	.634	-25.6095	8.5155

** The mean difference is significant at the 0.01 level

Table C.12: ANOVA Post-hoc tests for the Nasa Task Load Index - Group B - 99% Confidence Interval

WP - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-4.10100	4.63201	.998	-19.1771	10.9751
	3	-6.38850	4.63201	.952	-21.4646	8.6876
	4	-12.45050	4.63201	.213	-27.5266	2.6256
	5	-11.13100	4.63201	.370	-26.2071	3.9451
	6	-7.95750	4.63201	.825	-23.0336	7.1186
	7	-19.64350 *	4.63201	.002	-34.7196	-4.5674
	8	-10.85000	4.63201	.409	-25.9261	4.2261
	9	-16.38800 *	4.63201	.021	-31.4641	-1.3119
	10	-22.26350 *	4.63201	.000	-37.3396	-7.1874
	11	-22.80500 *	4.63201	.000	-37.8811	-7.7289
2	3	-2.28750	4.63201	1.000	-17.3636	12.7886
	4	-8.34950	4.63201	.777	-23.4256	6.7266
	5	-7.03000	4.63201	.912	-22.1061	8.0461
	6	-3.85650	4.63201	.999	-18.9326	11.2196
	7	-15.54250 *	4.63201	.037	-30.6186	-.4664
	8	-6.74900	4.63201	.932	-21.8251	8.3271
	9	-12.28700	4.63201	.229	-27.3631	2.7891
	10	-18.16250 *	4.63201	.006	-33.2386	-3.0864
	11	-18.70400 *	4.63201	.004	-33.7801	-3.6279
3	4	-6.06200	4.63201	.966	-21.1381	9.0141
	5	-4.74250	4.63201	.995	-19.8186	10.3336
	6	-1.56900	4.63201	1.000	-16.6451	13.5071
	7	-13.25500	4.63201	.143	-28.3311	1.8211
	8	-4.46150	4.63201	.997	-19.5376	10.6146
	9	-9.99950	4.63201	.537	-25.0756	5.0766
	10	-15.87500 *	4.63201	.030	-30.9511	-.7989
	11	-16.41650 *	4.63201	.020	-31.4926	-1.3404
4	5	1.31950	4.63201	1.000	-13.7566	16.3956
	6	4.49300	4.63201	.997	-10.5831	19.5691
	7	-7.19300	4.63201	.900	-22.2691	7.8831
	8	1.60050	4.63201	1.000	-13.4756	16.6766
	9	-3.93750	4.63201	.999	-19.0136	11.1386
	10	-9.81300	4.63201	.565	-24.8891	5.2631
	11	-10.35450	4.63201	.483	-25.4306	4.7216
5	6	3.17350	4.63201	1.000	-11.9026	18.2496
	7	-8.51250	4.63201	.756	-23.5886	6.5636
	8	.28100	4.63201	1.000	-14.7951	15.3571
	9	-5.25700	4.63201	.988	-20.3331	9.8191
	10	-11.13250	4.63201	.370	-26.2086	3.9436
	11	-11.67400	4.63201	.299	-26.7501	3.4021
6	7	-11.68600	4.63201	.297	-26.7621	3.3901
	8	-2.89250	4.63201	1.000	-17.9686	12.1836
	9	-8.43050	4.63201	.767	-23.5066	6.6456
	10	-14.30600	4.63201	.080	-29.3821	.7701
	11	-14.84750	4.63201	.058	-29.9236	.2286
7	8	8.79350	4.63201	.718	-6.2826	23.8696
	9	3.25550	4.63201	1.000	-11.8206	18.3316
	10	-2.62000	4.63201	1.000	-17.6961	12.4561
	11	-3.16150	4.63201	1.000	-18.2376	11.9146
8	9	-5.53800	4.63201	.982	-20.6141	9.5381
	10	-11.41350	4.63201	.332	-26.4896	3.6626
	11	-11.95500	4.63201	.266	-27.0311	3.1211
9	10	-5.87550	4.63201	.973	-20.9516	9.2006
	11	-6.41700	4.63201	.951	-21.4931	8.6591
10	11	-.54150	4.63201	1.000	-15.6176	14.5346

* The mean difference is significant at the 0.05 level

Table C.13: ANOVA Post-hoc tests for Workload Profile instrument - Group A - 95% Confidence Interval

WP - Tukey HSD					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-4.10100	4.63201	.998	-21.4971	13.2951
	3	-6.38850	4.63201	.952	-23.7846	11.0076
	4	-12.45050	4.63201	.213	-29.8466	4.9456
	5	-11.13100	4.63201	.370	-28.5271	6.2651
	6	-7.95750	4.63201	.825	-25.3536	9.4386
	7	-19.64350 **	4.63201	.002	-37.0396	-2.2474
	8	-10.85000	4.63201	.409	-28.2461	6.5461
	9	-16.38800	4.63201	.021	-33.7841	1.0081
	10	-22.26350 **	4.63201	.000	-39.6596	-4.8674
	11	-22.80500 **	4.63201	.000	-40.2011	-5.4089
2	3	-2.28750	4.63201	1.000	-19.6836	15.1086
	4	-8.34950	4.63201	.777	-25.7456	9.0466
	5	-7.03000	4.63201	.912	-24.4261	10.3661
	6	-3.85650	4.63201	.999	-21.2526	13.5396
	7	-15.54250	4.63201	.037	-32.9386	1.8536
	8	-6.74900	4.63201	.932	-24.1451	10.6471
	9	-12.28700	4.63201	.229	-29.6831	5.1091
	10	-18.16250 **	4.63201	.006	-35.5586	-.7664
	11	-18.70400 **	4.63201	.004	-36.1001	-1.3079
3	4	-6.06200	4.63201	.966	-23.4581	11.3341
	5	-4.74250	4.63201	.995	-22.1386	12.6536
	6	-1.56900	4.63201	1.000	-18.9651	15.8271
	7	-13.25500	4.63201	.143	-30.6511	4.1411
	8	-4.46150	4.63201	.997	-21.8576	12.9346
	9	-9.99950	4.63201	.537	-27.3956	7.3966
	10	-15.87500	4.63201	.030	-33.2711	1.5211
	11	-16.41650	4.63201	.020	-33.8126	.9796
4	5	1.31950	4.63201	1.000	-16.0766	18.7156
	6	4.49300	4.63201	.997	-12.9031	21.8891
	7	-7.19300	4.63201	.900	-24.5891	10.2031
	8	1.60050	4.63201	1.000	-15.7956	18.9966
	9	-3.93750	4.63201	.999	-21.3336	13.4586
	10	-9.81300	4.63201	.565	-27.2091	7.5831
	11	-10.35450	4.63201	.483	-27.7506	7.0416
5	6	3.17350	4.63201	1.000	-14.2226	20.5696
	7	-8.51250	4.63201	.756	-25.9086	8.8836
	8	.28100	4.63201	1.000	-17.1151	17.6771
	9	-5.25700	4.63201	.988	-22.6531	12.1391
	10	-11.13250	4.63201	.370	-28.5286	6.2636
	11	-11.67400	4.63201	.299	-29.0701	5.7221
6	7	-11.68600	4.63201	.297	-29.0821	5.7101
	8	-2.89250	4.63201	1.000	-20.2886	14.5036
	9	-8.43050	4.63201	.767	-25.8266	8.9656
	10	-14.30600	4.63201	.080	-31.7021	3.0901
	11	-14.84750	4.63201	.058	-32.2436	2.5486
7	8	8.79350	4.63201	.718	-8.6026	26.1896
	9	3.25550	4.63201	1.000	-14.1406	20.6516
	10	-2.62000	4.63201	1.000	-20.0161	14.7761
	11	-3.16150	4.63201	1.000	-20.5576	14.2346
8	9	-5.53800	4.63201	.982	-22.9341	11.8581
	10	-11.41350	4.63201	.332	-28.8096	5.9826
	11	-11.95500	4.63201	.266	-29.3511	5.4411
9	10	-5.87550	4.63201	.973	-23.2716	11.5206
	11	-6.41700	4.63201	.951	-23.8131	10.9791
10	11	-.54150	4.63201	1.000	-17.9376	16.8546

** The mean difference is significant at the 0.01 level

Table C.14: ANOVA Post-hoc tests for Workload Profile instrument - Group A - 99% Confidence Interval

WP - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	10.07278	3.86761	.253	-2.5188	22.6644
	3	7.81800	3.76446	.595	-4.4377	20.0737
	4	.84350	3.76446	1.000	-11.4122	13.0992
	5	2.89400	3.76446	1.000	-9.3617	15.1497
	6	.20395	3.81367	1.000	-12.2120	12.6199
	7	-5.75750	3.76446	.908	-18.0132	6.4982
	8	.63000	3.76446	1.000	-11.6257	12.8857
	9	-2.90250	3.76446	1.000	-15.1582	9.3532
	10	-11.55079	3.81367	.094	-23.9668	.8652
	11	-8.82711	3.81367	.428	-21.2431	3.5889
2	3	-2.25478	3.86761	1.000	-14.8464	10.3368
	4	-9.22928	3.86761	.381	-21.8209	3.3623
	5	-7.17878	3.86761	.745	-19.7704	5.4128
	6	-9.86883	3.91553	.299	-22.6164	2.8787
	7	-15.83028 *	3.86761	.003	-28.4219	-3.2387
	8	-9.44278	3.86761	.346	-22.0344	3.1488
	9	-12.97528 *	3.86761	.037	-25.5669	-.3837
	10	-21.62357 *	3.91553	.000	-34.3711	-8.8760
	11	-18.89988 *	3.91553	.000	-31.6475	-6.1523
3	4	-6.97450	3.76446	.747	-19.2302	5.2812
	5	-4.92400	3.76446	.967	-17.1797	7.3317
	6	-7.61405	3.81367	.652	-20.0300	4.8019
	7	-13.57550 *	3.76446	.017	-25.8312	-1.3198
	8	-7.18800	3.76446	.710	-19.4437	5.0677
	9	-10.72050	3.76446	.148	-22.9762	1.5352
	10	-19.36879 *	3.81367	.000	-31.7848	-6.9528
	11	-16.64511 *	3.81367	.001	-29.0611	-4.2291
4	5	2.05050	3.76446	1.000	-10.2052	14.3062
	6	-.63955	3.81367	1.000	-13.0555	11.7764
	7	-6.60100	3.76446	.806	-18.8567	5.6547
	8	-.21350	3.76446	1.000	-12.4692	12.0422
	9	-3.74600	3.76446	.996	-16.0017	8.5097
	10	-12.39429	3.81367	.051	-24.8103	.0217
	11	-9.67061	3.81367	.290	-22.0866	2.7454
5	6	-2.69005	3.81367	1.000	-15.1060	9.7259
	7	-8.65150	3.76446	.439	-20.9072	3.6042
	8	-2.26400	3.76446	1.000	-14.5197	9.9917
	9	-5.79650	3.76446	.905	-18.0522	6.4592
	10	-14.44479 *	3.81367	.009	-26.8608	-2.0288
	11	-11.72111	3.81367	.083	-24.1371	.6949
6	7	-5.96145	3.81367	.896	-18.3774	6.4545
	8	.42605	3.81367	1.000	-11.9899	12.8420
	9	-3.10645	3.81367	.999	-15.5224	9.3095
	10	-11.75474	3.86225	.090	-24.3289	.8194
	11	-9.03105	3.86225	.412	-21.6052	3.5431
7	8	6.38750	3.76446	.836	-5.8682	18.6432
	9	2.85500	3.76446	1.000	-9.4007	15.1107
	10	-5.79329	3.81367	.912	-18.2093	6.6227
	11	-3.06961	3.81367	.999	-15.4856	9.3464
8	9	-3.53250	3.76446	.997	-15.7882	8.7232
	10	-12.18079	3.81367	.060	-24.5968	.2352
	11	-9.45711	3.81367	.323	-21.8731	2.9589
9	10	-8.64829	3.81367	.460	-21.0643	3.7677
	11	-5.92461	3.81367	.899	-18.3406	6.4914
10	11	2.72368	3.86225	1.000	-9.8504	15.2978

* The mean difference is significant at the 0.05 level

Table C.15: ANOVA Post-hoc tests for Workload Profile instrument - Group B - 95% Confidence Interval

WP - Tukey HSD					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	10.07278	3.86761	.253	-4.4582	24.6037
	3	7.81800	3.76446	.595	-6.3254	21.9614
	4	.84350	3.76446	1.000	-13.2999	14.9869
	5	2.89400	3.76446	1.000	-11.2494	17.0374
	6	.20395	3.81367	1.000	-14.1243	14.5322
	7	-5.75750	3.76446	.908	-19.9009	8.3859
	8	.63000	3.76446	1.000	-13.5134	14.7734
	9	-2.90250	3.76446	1.000	-17.0459	11.2409
	10	-11.55079	3.81367	.094	-25.8791	2.7775
	11	-8.82711	3.81367	.428	-23.1554	5.5012
2	3	-2.25478	3.86761	1.000	-16.7857	12.2762
	4	-9.22928	3.86761	.381	-23.7602	5.3017
	5	-7.17878	3.86761	.745	-21.7097	7.3522
	6	-9.86883	3.91553	.299	-24.5798	4.8422
	7	-15.83028 **	3.86761	.003	-30.3612	-1.2993
	8	-9.44278	3.86761	.346	-23.9737	5.0882
	9	-12.97528	3.86761	.037	-27.5062	1.5557
	10	-21.62357 **	3.91553	.000	-36.3346	-6.9126
	11	-18.89988 **	3.91553	.000	-33.6109	-4.1889
3	4	-6.97450	3.76446	.747	-21.1179	7.1689
	5	-4.92400	3.76446	.967	-19.0674	9.2194
	6	-7.61405	3.81367	.652	-21.9423	6.7142
	7	-13.57550	3.76446	.017	-27.7189	.5679
	8	-7.18800	3.76446	.710	-21.3314	6.9554
	9	-10.72050	3.76446	.148	-24.8639	3.4229
	10	-19.36879 **	3.81367	.000	-33.6971	-5.0405
	11	-16.64511 **	3.81367	.001	-30.9734	-2.3168
4	5	2.05050	3.76446	1.000	-12.0929	16.1939
	6	-.63955	3.81367	1.000	-14.9678	13.6887
	7	-6.60100	3.76446	.806	-20.7444	7.5424
	8	-.21350	3.76446	1.000	-14.3569	13.9299
	9	-3.74600	3.76446	.996	-17.8894	10.3974
	10	-12.39429	3.81367	.051	-26.7226	1.9340
	11	-9.67061	3.81367	.290	-23.9989	4.6577
5	6	-2.69005	3.81367	1.000	-17.0183	11.6382
	7	-8.65150	3.76446	.439	-22.7949	5.4919
	8	-2.26400	3.76446	1.000	-16.4074	11.8794
	9	-5.79650	3.76446	.905	-19.9399	8.3469
	10	-14.44479 **	3.81367	.009	-28.7731	-.1165
	11	-11.72111	3.81367	.083	-26.0494	2.6072
6	7	-5.96145	3.81367	.896	-20.2897	8.3668
	8	.42605	3.81367	1.000	-13.9022	14.7543
	9	-3.10645	3.81367	.999	-17.4347	11.2218
	10	-11.75474	3.86225	.090	-26.2656	2.7561
	11	-9.03105	3.86225	.412	-23.5419	5.4798
7	8	6.38750	3.76446	.836	-7.7559	20.5309
	9	2.85500	3.76446	1.000	-11.2884	16.9984
	10	-5.79329	3.81367	.912	-20.1216	8.5350
	11	-3.06961	3.81367	.999	-17.3979	11.2587
8	9	-3.53250	3.76446	.997	-17.6759	10.6109
	10	-12.18079	3.81367	.060	-26.5091	2.1475
	11	-9.45711	3.81367	.323	-23.7854	4.8712
9	10	-8.64829	3.81367	.460	-22.9766	5.6800
	11	-5.92461	3.81367	.899	-20.2529	8.4037
10	11	2.72368	3.86225	1.000	-11.7871	17.2345

** The mean difference is significant at the 0.01 level

Table C.16: ANOVA Post-hoc tests for Workload Profile instrument - Group B - 99% Confidence Interval

MWL_{Def}^{NI} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-11.05050	4.15649	.226	-24.5803	2.4793
	3	-9.33650	4.15649	.475	-22.8663	4.1933
	4	-14.07750 *	4.15649	.034	-27.6073	-.5477
	5	-13.08283	4.27038	.086	-26.9834	.8177
	6	-14.26150 *	4.15649	.029	-27.7913	-.7317
	7	-18.88600 *	4.15649	.000	-32.4158	-5.3562
	8	-18.96950 *	4.15649	.000	-32.4993	-5.4397
	9	-26.70300 *	4.15649	.000	-40.2328	-13.1732
	10	-31.65400 *	4.15649	.000	-45.1838	-18.1242
	11	-32.73000 *	4.15649	.000	-46.2598	-19.2002
2	3	1.71400	4.15649	1.000	-11.8158	15.2438
	4	-3.02700	4.15649	1.000	-16.5568	10.5028
	5	-2.03233	4.27038	1.000	-15.9329	11.8682
	6	-3.21100	4.15649	1.000	-16.7408	10.3188
	7	-7.83550	4.15649	.726	-21.3653	5.6943
	8	-7.91900	4.15649	.713	-21.4488	5.6108
	9	-15.65250 *	4.15649	.010	-29.1823	-2.1227
	10	-20.60350 *	4.15649	.000	-34.1333	-7.0737
	11	-21.67950 *	4.15649	.000	-35.2093	-8.1497
3	4	-4.74100	4.15649	.988	-18.2708	8.7888
	5	-3.74633	4.27038	.999	-17.6469	10.1542
	6	-4.92500	4.15649	.984	-18.4548	8.6048
	7	-9.54950	4.15649	.440	-23.0793	3.9803
	8	-9.63300	4.15649	.426	-23.1628	3.8968
	9	-17.36650 *	4.15649	.002	-30.8963	-3.8367
	10	-22.31750 *	4.15649	.000	-35.8473	-8.7877
	11	-23.39350 *	4.15649	.000	-36.9233	-9.8637
4	5	.99467	4.27038	1.000	-12.9059	14.8952
	6	-.18400	4.15649	1.000	-13.7138	13.3458
	7	-4.80850	4.15649	.986	-18.3383	8.7213
	8	-4.89200	4.15649	.984	-18.4218	8.6378
	9	-12.62550	4.15649	.092	-26.1553	.9043
	10	-17.57650 *	4.15649	.002	-31.1063	-4.0467
	11	-18.65250 *	4.15649	.001	-32.1823	-5.1227
5	6	-1.17867	4.27038	1.000	-15.0792	12.7219
	7	-5.80317	4.27038	.957	-19.7037	8.0974
	8	-5.88667	4.27038	.952	-19.7872	8.0139
	9	-13.62017	4.27038	.060	-27.5207	.2804
	10	-18.57117 *	4.27038	.001	-32.4717	-4.6706
	11	-19.64717 *	4.27038	.000	-33.5477	-5.7466
6	7	-4.62450	4.15649	.990	-18.1543	8.9053
	8	-4.70800	4.15649	.988	-18.2378	8.8218
	9	-12.44150	4.15649	.103	-25.9713	1.0883
	10	-17.39250 *	4.15649	.002	-30.9223	-3.8627
	11	-18.46850 *	4.15649	.001	-31.9983	-4.9387
7	8	-.08350	4.15649	1.000	-13.6133	13.4463
	9	-7.81700	4.15649	.729	-21.3468	5.7128
	10	-12.76800	4.15649	.084	-26.2978	.7618
	11	-13.84400 *	4.15649	.040	-27.3738	-.3142
8	9	-7.73350	4.15649	.742	-21.2633	5.7963
	10	-12.68450	4.15649	.088	-26.2143	.8453
	11	-13.76050 *	4.15649	.042	-27.2903	-.2307
9	10	-4.95100	4.15649	.983	-18.4808	8.5788
	11	-6.02700	4.15649	.934	-19.5568	7.5028
10	11	-1.07600	4.15649	1.000	-14.6058	12.4538

* The mean difference is significant at the 0.05 level

Table C.17: ANOVA Post-hoc tests for the instance MWL_{def}^{NI} - Group A - 95% Confidence Interval

MWL^{NI}_{Def} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-11.05050	4.15649	.226	-26.6631	4.5621
	3	-9.33650	4.15649	.475	-24.9491	6.2761
	4	-14.07750	4.15649	.034	-29.6901	1.5351
	5	-13.08283	4.27038	.086	-29.1232	2.9576
	6	-14.26150	4.15649	.029	-29.8741	1.3511
	7	-18.88600 *	4.15649	.000	-34.4986	-3.2734
	8	-18.96950 *	4.15649	.000	-34.5821	-3.3569
	9	-26.70300 *	4.15649	.000	-42.3156	-11.0904
	10	-31.65400 *	4.15649	.000	-47.2666	-16.0414
	11	-32.73000 *	4.15649	.000	-48.3426	-17.1174
2	3	1.71400	4.15649	1.000	-13.8986	17.3266
	4	-3.02700	4.15649	1.000	-18.6396	12.5856
	5	-2.03233	4.27038	1.000	-18.0727	14.0081
	6	-3.21100	4.15649	1.000	-18.8236	12.4016
	7	-7.83550	4.15649	.726	-23.4481	7.7771
	8	-7.91900	4.15649	.713	-23.5316	7.6936
	9	-15.65250 *	4.15649	.010	-31.2651	-.0399
	10	-20.60350 *	4.15649	.000	-36.2161	-4.9909
	11	-21.67950 *	4.15649	.000	-37.2921	-6.0669
3	4	-4.74100	4.15649	.988	-20.3536	10.8716
	5	-3.74633	4.27038	.999	-19.7867	12.2941
	6	-4.92500	4.15649	.984	-20.5376	10.6876
	7	-9.54950	4.15649	.440	-25.1621	6.0631
	8	-9.63300	4.15649	.426	-25.2456	5.9796
	9	-17.36650 *	4.15649	.002	-32.9791	-1.7539
	10	-22.31750 *	4.15649	.000	-37.9301	-6.7049
	11	-23.39350 *	4.15649	.000	-39.0061	-7.7809
4	5	.99467	4.27038	1.000	-15.0457	17.0351
	6	-.18400	4.15649	1.000	-15.7966	15.4286
	7	-4.80850	4.15649	.986	-20.4211	10.8041
	8	-4.89200	4.15649	.984	-20.5046	10.7206
	9	-12.62550	4.15649	.092	-28.2381	2.9871
	10	-17.57650 *	4.15649	.002	-33.1891	-1.9639
	11	-18.65250 *	4.15649	.001	-34.2651	-3.0399
5	6	-1.17867	4.27038	1.000	-17.2191	14.8617
	7	-5.80317	4.27038	.957	-21.8436	10.2372
	8	-5.88667	4.27038	.952	-21.9271	10.1537
	9	-13.62017	4.27038	.060	-29.6606	2.4202
	10	-18.57117 *	4.27038	.001	-34.6116	-2.5308
	11	-19.64717 *	4.27038	.000	-35.6876	-3.6068
6	7	-4.62450	4.15649	.990	-20.2371	10.9881
	8	-4.70800	4.15649	.988	-20.3206	10.9046
	9	-12.44150	4.15649	.103	-28.0541	3.1711
	10	-17.39250 *	4.15649	.002	-33.0051	-1.7799
	11	-18.46850 *	4.15649	.001	-34.0811	-2.8559
7	8	-.08350	4.15649	1.000	-15.6961	15.5291
	9	-7.81700	4.15649	.729	-23.4296	7.7956
	10	-12.76800	4.15649	.084	-28.3806	2.8446
	11	-13.84400	4.15649	.040	-29.4566	1.7686
8	9	-7.73350	4.15649	.742	-23.3461	7.8791
	10	-12.68450	4.15649	.088	-28.2971	2.9281
	11	-13.76050	4.15649	.042	-29.3731	1.8521
9	10	-4.95100	4.15649	.983	-20.5636	10.6616
	11	-6.02700	4.15649	.934	-21.6396	9.5856
10	11	-1.07600	4.15649	1.000	-16.6886	14.5366

** The mean difference is significant at the 0.01 level

Table C.18: ANOVA Post-hoc tests for the instance MWL^{NI}_{def} - Group A - 99% Confidence Interval

MWL^{NI}_{Def} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	6.87050	4.69217	.930	-8.4064	22.1474
	3	-.03400	4.56702	1.000	-14.9035	14.8355
	4	2.39800	4.56702	1.000	-12.4715	17.2675
	5	8.12900	4.56702	.791	-6.7405	22.9985
	6	4.31650	4.56702	.997	-10.5530	19.1860
	7	-6.19350	4.56702	.957	-21.0630	8.6760
	8	-13.43476	4.62672	.129	-28.4986	1.6291
	9	-6.59244	4.76425	.951	-22.1040	8.9192
	10	-15.98200 *	4.56702	.024	-30.8515	-1.1125
	11	-7.37100	4.56702	.875	-22.2405	7.4985
2	3	-6.90450	4.69217	.928	-22.1814	8.3724
	4	-4.47250	4.69217	.997	-19.7494	10.8044
	5	1.25850	4.69217	1.000	-14.0184	16.5354
	6	-2.55400	4.69217	1.000	-17.8309	12.7229
	7	-13.06400	4.69217	.172	-28.3409	2.2129
	8	-20.30526 *	4.75030	.001	-35.7714	-4.8391
	9	-13.46294	4.88434	.183	-29.3655	2.4397
	10	-22.85250 *	4.69217	.000	-38.1294	-7.5756
	11	-14.24150	4.69217	.092	-29.5184	1.0354
3	4	2.43200	4.56702	1.000	-12.4375	17.3015
	5	8.16300	4.56702	.786	-6.7065	23.0325
	6	4.35050	4.56702	.997	-10.5190	19.2200
	7	-6.15950	4.56702	.959	-21.0290	8.7100
	8	-13.40076	4.62672	.132	-28.4646	1.6631
	9	-6.55844	4.76425	.953	-22.0700	8.9532
	10	-15.94800 *	4.56702	.024	-30.8175	-1.0785
	11	-7.33700	4.56702	.878	-22.2065	7.5325
4	5	5.73100	4.56702	.975	-9.1385	20.6005
	6	1.91850	4.56702	1.000	-12.9510	16.7880
	7	-8.59150	4.56702	.729	-23.4610	6.2780
	8	-15.83276 *	4.62672	.030	-30.8966	-.7689
	9	-8.99044	4.76425	.725	-24.5020	6.5212
	10	-18.38000 *	4.56702	.004	-33.2495	-3.5105
	11	-9.76900	4.56702	.551	-24.6385	5.1005
5	6	-3.81250	4.56702	.999	-18.6820	11.0570
	7	-14.32250	4.56702	.070	-29.1920	.5470
	8	-21.56376 *	4.62672	.000	-36.6276	-6.4999
	9	-14.72144	4.76425	.080	-30.2330	.7902
	10	-24.11100 *	4.56702	.000	-38.9805	-9.2415
	11	-15.50000 *	4.56702	.033	-30.3695	-.6305
6	7	-10.51000	4.56702	.437	-25.3795	4.3595
	8	-17.75126 *	4.62672	.008	-32.8151	-2.6874
	9	-10.90894	4.76425	.445	-26.4205	4.6027
	10	-20.29850 *	4.56702	.001	-35.1680	-5.4290
	11	-11.68750	4.56702	.278	-26.5570	3.1820
7	8	-7.24126	4.62672	.895	-22.3051	7.8226
	9	-.39894	4.76425	1.000	-15.9105	15.1127
	10	-9.78850	4.56702	.548	-24.6580	5.0810
	11	-1.17750	4.56702	1.000	-16.0470	13.6920
8	9	6.84232	4.82151	.942	-8.8557	22.5403
	10	-2.54724	4.62672	1.000	-17.6111	12.5166
	11	6.06376	4.62672	.966	-9.0001	21.1276
9	10	-9.38956	4.76425	.669	-24.9012	6.1220
	11	-.77856	4.76425	1.000	-16.2902	14.7330
10	11	8.61100	4.56702	.726	-6.2585	23.4805

* The mean difference is significant at the 0.05 level

Table C.19: ANOVA Post-hoc tests for the instance MWL^{NI}_{def} - Group B - 95% Confidence Interval

MWL^{NI}_{Def} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	6.87050	4.692 6	.930	-10.7598	24.5008
	3	-.03400	4.56702	1.000	-17.1941	17.1261
	4	2.39800	4.56702	1.000	-14.76 10	19.5581
	5	8.12900	4.56702	.791	-9.0311	25.2891
	6	4.31650	4.56702	.997	-12.8436	21.4766
	7	-6.19350	4.56702	.957	-23.3536	10.9666
	8	-13.43476	4.62672	.129	-30.8192	3.9497
	9	-6.59244	4.76425	.951	-24.4936	11.3087
	10	-15.98200	4.56702	.024	-33.14 10	1.1781
	11	-7.37100	4.56702	.875	-24.5311	9.7891
2	3	-6.90450	4.692 6	.928	-24.5348	10.7258
	4	-4.47250	4.692 6	.997	-22.1028	13.1578
	5	1.25850	4.692 6	1.000	-16.37 7	18.8888
	6	-2.55400	4.692 6	1.000	-20.1843	15.0763
	7	-13.06400	4.692 6	.172	-30.6943	4.5663
	8	-20.30526 *	4.75030	.001	-38.1540	-2.4565
	9	-13.46294	4.88434	.183	-31.8154	4.8895
	10	-22.85250 *	4.692 6	.000	-40.4828	-5.2222
	11	-14.24150	4.692 6	.092	-31.87 7	3.3888
3	4	2.43200	4.56702	1.000	-14.7281	19.5921
	5	8.16300	4.56702	.786	-8.9971	25.3231
	6	4.35050	4.56702	.997	-12.8096	21.5106
	7	-6.15950	4.56702	.959	-23.3196	11.0006
	8	-13.40076	4.62672	.132	-30.7852	3.9837
	9	-6.55844	4.76425	.953	-24.4596	11.3427
	10	-15.94800	4.56702	.024	-33.1081	1.2121
	11	-7.33700	4.56702	.878	-24.4971	9.8231
4	5	5.73100	4.56702	.975	-11.4291	22.8911
	6	1.91850	4.56702	1.000	-15.24 5	19.0786
	7	-8.59150	4.56702	.729	-25.75 5	8.5686
	8	-15.83276	4.62672	.030	-33.2172	1.5517
	9	-8.99044	4.76425	.725	-26.89 5	8.9107
	10	-18.38000 *	4.56702	.004	-35.5401	-1.2199
	11	-9.76900	4.56702	.551	-26.9291	7.3911
5	6	-3.81250	4.56702	.999	-20.9726	13.3476
	7	-14.32250	4.56702	.070	-31.4826	2.8376
	8	-21.56376 *	4.62672	.000	-38.9482	-4.1793
	9	-14.72144	4.76425	.080	-32.6226	3.1797
	10	-24.11100 *	4.56702	.000	-41.2711	-6.9509
	11	-15.50000	4.56702	.033	-32.6601	1.6601
6	7	-10.51000	4.56702	.437	-27.6701	6.6501
	8	-17.75126 *	4.62672	.008	-35.1357	-.3668
	9	-10.90894	4.76425	.445	-28.8101	6.9922
	10	-20.29850 *	4.56702	.001	-37.4586	-3.1384
	11	-11.68750	4.56702	.278	-28.8476	5.4726
7	8	-7.24126	4.62672	.895	-24.6257	10.1432
	9	-.39894	4.76425	1.000	-18.3001	17.5022
	10	-9.78850	4.56702	.548	-26.9486	7.3716
	11	-1.17750	4.56702	1.000	-18.3376	15.9826
8	9	6.84232	4.82151	.942	-11.2740	24.9586
	10	-2.54724	4.62672	1.000	-19.93 6	14.8372
	11	6.06376	4.62672	.966	-11.3207	23.4482
9	10	-9.38956	4.76425	.669	-27.2907	8.5116
	11	-.77856	4.76425	1.000	-18.6797	17.1226
10	11	8.61100	4.56702	.726	-8.5491	25.7711

** The mean difference is significant at the 0.01 level

Table C.20: ANOVA Post-hoc tests for the instance MWL^{NI}_{def} - Group B - 99% Confidence Interval

MWL_{Def} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-11.17616	4.46065	.308	-25.6961	3.3437
	3	-9.89016	4.46065	.495	-24.4101	4.6297
	4	-15.01066 *	4.46065	.036	-29.5306	-.4908
	5	-11.91421	4.51747	.237	-26.6191	2.7907
	6	-15.84266 *	4.46065	.020	-30.3626	-1.3228
	7	-22.21916 *	4.46065	.000	-36.7391	-7.6993
	8	-20.96366 *	4.46065	.000	-35.4836	-6.4438
	9	-29.64866 *	4.46065	.000	-44.1686	-15.1288
	10	-34.18666 *	4.46065	.000	-48.7066	-19.6668
	11	-35.08516 *	4.46065	.000	-49.6051	-20.5653
2	3	1.28600	4.40309	1.000	-13.0465	15.6185
	4	-3.83450	4.40309	.999	-18.1670	10.4980
	5	-.73805	4.46065	1.000	-15.2580	13.7818
	6	-4.66650	4.40309	.993	-18.9990	9.6660
	7	-11.04300	4.40309	.306	-25.3755	3.2895
	8	-9.78750	4.40309	.491	-24.1200	4.5450
	9	-18.47250 *	4.40309	.002	-32.8050	-4.1400
	10	-23.01050 *	4.40309	.000	-37.3430	-8.6780
	11	-23.90900 *	4.40309	.000	-38.2415	-9.5765
3	4	-5.12050	4.40309	.986	-19.4530	9.2120
	5	-2.02405	4.46065	1.000	-16.5440	12.4958
	6	-5.95250	4.40309	.958	-20.2850	8.3800
	7	-12.32900	4.40309	.165	-26.6615	2.0035
	8	-11.07350	4.40309	.302	-25.4060	3.2590
	9	-19.75850 *	4.40309	.001	-34.0910	-5.4260
	10	-24.29650 *	4.40309	.000	-38.6290	-9.9640
	11	-25.19500 *	4.40309	.000	-39.5275	-10.8625
4	5	3.09645	4.46065	1.000	-11.4235	17.6163
	6	-.83200	4.40309	1.000	-15.1645	13.5005
	7	-7.20850	4.40309	.864	-21.5410	7.1240
	8	-5.95300	4.40309	.958	-20.2855	8.3795
	9	-14.63800 *	4.40309	.041	-28.9705	-.3055
	10	-19.17600 *	4.40309	.001	-33.5085	-4.8435
	11	-20.07450 *	4.40309	.000	-34.4070	-5.7420
5	6	-3.92845	4.46065	.998	-18.4483	10.5915
	7	-10.30495	4.46065	.431	-24.8248	4.2150
	8	-9.04945	4.46065	.629	-23.5693	5.4705
	9	-17.73445 *	4.46065	.005	-32.2543	-3.2145
	10	-22.27245 *	4.46065	.000	-36.7923	-7.7525
	11	-23.17095 *	4.46065	.000	-37.6908	-8.6510
6	7	-6.37650	4.40309	.934	-20.7090	7.9560
	8	-5.12100	4.40309	.986	-19.4535	9.2115
	9	-13.80600	4.40309	.070	-28.1385	.5265
	10	-18.34400 *	4.40309	.002	-32.6765	-4.0115
	11	-19.24250 *	4.40309	.001	-33.5750	-4.9100
7	8	1.25550	4.40309	1.000	-13.0770	15.5880
	9	-7.42950	4.40309	.840	-21.7620	6.9030
	10	-11.96750	4.40309	.199	-26.3000	2.3650
	11	-12.86600	4.40309	.123	-27.1985	1.4665
8	9	-8.68500	4.40309	.668	-23.0175	5.6475
	10	-13.22300	4.40309	.100	-27.5555	1.1095
	11	-14.12150	4.40309	.057	-28.4540	.2110
9	10	-4.53800	4.40309	.994	-18.8705	9.7945
	11	-5.43650	4.40309	.978	-19.7690	8.8960
10	11	-.89850	4.40309	1.000	-15.2310	13.4340

* The mean difference is significant at the 0.05 level

Table C.21: ANOVA Post-hoc tests for the instance MWL_{def} of the defeasible framework - Group A - 95% Confidence Interval

MWL_{Def} - Tukey HSD					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	-11.17616	4.46065	.308	-27.9312	5.5789
	3	-9.89016	4.46065	.495	-26.6452	6.8649
	4	-15.01066	4.46065	.036	-31.7657	1.7444
	5	-11.91421	4.51747	.237	-28.8827	5.0543
	6	-15.84266	4.46065	.020	-32.5977	.9124
	7	-22.21916 **	4.46065	.000	-38.9742	-5.4641
	8	-20.96366 **	4.46065	.000	-37.7187	-4.2086
	9	-29.64866 **	4.46065	.000	-46.4037	-12.8936
	10	-34.18666 **	4.46065	.000	-50.9417	-17.4316
	11	-35.08516 **	4.46065	.000	-51.8402	-18.3301
2	3	1.28600	4.40309	1.000	-15.2529	17.8249
	4	-3.83450	4.40309	.999	-20.3734	12.7044
	5	-.73805	4.46065	1.000	-17.4931	16.0170
	6	-4.66650	4.40309	.993	-21.2054	11.8724
	7	-11.04300	4.40309	.306	-27.5819	5.4959
	8	-9.78750	4.40309	.491	-26.3264	6.7514
	9	-18.47250 **	4.40309	.002	-35.0114	-1.9336
	10	-23.01050 **	4.40309	.000	-39.5494	-6.4716
	11	-23.90900 **	4.40309	.000	-40.4479	-7.3701
3	4	-5.12050	4.40309	.986	-21.6594	11.4184
	5	-2.02405	4.46065	1.000	-18.7791	14.7310
	6	-5.95250	4.40309	.958	-22.4914	10.5864
	7	-12.32900	4.40309	.165	-28.8679	4.2099
	8	-11.07350	4.40309	.302	-27.6124	5.4654
	9	-19.75850 **	4.40309	.001	-36.2974	-3.2196
	10	-24.29650 **	4.40309	.000	-40.8354	-7.7576
	11	-25.19500 **	4.40309	.000	-41.7339	-8.6561
4	5	3.09645	4.46065	1.000	-13.6586	19.8515
	6	-.83200	4.40309	1.000	-17.3709	15.7069
	7	-7.20850	4.40309	.864	-23.7474	9.3304
	8	-5.95300	4.40309	.958	-22.4919	10.5859
	9	-14.63800	4.40309	.041	-31.1769	1.9009
	10	-19.17600 **	4.40309	.001	-35.7149	-2.6371
	11	-20.07450 **	4.40309	.000	-36.6134	-3.5356
5	6	-3.92845	4.46065	.998	-20.6835	12.8266
	7	-10.30495	4.46065	.431	-27.0600	6.4501
	8	-9.04945	4.46065	.629	-25.8045	7.7056
	9	-17.73445 **	4.46065	.005	-34.4895	-.9794
	10	-22.27245 **	4.46065	.000	-39.0275	-5.5174
	11	-23.17095 **	4.46065	.000	-39.9260	-6.4159
6	7	-6.37650	4.40309	.934	-22.9154	10.1624
	8	-5.12100	4.40309	.986	-21.6599	11.4179
	9	-13.80600	4.40309	.070	-30.3449	2.7329
	10	-18.34400 **	4.40309	.002	-34.8829	-1.8051
	11	-19.24250 **	4.40309	.001	-35.7814	-2.7036
7	8	1.25550	4.40309	1.000	-15.2834	17.7944
	9	-7.42950	4.40309	.840	-23.9684	9.1094
	10	-11.96750	4.40309	.199	-28.5064	4.5714
	11	-12.86600	4.40309	.123	-29.4049	3.6729
8	9	-8.68500	4.40309	.668	-25.2239	7.8539
	10	-13.22300	4.40309	.100	-29.7619	3.3159
	11	-14.12150	4.40309	.057	-30.6604	2.4174
9	10	-4.53800	4.40309	.994	-21.0769	12.0009
	11	-5.43650	4.40309	.978	-21.9754	11.1024
10	11	-.89850	4.40309	1.000	-17.4374	15.6404

** The mean difference is significant at the 0.01 level

Table C.22: ANOVA Post-hoc tests for the instance MWL_{def} of the defeasible framework - Group A - 99% Confidence Interval

MWL_{Def} - Tukey HSD					95% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	13.95350	4.31563	.053	-.0959	28.0029
	3	14.21050 *	4.31563	.045	.1611	28.2599
	4	6.55300	4.31563	.912	-7.4964	20.6024
	5	8.12550	4.31563	.728	-5.9239	22.1749
	6	5.30066	4.37205	.981	-8.9324	19.5337
	7	-3.53972	4.43389	.999	-17.9741	10.8947
	8	.15329	4.37205	1.000	-14.0798	14.3864
	9	-7.08700	4.31563	.862	-21.1364	6.9624
	10	-14.93300 *	4.31563	.027	-28.9824	-.8836
	11	-10.85700	4.31563	.302	-24.9064	3.1924
2	3	.25700	4.31563	1.000	-13.7924	14.3064
	4	-7.40050	4.31563	.826	-21.4499	6.6489
	5	-5.82800	4.31563	.958	-19.8774	8.2214
	6	-8.65284	4.37205	.664	-22.8859	5.5802
	7	-17.49322 *	4.43389	.005	-31.9276	-3.0588
	8	-13.80021	4.37205	.066	-28.0333	.4329
	9	-21.04050 *	4.31563	.000	-35.0899	-6.9911
	10	-28.88650 *	4.31563	.000	-42.9359	-14.8371
	11	-24.81050 *	4.31563	.000	-38.8599	-10.7611
3	4	-7.65750	4.31563	.794	-21.7069	6.3919
	5	-6.08500	4.31563	.945	-20.1344	7.9644
	6	-8.90984	4.37205	.623	-23.1429	5.3232
	7	-17.75022 *	4.43389	.004	-32.1846	-3.3158
	8	-14.05721	4.37205	.056	-28.2903	.1759
	9	-21.29750 *	4.31563	.000	-35.3469	-7.2481
	10	-29.14350 *	4.31563	.000	-43.1929	-15.0941
	11	-25.06750 *	4.31563	.000	-39.1169	-11.0181
4	5	1.57250	4.31563	1.000	-12.4769	15.6219
	6	-1.25234	4.37205	1.000	-15.4854	12.9807
	7	-10.09272	4.43389	.454	-24.5271	4.3417
	8	-6.39971	4.37205	.930	-20.6328	7.8334
	9	-13.64000	4.31563	.066	-27.6894	.4094
	10	-21.48600 *	4.31563	.000	-35.5354	-7.4366
	11	-17.41000 *	4.31563	.004	-31.4594	-3.3606
5	6	-2.82484	4.37205	1.000	-17.0579	11.4082
	7	-11.66522	4.43389	.240	-26.0996	2.7692
	8	-7.97221	4.37205	.765	-22.2053	6.2609
	9	-15.21250 *	4.31563	.022	-29.2619	-1.1631
	10	-23.05850 *	4.31563	.000	-37.1079	-9.0091
	11	-18.98250 *	4.31563	.001	-33.0319	-4.9331
6	7	-8.84038	4.48882	.670	-23.4536	5.7728
	8	-5.14737	4.42775	.986	-19.5618	9.2670
	9	-12.38766	4.37205	.153	-26.6207	1.8454
	10	-20.23366 *	4.37205	.000	-34.4667	-6.0006
	11	-16.15766 *	4.37205	.012	-30.3907	-1.9246
7	8	3.69301	4.48882	.999	-10.9202	18.3062
	9	-3.54728	4.43389	.999	-17.9817	10.8871
	10	-11.39328	4.43389	.272	-25.8277	3.0411
	11	-7.31728	4.43389	.858	-21.7517	7.1171
8	9	-7.24029	4.37205	.856	-21.4734	6.9928
	10	-15.08629 *	4.37205	.028	-29.3194	-.8532
	11	-11.01029	4.37205	.300	-25.2434	3.2228
9	10	-7.84600	4.31563	.768	-21.8954	6.2034
	11	-3.77000	4.31563	.999	-17.8194	10.2794
10	11	4.07600	4.31563	.997	-9.9734	18.1254

* The mean difference is significant at the 0.05 level

Table C.23: ANOVA Post-hoc tests for the instance MWL_{def} of the defeasible framework - Group B - 95% Confidence Interval

MWL_{Def} - Tukey HSD					99% Confidence Interval	
(I) task	(J) task	Mean Diff. (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1	2	13.95350	4.31563	.053	-2.2594	30.1664
	3	14.21050	4.31563	.045	-2.0024	30.4234
	4	6.55300	4.31563	.912	-9.6599	22.7659
	5	8.12550	4.31563	.728	-8.0874	24.3384
	6	5.30066	4.37205	.981	-11.1242	21.7255
	7	-3.53972	4.43389	.999	-20.1969	13.1175
	8	.15329	4.37205	1.000	-16.2716	16.5782
	9	-7.08700	4.31563	.862	-23.2999	9.1259
	10	-14.93300	4.31563	.027	-31.1459	1.2799
	11	-10.85700	4.31563	.302	-27.0699	5.3559
2	3	.25700	4.31563	1.000	-15.9559	16.4699
	4	-7.40050	4.31563	.826	-23.6134	8.8124
	5	-5.82800	4.31563	.958	-22.0409	10.3849
	6	-8.65284	4.37205	.664	-25.0777	7.7720
	7	-17.49322 **	4.43389	.005	-34.1504	-.8360
	8	-13.80021	4.37205	.066	-30.2251	2.6247
	9	-21.04050 **	4.31563	.000	-37.2534	-4.8276
	10	-28.88650 **	4.31563	.000	-45.0994	-12.6736
	11	-24.81050 **	4.31563	.000	-41.0234	-8.5976
3	4	-7.65750	4.31563	.794	-23.8704	8.5554
	5	-6.08500	4.31563	.945	-22.2979	10.1279
	6	-8.90984	4.37205	.623	-25.3347	7.5150
	7	-17.75022 **	4.43389	.004	-34.4074	-1.0930
	8	-14.05721	4.37205	.056	-30.4821	2.3677
	9	-21.29750 **	4.31563	.000	-37.5104	-5.0846
	10	-29.14350 **	4.31563	.000	-45.3564	-12.9306
	11	-25.06750 **	4.31563	.000	-41.2804	-8.8546
4	5	1.57250	4.31563	1.000	-14.6404	17.7854
	6	-1.25234	4.37205	1.000	-17.6772	15.1725
	7	-10.09272	4.43389	.454	-26.7499	6.5645
	8	-6.39971	4.37205	.930	-22.8246	10.0252
	9	-13.64000	4.31563	.066	-29.8529	2.5729
	10	-21.48600 **	4.31563	.000	-37.6989	-5.2731
	11	-17.41000 **	4.31563	.004	-33.6229	-1.1971
5	6	-2.82484	4.37205	1.000	-19.2497	13.6000
	7	-11.66522	4.43389	.240	-28.3224	4.9920
	8	-7.97221	4.37205	.765	-24.3971	8.4527
	9	-15.21250	4.31563	.022	-31.4254	1.0004
	10	-23.05850 **	4.31563	.000	-39.2714	-6.8456
	11	-18.98250 **	4.31563	.001	-35.1954	-2.7696
6	7	-8.84038	4.48882	.670	-25.7039	8.0232
	8	-5.14737	4.42775	.986	-21.7815	11.4868
	9	-12.38766	4.37205	.153	-28.8125	4.0372
	10	-20.23366 **	4.37205	.000	-36.6585	-3.8088
	11	-16.15766	4.37205	.012	-32.5825	.2672
7	8	3.69301	4.48882	.999	-13.1706	20.5566
	9	-3.54728	4.43389	.999	-20.2045	13.1099
	10	-11.39328	4.43389	.272	-28.0505	5.2639
	11	-7.31728	4.43389	.858	-23.9745	9.3399
8	9	-7.24029	4.37205	.856	-23.6652	9.1846
	10	-15.08629	4.37205	.028	-31.5112	1.3386
	11	-11.01029	4.37205	.300	-27.4352	5.4146
9	10	-7.84600	4.31563	.768	-24.0589	8.3669
	11	-3.77000	4.31563	.999	-19.9829	12.4429
10	11	4.07600	4.31563	.997	-12.1369	20.2889

** The mean difference is significant at the 0.01 level

Table C.24: ANOVA Post-hoc tests for the instance MWL_{def} of the defeasible framework- Group B - 99% Confidence Interval

task	2	3	4	5	6	7	8	9	10	11
1	$NASA^{*}$		$NASA^{*}_{**}$ $def-NI^{*}$ def^{*}	$NASA^{*}_{**}$	$NASA^{*}_{**}$ $def-NI^{*}$ def^{*}	$NASA^{*}_{**}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}
2						WP^{*}		$NASA^{*}$ $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}
3								$NASA^{*}$ $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}$ WP^{*} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*} $def-NI^{*}_{**}$ def^{*}_{**}
4								def^{*}	$def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ $def-NI^{*}_{**}$ def^{*}_{**}
5								def^{*}_{**}	$def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ $def-NI^{*}_{**}$ def^{*}_{**}
6									$def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}$ $def-NI^{*}_{**}$ def^{*}_{**}
7								$NASA^{*}$		$NASA^{*}_{**}$ $def-NI^{*}$
8										$NASA^{*}$ $def-NI^{*}$
9										$NASA^{*}$
10										

* $p < 0.05$ ** $p < 0.01$

Table C.25: Post-hoc results of the ANOVA procedure for the mental workload assessment instruments - Group A

task	2	3	4	5	6	7	8	9	10	11
1		def^{*}							$def-NI^{*}$ def^{*}	
2		$NASA^{*}_{**}$				WP^{*}_{**} def^{*}_{**}	$def-NI^{*}_{**}$	WP^{*} def^{*}_{**}	WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} def^{*}_{**}
3			$NASA^{*}$		$NASA^{*}_{**}$	$NASA^{*}_{**}$ WP^{*} def^{*}_{**}	$NASA^{*}_{**}$	$NASA^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} $def-NI^{*}$ def^{*}_{**}	$NASA^{*}_{**}$ WP^{*}_{**} def^{*}_{**}
4							$def-NI^{*}$		$def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ def^{*}_{**}
5							$def-NI^{*}_{**}$	def^{*}	$NASA^{*}$ WP^{*}_{**} $def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}_{**}$ $def-NI^{*}$ def^{*}_{**}
6							$def-NI^{*}_{**}$		$def-NI^{*}_{**}$ def^{*}_{**}	$NASA^{*}$ def^{*}
7										$NASA^{*}$
8									def^{*}	
9										
10										

* $p < 0.05$ ** $p < 0.01$

Table C.26: Post-hoc results of the ANOVA procedure for the mental workload assessment instruments - Group B

C.6 Multicollinearity of each mental workload attribute

	mental	temporal	psychological	performance	effort	central	response	visual	auditory	spatial	verbal	manual	speech	arousal	bias	intention	knowledge	parallelism	skill
mental	1	.29	.31	-.07	.63	.55	.35	.33	.09	.27	.31	.22	.03	.06	.11	.03	-.17	.21	-.02
temporal	.29	1	.40	-.21	.27	.31	.29	.16	.13	.19	.09	.22	.07	.04	.13	.09	-.03	.31	.02
psychological	.31	.40	1	-.20	.28	.33	.33	.23	.14	.29	.14	.26	.11	-.03	.24	-.09	-.01	.16	.04
performance	-.07	-.21	-.20	1	-.15	-.13	.00	.11	-.35	-.09	-.14	.17	-.16	.13	-.13	.28	.17	-.02	.28
effort	.63	.27	.28	-.15	1	.59	.28	.24	.24	.19	.31	.11	.16	.15	.18	.08	-.14	.05	-.05
central	.55	.31	.33	-.13	.59	1	.47	.38	.18	.37	.27	.28	.21	.07	.26	.03	-.03	.29	.09
response	.35	.29	.33	.00	.28	.47	1	.20	.08	.34	.13	.51	.15	-.11	.08	.08	.12	.28	.22
visual	.33	.16	.23	.11	.24	.38	.20	1	-.03	.23	.18	.31	.09	.22	.10	.00	.11	.18	.18
auditory	.09	.13	.14	-.35	.24	.18	.08	-.03	1	.13	.39	-.16	.27	.00	.12	.06	-.17	.07	-.12
spatial	.27	.19	.29	-.09	.19	.37	.34	.23	.13	1	.11	.17	.15	-.09	.38	-.12	-.09	.23	-.09
verbal	.31	.09	.14	-.14	.31	.27	.13	.18	.39	.11	1	.03	.13	.05	-.01	.00	-.09	.02	-.08
manual	.22	.22	.26	.17	.11	.28	.51	.31	-.16	.17	.03	1	.07	.00	-.05	.09	.19	.31	.34
speech	.03	.07	.11	-.16	.16	.21	.15	.09	.27	.15	.13	.07	1	-.07	.14	-.04	.02	.09	.04
arousal	.06	.04	-.03	.13	.15	.07	-.11	.22	.00	-.09	.05	.00	-.07	1	.08	.14	.12	.11	.13
bias	.11	.13	.24	-.13	.18	.26	.08	.10	.12	.38	-.01	-.05	.14	.08	1	-.14	-.05	.21	-.06
intention	.03	.09	-.09	.28	.08	.03	.08	.00	.06	-.12	.00	.09	-.04	.14	-.14	1	.00	.07	.16
knowledge	-.17	-.03	-.01	.17	-.14	-.03	.12	.11	-.17	-.09	-.09	.19	.02	.12	-.05	.00	1	-.01	.50
parallelism	.21	.31	.16	-.02	.05	.29	.28	.18	.07	.23	.02	.31	.09	.11	.21	.07	-.01	1	.08
skill	-.02	.02	.04	.28	-.05	.09	.22	.18	-.12	-.09	-.08	.34	.04	.13	-.06	.16	.50	.08	1

Table C.27: Inter-correlations among mental workload attributes - Group A

	mental	temporal	psychological	performance	effort	central	response	visual	auditory	spatial	verbal	manual	speech	arousal	bias	intention	knowledge	parallelism	skill
mental	1	.22	.28	-.22	.73	.72	.35	.41	.16	.18	.42	.28	.17	-.04	.30	.23	-.13	.25	-.13
temporal	.22	1	.30	-.18	.26	.19	.23	.25	.23	.18	.03	.14	.27	-.03	.17	.16	-.08	.20	-.12
psychological	.28	.30	1	-.39	.35	.29	.21	.30	.17	.34	.15	.09	.20	-.06	.37	.04	-.19	.12	-.18
performance	-.22	-.18	-.39	1	-.21	-.21	-.08	-.09	-.35	-.19	-.20	.06	-.22	.14	-.20	.09	.22	.06	.31
effort	.73	.26	.35	-.21	1	.58	.19	.43	.25	.16	.37	.12	.17	.09	.30	.20	-.23	.16	-.21
central	.72	.19	.29	-.21	.58	1	.39	.41	.13	.26	.23	.30	.21	-.01	.37	.26	-.18	.27	-.08
response	.35	.23	.21	-.08	.19	.39	1	.34	-.11	.23	.15	.56	.10	-.04	.24	.16	-.16	.36	-.05
visual	.41	.25	.30	-.09	.43	.41	.34	1	.07	.30	.17	.29	.12	.01	.33	.24	-.17	.21	-.07
auditory	.16	.23	.17	-.35	.25	.13	-.11	.07	1	.26	.25	-.22	.24	.06	.16	.14	-.17	-.06	-.26
spatial	.18	.18	.34	-.19	.16	.26	.23	.30	.26	1	.06	.18	.42	-.07	.48	.10	-.02	.29	-.06
verbal	.42	.03	.15	-.20	.37	.23	.15	.17	.25	.06	1	.01	.09	.17	.08	.23	-.10	.05	-.20
manual	.28	.14	.09	.06	.12	.30	.56	.29	-.22	.18	.01	1	.01	-.08	.10	.21	-.01	.37	.09
speech	.17	.27	.20	-.22	.17	.21	.10	.12	.24	.42	.09	.01	1	.04	.31	.05	-.01	.27	-.05
arousal	-.04	-.03	-.06	.14	.09	-.01	-.04	.01	.06	-.07	.17	-.08	.04	1	.05	.19	-.10	.03	-.01
bias	.30	.17	.37	-.20	.30	.37	.24	.33	.16	.48	.08	.10	.31	.05	1	.11	-.17	.38	-.19
intention	.23	.16	.04	.09	.20	.26	.16	.24	.14	.10	.23	.21	.05	.19	.11	1	-.03	.16	.02
knowledge	-.13	-.08	-.19	.22	-.23	-.18	-.16	-.17	-.17	-.02	-.10	-.01	-.01	-.10	-.17	-.03	1	.01	.60
parallelism	.25	.20	.12	.06	.16	.27	.36	.21	-.06	.29	.05	.37	.27	.03	.38	.16	.01	1	.04
skill	-.13	-.12	-.18	.31	-.21	-.08	-.05	-.07	-.26	-.06	-.20	.09	-.05	-.01	-.19	.02	.60	.04	1

Table C.28: Inter-correlations among mental workload attributes - Group B

C.7 Shapiro-Wilk test of normality for the mental workload attributes

	1		2		3		4		5		6		7		8		9		10		11	
Attribute	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig
mental	.917	.089	.930	.155	.813	.001	.950	.370	.950	.366	.947	.323	.957	.489	.939	.234	.899	.040	.961	.568	.926	.132
temporal	.806	.001	.841	.004	.854	.006	.928	.144	.894	0.32	.963	.611	.863	.009	.908	.058	.946	.316	969	.730	.921	.103
psychological	.718	.000	.894	.032	.925	.123	.864	.009	.906	.053	.902	.045	.916	.083	.887	.024	.918	.090	.961	.554	.960	.546
performance	.854	.006	.860	.008	.725	.000	.923	.113	.921	.103	.930	.156	.807	.001	.940	.244	.879	.017	.927	.134	.797	.001
effort	.894	.032	.915	.078	.905	.050	.894	.032	.959	.528	.909	.060	.871	.012	.934	.182	.965	.643	.974	.838	.895	.033
central	.887	.024	.913	.074	.886	.023	.887	.023	.952	.392	.950	.368	.887	.023	.918	.092	.887	.023	.953	.417	.945	.294
response	.921	.105	.775	.000	.885	.022	.925	.125	.910	.063	.494	.358	.831	.003	.869	.011	.821	.002	.950	.373	.906	.052
visual	.951	.388	.898	.038	.911	.067	.879	.017	.788	.001	.932	.168	.953	.416	.967	.684	.825	.002	.885	.021	.962	.588
auditory	.687	.000	.640	.000	.884	.021	.648	.000	.712	.000	.625	.000	.642	.000	.911	.065	.882	.019	.823	.002	.888	.025
spatial	.855	.007	.822	.002	.886	.022	.911	.066	.957	.494	.904	.050	.846	.005	.880	.018	.863	.009	.918	.089	.879	.017
verbal	.857	.007	.859	.008	.867	.011	.838	.003	.882	.019	.860	.008	.796	.001	.919	.094	.833	.003	.951	.390	.880	.018
manual	.921	.103	.882	.019	.935	.196	.926	.129	.949	.346	.932	.167	.809	.001	.854	.006	.782	.000	.914	.076	.878	.016
speech	.775	.000	.670	.000	.808	.001	.692	.000	.742	.000	.702	.000	737	.000	.853	.006	.746	.000	.786	.001	.802	.001
arousal	.930	.156	.926	.129	.951	.386	.951	.383	.972	.790	.925	.124	.962	.595	.914	.077	.963	.607	.959	.520	.933	.176
bias	.653	.000	.810	.001	.721	.000	.711	.000	.708	.000	683	.000	796	.001	.908	.058	.906	.053	.787	.001	797	.001
intention	.901	.043	.908	.057	.869	.011	.923	.113	.953	.415	.954	.423	.914	.077	.915	.079	.894	.032	.923	.114	.958	.503
knowledge	.812	.001	.874	.014	.822	.002	.932	.171	.933	.175	.801	.001	.903	.046	.872	.013	.927	.132	.937	.213	.926	.129
parallelism	.694	.000	.634	.000	.928	.144	.745	.000	.862	.009	.752	.000	.789	.001	.756	.000	.753	.000	.847	.005	.767	.000
skill	.930	.157	.897	.036	.884	.020	.951	.390	.927	.138	.958	.506	.926	.131	.876	.015	.912	.068	.957	.479	.936	.199

Table C.29: Shapiro-Wilk normality tests of the mental workload attributes - Group A

	1		2		3		4		5		6		7		8		9		10		11	
Attribute	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig	Stat	Sig
mental	.947	.331	.941	.252	.876	.015	.894	.032	.917	.085	.919	.096	.846	.005	.950	.368	.883	.020	.872	.013	.898	.038
temporal	.949	.345	.850	.005	.904	.049	.888	.025	.845	.004	.910	.063	.915	.778	.929	.149	.914	.076	.904	.049	.946	.309
psychological	.927	.136	.831	.003	.769	.000	.888	.024	.844	.004	.887	.024	.918	.090	.922	.108	.933	.180	.903	.048	.974	.837
performance	.726	.000	.960	.538	.725	.000	.887	.024	.873	.013	.834	.003	.949	.359	.891	.028	.952	.400	.861	.008	.778	.000
effort	.963	.596	.827	.002	.828	.002	.912	.068	.903	.048	.933	.174	.854	.006	.945	.303	.915	.078	.956	.469	.898	.038
central	.977	.892	.909	.061	.877	.015	.900	.042	.841	.004	.928	.141	.853	.006	.935	.196	.858	.007	.952	.392	.956	.467
response	.947	.319	.931	.160	.919	.095	.931	.160	.922	.108	.935	.190	.854	.006	.903	.047	.916	.085	.954	.437	.869	.011
visual	.915	.079	.918	.091	.917	.088	.900	.041	.951	.378	.818	.002	.892	.029	.923	.114	.899	.040	.962	.594	.973	.819
auditory	.546	.000	.719	.000	960	.538	.543	.000	.718	.000	541	.000	834	.003	.817	.012	.848	.005	.862	.009	.891	.028
spatial	.840	.004	.821	.002	.866	.010	.806	.001	.804	.001	.875	.014	.748	.000	840	.004	.869	.011	.923	.114	.886	.022
verbal	.955	.445	.888	.025	.837	.003	.906	.053	.927	.132	.885	.022	.937	.209	.852	.006	.845	.004	.896	.035	.832	.003
manual	.952	.395	.794	.001	.960	.535	.927	.137	.905	.052	.912	.069	.692	.000	.944	.285	.893	.030	.939	.233	.942	.266
speech	.469	.000	.717	.000	595	.000	.738	.000	.543	.000	.693	.000	.784	.001	.707	.000	.771	.000	.728	.000	.855	.007
arousal	.921	.102	.965	.644	.899	.039	.951	.390	.946	.310	.915	.081	.936	.197	.853	.006	.864	.009	.918	.089	.930	.152
bias	.538	.000	.730	.000	.753	.000	.744	.000	.604	.000	.690	.000	.833	.003	.843	.004	.889	.025	.805	.001	.727	.000
intention	.854	.006	.906	.053	.920	.098	.614	.077	.885	.022	.969	.724	.930	.153	.910	.065	.898	.038	.976	.866	.940	.244
knowledge	.895	.034	.935	.192	.836	.003	.928	.138	.810	.001	.836	.003	.842	.004	.866	.010	.867	.011	.853	.006	.910	.064
parallelism	.582	.000	568	.000	876	.015	.787	.001	.785	.001	.688	.000	.868	.011	.854	.006	.686	.000	.929	.147	.870	.012
skill	.910	.063	.847	.005	.895	.033	.922	.106	.911	.066	.910	.065	.900	.041	.918	.092	.879	.017	.943	.273	.910	.063

Table C.30: Shapiro-Wilk normality tests of the mental workload attributes - Group B

C.8 Likelihood ratio tests for the multinomial logistic regression

	Model fitting criteria	Likelihood ratio tests		
effect(s)	-2Log likelihood of reduced model	Chi-square	df	Sig
Intercept	2287.548	28641	21	0.123
effort	2333.009	74.101	21	0.000
psychological	2303.701	44.793	21	0.002
mental	2294.018	35.111	21	0.027
temporal	2376.493	117.586	21	0.000
performance	2360.125	101.217	21	0.000

The chi-square statistic is the difference in -2 log-likelihoods between the final and a reduced model that is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Table C.31: Likelihood ratio tests of the multinomial logistic regression with the attributes of the NASATLX

	Model fitting criteria	Likelihood ratio tests		
effect(s)	-2Log likelihood of reduced model	Chi-square	df	Sig
Intercept	1858.838	84.953	21	0.000
speech	1828.189	54.304	21	0.000
verbal	1856.830	82.945	21	0.000
auditory	2129.535	355.650	21	0.000
response	1832.535	58.504	21	0.000
central	1847.477	73.592	21	0.000
visual	1820.956	47.071	21	0.001
spatial	1831.489	57.604	21	0.000
manual	1843.307	69.423	21	0.000

The chi-square statistic is the difference in -2 log-likelihoods between the final and a reduced model that is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Table C.32: Likelihood ratio tests of the multinomial logistic regression with the attributes of the WP

	Model fitting criteria	Likelihood ratio tests		
effect(s)	-2Log likelihood of reduced model	Chi-square	df	Sig
Intercept	1228.870	40.302	21	0.000
skill	1227.784	39.216	21	0.009
knowledge	1241.983	53.415	21	0.000
bias	1243.347	54.780	21	0.000
speech	1244.528	55.960	21	0.000
verbal	1250.175	61.607	21	0.000
auditory	1499.734	311.166	21	0.000
response	1245.445	56.877	21	0.000
effort	1270.755	82.187	21	0.000
psychological	1234.922	46.355	21	0.001
temporal	1300.782	112.214	21	0.000
performance	1243.812	55.244	21	0.000
central	1224.444	35.877	21	0.023
visual	1247.265	58.697	21	0.000
spatial	1239.049	50.481	21	0.000
manual	1247.037	58.469	21	0.000
arousal	1226.932	38.364	21	0.012
parallelism	1289.076	100.509	21	0.000

The chi-square statistic is the difference in -2 log-likelihoods between the final and a reduced model that is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Table C.33: Likelihood ratio tests the multinomial logistic regression with the attributes of the instances of the defeasible framework (MWL_{def} and MWL^{NI}_{def})

C.9 Predictions of the multinomial logistic regressions model

		Group A											GroupB										
	Observed	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9	10	11	% correct
Group A	1	8	0	2	0	0	1	2	1	0	0	0	0	1	3	0	2	0	0	0	0	0	0	40.0
	2	4	0	0	3	2	1	2	2	0	0	1	0	0	0	1	0	0	0	2	0	1	1	0.0
	3	3	1	2	0	0	0	0	2	0	1	1	0	0	3	0	0	4	0	0	0	1	2	10.0
	4	2	0	0	5	1	0	4	1	2	0	0	0	0	0	0	2	0	0	0	2	1	0	25.0
	5	1	0	0	3	5	0	0	0	0	1	0	2	1	0	1	1	0	2	2	0	0	1	25.0
	6	1	0	2	1	1	0	2	0	0	3	1	0	0	1	0	0	2	1	0	0	3	2	0.0
	7	0	0	0	1	1	0	8	2	0	0	1	0	0	1	0	1	1	1	0	0	1	2	40.0
	8	2	0	0	0	0	0	3	2	0	2	2	0	0	0	0	2	1	1	1	0	1	3	10.0
	9	0	1	0	1	1	2	3	0	5	2	1	0	0	0	0	1	0	0	0	1	2	0	25.0
	10	0	0	0	0	1	1	0	0	1	4	3	1	0	1	0	1	1	0	3	0	3	0	20.0
	11	0	0	1	0	0	0	0	2	1	1	11	0	0	0	0	0	0	0	1	0	1	2	55.0
Group B	1	4	1	0	0	1	0	0	1	0	2	2	3	0	2	0	1	1	0	1	0	0	1	15.0
	2	3	2	0	4	0	0	4	3	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0.0
	3	5	0	1	0	0	0	0	0	0	0	0	2	0	9	0	0	1	1	1	0	0	0	45.0
	4	1	0	1	3	1	0	1	2	2	0	0	0	1	0	1	5	2	0	0	0	0	0	5.0
	5	4	0	0	5	1	1	0	1	1	1	0	0	0	2	0	3	0	0	0	0	1	0	15.0
	6	1	0	2	0	1	1	1	1	1	2	0	0	0	3	1	0	3	0	0	0	2	1	15.0
	7	1	1	0	0	2	0	3	1	4	0	0	0	1	1	1	0	0	4	0	1	0	0	20.0
	8	3	0	3	1	1	1	0	2	1	1	2	1	0	1	0	0	0	0	0	0	0	3	0.0
	9	2	1	0	0	1	0	2	1	1	2	1	0	1	0	0	1	1	1	1	1	3	0	5.0
	10	0	0	2	0	2	1	0	0	3	2	1	1	0	0	2	0	2	0	0	0	4	0	20.0
	11	0	0	0	0	0	0	0	1	2	1	8	0	0	0	0	0	0	0	1	0	1	6	30.0
	Tot. %	10.2	1.6	3.6	6.1	5.0	2.0	8.0	5.7	5.7	5.7	8.0	2.3	1.1	6.1	1.6	4.8	4.5	2.7	3.0	1.1	5.7	5.5	19.1%

Table C.34: Predicted task membership by the multinomial logistic regression with the attributes of the NASATLX

		Group A											GroupB										
	Observed	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9	10	11	% correct
Group A	1	9	1	1	0	0	2	0	2	0	1	0	1	1	1	0	1	0	0	0	0	0	0	45.0
	2	3	8	0	0	0	0	0	2	0	1	0	1	5	0	0	0	0	0	0	0	0	0	40.0
	3	3	1	8	0	0	1	0	2	0	0	0	0	0	4	0	0	0	0	0	0	1	0	40.0
	4	1	1	0	2	2	1	0	1	0	3	0	4	1	0	1	0	1	0	0	1	1	0	10.0
	5	1	2	1	2	3	2	0	0	1	1	0	1	0	0	1	0	1	0	0	2	2	0	15.0
	6	3	1	0	2	1	3	0	1	0	1	0	0	4	0	1	1	1	0	0	0	1	0	15.0
	7	0	0	1	0	0	0	10	1	0	0	1	0	0	0	0	0	0	4	0	0	0	3	50.0
	8	0	0	1	0	0	0	2	6	1	0	1	0	2	2	0	0	0	0	2	2	1	0	30.0
	9	1	0	0	0	0	0	1	2	9	0	0	0	0	0	0	0	0	0	1	3	2	1	45.0
	10	0	2	1	1	3	0	0	0	0	8	0	1	0	0	0	2	0	0	2	0	0	0	40.0
	11	0	0	1	0	0	0	3	1	1	0	7	0	0	0	0	0	0	4	0	1	0	2	35.0
Group B	1	1	2	0	1	0	1	0	0	0	1	0	4	3	0	0	1	3	0	1	0	2	0	20.0
	2	1	4	0	0	1	0	0	0	0	0	0	1	12	0	0	1	0	0	0	0	0	0	60.0
	3	0	1	3	0	0	0	0	1	0	0	0	0	0	13	0	0	0	0	0	1	1	0	65.0
	4	2	0	1	0	0	0	0	0	1	2	0	4	2	0	2	5	0	0	0	0	0	1	10.0
	5	3	0	2	0	0	1	0	1	1	1	0	3	0	0	2	5	1	0	0	0	0	0	25.0
	6	1	0	0	1	1	2	0	0	0	3	0	0	1	0	1	3	6	0	1	0	0	0	30.0
	7	0	0	0	0	0	0	2	0	2	0	1	0	0	2	0	0	0	9	0	1	0	3	45.0
	8	2	0	1	0	0	0	0	2	0	1	1	1	0	1	4	1	0	0	0	3	3	0	0.0
	9	0	1	0	0	0	0	0	1	3	1	0	1	0	0	0	0	1	1	0	8	1	2	40.0
	10	0	0	0	1	2	1	0	1	2	5	0	1	0	0	2	1	1	0	0	0	3	0	15.0
	11	0	0	0	0	0	0	1	0	0	0	3	0	0	1	0	0	0	7	0	1	0	7	35.0
	Tot. %	7.0	5.5	4.8	2.3	3.0	3.2	4.3	5.5	4.8	6.6	3.2	5.2	7.0	5.5	3.2	4.8	3.4	5.7	1.6	5.2	4.1	4.3	32.3%

Table C.35: Predicted task membership by the multinomial logistic regression with the attributes of the WP

		Group A											GroupB										
	Observed	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9	10	11	% correct
Group A	1	12	0	0	1	0	2	0	0	1	0	0	0	1	2	0	1	0	0	0	0	0	0	60.0
	2	1	13	0	0	1	1	0	1	1	0	0	0	2	0	0	0	0	0	0	0	0	0	65.0
	3	0	0	8	0	0	1	0	4	0	1	0	1	1	4	0	0	0	0	0	0	0	0	40.0
	4	0	0	0	7	5	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	1	0	35.0
	5	1	1	0	3	8	0	0	2	0	0	0	1	1	0	1	0	0	0	0	1	1	0	40.0
	6	3	1	1	1	1	6	0	1	0	0	0	1	1	0	0	0	3	0	1	0	0	0	30.0
	7	0	0	0	1	0	0	12	1	0	0	1	0	0	0	0	0	0	5	0	0	0	0	60.0
	8	0	1	1	0	1	0	2	10	1	0	1	0	0	0	0	0	0	0	2	1	0	0	50.0
	9	1	0	0	0	0	0	1	1	14	1	0	0	0	0	0	0	0	0	1	1	0	0	70.0
	10	0	0	0	1	0	0	0	1	0	8	0	0	0	0	1	0	0	0	1	1	6	1	40.0
	11	0	0	1	0	0	0	0	0	2	0	13	0	0	0	0	0	0	0	1	0	0	3	65.0
Group B	1	0	0	1	1	1	0	0	0	0	0	0	11	0	0	2	1	1	0	1	1	0	0	55.0
	2	0	3	0	0	0	0	0	0	0	0	0	0	15	0	0	0	2	0	0	0	0	0	75.0
	3	0	0	3	0	0	0	0	0	0	0	0	0	0	16	1	0	0	0	0	0	0	0	80.0
	4	2	0	0	1	0	0	0	1	0	0	0	0	1	0	7	3	3	0	1	0	0	1	35.0
	5	2	0	1	1	1	0	0	0	0	0	0	2	0	0	2	8	1	0	1	0	1	0	40.0
	6	1	0	0	3	0	2	0	0	0	2	0	2	0	0	0	1	7	0	2	0	0	0	35.0
	7	0	0	0	0	0	0	6	0	0	0	0	0	0	1	0	0	0	12	0	0	0	1	60.0
	8	1	0	0	0	0	0	0	1	1	3	0	1	0	1	0	0	0	0	9	0	1	2	45.0
	9	0	0	0	0	0	0	0	1	4	0	1	0	0	0	0	0	0	2	0	12	0	0	60.0
	10	0	0	1	0	1	0	0	0	0	4	0	0	0	0	0	1	0	0	1	0	12	0	60.0
	11	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	2	0	14	70.0
	Tot. %	5.5	4.3	3.9	4.5	4.3	3.0	5.0	5.7	5.5	4.5	4.3	4.5	5.2	5.5	3.2	3.6	4.1	4.3	4.8	4.3	5.0	5.0	53.2 %

Table C.36: Predicted task membership by the multinomial logistic regression with the attributes of the new instances of the defeasible framework (MWL_{def} and MWL_{def}^{NI})

C.10 Step summaries of the multinomial logistic procedure

			Model fitting criteria	Likelihood ratio tests		
Model	Action	effect(s)	-2Log likelihood of reduced model	Chi-square [a]	df	Sig
0	Entered	Intercept	2720.117			
1	Entered	temporal	2579.573	140.544	21	0.000
2	Entered	effort	2446.946	132.627	21	0.000
3	Entered	performance	2337.516	109.430	21	0.000
4	Entered	psychological	2294.018	43.497	21	0.003
5	Entered	mental	2258.907	35.111	21	0.027

Stepwise Method: Forward Entry

[a] The chi-square for entry is based on the likelihood ratio test.

Table C.37: Step summary of the multinomial logistic regression with the attributes of the NASATLX

Model	Action	effect(s)	Model fitting criteria	Likelihood ratio tests		
			-2Log likelihood of reduced model	Chi-square [a]	df	Sig
0	Entered	Intercept	2720.117			
1	Entered	auditory	2312.625	407.493	21	0.000
2	Entered	central	2190.580	122.044	21	0.000
3	Entered	manual	2083.952	106.628	21	0.000
4	Entered	verbal	1987.693	96.259	21	0.000
5	Entered	spatial	1933.067	54.626	21	0.000
6	Entered	response	1872.922	60.144	21	0.000
7	Entered	speech	1820.956	51.967	21	0.000
8	Entered	visual	1773.885	47.071	21	0.001

Stepwise Method: Forward Entry [a]The chi-square for entry is based on the likelihood ratio test.

Table C.38: Step summary of the multinomial logistic regression with the attributes of the WP instrument

Model	Action	effect(s)	Model fitting criteria	Likelihood ratio tests		
			-2Log likelihood of reduced model	Chi-square [a]	df	Sig
0	Entered	Intercept	2720.117			
1	Entered	auditory	2312.625	407.493	21	0.000
2	Entered	parallelism	2165.939	146.686	21	0.000
3	Entered	temporal	2051.458	114.481	21	0.000
4	Entered	effort	1934.399	117.059	21	0.000
5	Entered	manual	1839.024	95.375	21	0.000
6	Entered	bias	1750.272	88.751	21	0.000
7	Entered	verbal	1674.376	75.896	21	0.000
8	Entered	knowledge	1617.276	57.099	21	0.000
9	Entered	speech	1559.886	57.390	21	0.000
10	Entered	performance	1504.388	55.498	21	0.000
11	Entered	visual	1444.192	60.196	21	0.000
12	Entered	response	1396.445	47.747	21	0.001
13	Entered	spatial	1347.588	48.857	21	0.001
14	Entered	psychological	1303.836	43.752	21	0.003
15	Entered	skill	1262.599	41.237	21	0.005
16	Entered	arousal	1224.444	38.154	21	0.012
17	Entered	central	1188.568	35.877	21	0.023

Stepwise Method: Forward Entry [a]The chi-square for entry is based on the likelihood ratio test.

Table C.39: Step summary of the multinomial logistic regression with the attributes of the instances of the defeasible framework (MWL_{def} and MWL_{def}^{NI})

Appendix D

D.1 Consent form

D.1.1 Study Information

The explosion of the Internet as a collaborative, accessible platform and data source is rapidly changing the way in which people access, seek, publish, and consume information. Emerging applications shape our cognitive development, with fresh modes of interaction evolving new and interesting behaviours rich in useful data. User exchanges with and through websites and popular communications channels, including Twitter, Facebook, email, and instant messaging applications, are constantly monitored and mined for pertinent patterns to bring such knowledge to bear in the improvement of services and more accurately targeted content delivery. Such data is explicit in detailing a user's interaction with specific web pages or users, but is currently deficient in elucidating on a user's original goals or intentions. Use of the internet is, in actuality, a complex cognitive activity involving auditory, tactile, and visual human modalities rather than the mere external physical channels we presently monitor. The analysis and prediction of a user's cognitive engagement whilst involved in such activity is the main focus and inspiration for this study. To this end, we aim to introduce the concept of Human Mental Workload - currently in use by Psychological, Neuro, and Cognitive Sciences - to the field of Computer Science, as it becomes increasingly connected to Social and Behavioural research. The objective measurement of human mental workload during task execution online enables aggregate behaviour analysis, useful toward the study of collective intelligence in large groups of users. In this study your behaviour while surfing the World Wide Web will be monitored, implicitly gathered by a non-invasive piece of software, and stored in a database for future analysis. The study aims to capture detailed interaction information for performed actions to automatically assess user engagement. All actions you perform over webpages will be recorded and saved, these include clicking, scrolling, mouse gestures and movements, and keyboard usage. However, data input in online forms will not be collected. Your job is to naturally interact with the Web. You will be asked to participate in a subjective questionnaire to obtain feedback on your experience with the execution of online tasks.

You can leave the study or request a break at any time. This study is conducted in accordance with the School of Computer Science and Statistics at Trinity College Dublin, along with its ethics guidelines. Your rights as a participant, including the right to withdraw at any point without penalty, are ensured. It is anticipated that the findings of the study will be written up for publication in a peer-reviewed journal and presented at international conferences. All results will be anonymised and it will not be possible to identify individual participants' data.

D.1.2 Frequently Asked Questions

1. Is the study anonymous? Yes, the study is totally anonymous, collected data will not be linked to your identity.
2. Will my user experience while surfing the Web be altered by the monitoring technology? No, your experience while surfing the Web will not change. The monitoring technology will be completely invisible and non-invasive.
3. Will data I input online such as logins and passwords be captured and stored somewhere? No, data entered online such as web-form input, logins, passwords, or email addresses, will not be recorded.
4. How is my privacy guaranteed? Your personal data will not be stored. Your logins, password, and any data entered into forms or over social networks, wikis, blogs, or other resources will never be recorded.
5. Will recorded data be linked to me? No, recorded data will never be linked to you. Our software will randomly generate a code to identify your Web interactions over time, but this code can never be associated with your personal data, computer IP, or computer MAC address as we never store such info..
6. Is the captured data stored in a public database? No, the captured data will be stored in a local password-protected database behind proxy machines and firewalls within the Distributed Systems Group's network at Trinity College Dublin.
7. Who will have access to stored participants' data, and what about confidentiality? Only the researcher of this study will have access to your interaction data, exclusively for research purposes. The researcher will never be able to associate any stored data with the identity of a specific user, as this information is never stored. No one else will have the right to access any stored information.
8. What does the software look like? The software solution is totally transparent. It comes as either a proxy - meaning you need only configure your browser to point to the proxy - or a plug-in/add-on for your favourite browser (Firefox or Chrome) that is installed only once. Nothing further is required.

D.1.3 Consent form

Name of Participant: Sex: Male/Female Date of birth:
Researcher: Luca Longo

I consent to participate in this study. I am satisfied with the instructions I have been given so far and I expect to have any further information requested regarding the study supplied to me at the end of the experiment. I have been informed that the confidentiality of the data I provide will be safeguarded. I am free to ask any questions at any time before and during the study. I have been provided with a copy of this form and the participant information sheet.

I understand that my behaviour while surfing the World Wide Web will be monitored and stored for statistical analysis. The aim of the study is to capture interaction actions for automatically assessing indexes of user's engagement. Technically, my web-interaction will be saved and all the actions I perform over web-pages, such as clicking, scrolling, mouse movements and keyboard usage, will be recorded. All the data I will send through web-forms will not be saved in anyhow but just the action will be recorded.

I have not been coerced in any way to participate in this study and I understand that I may terminate my participation in the study at any point should I so wish. I am at least 18 years of age.

D.1.4 Data Protection

I agree to the University processing personal data that I have supplied. I agree to the processing of such data for any purposes connected with the Research Project as outlined to me.

Also, I understand that my participation is entirely voluntary, that I may refuse to answer any question and may withdraw at any time without prejudice. I agree to Luca Longo and Trinity College, University of Dublin storing of any data which results from this project. I agree to the processing of such data for any purposes connected with the research project as outlined to me. I understand that my participation is fully anonymous, no personal details will be recorded, no images or video will be stored and all information collected will remain confidential. I have been provided with an information letter which outlines the activities I will take part in, how data will be collected and stored and how I can contact the researcher. I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity. In the extremely unlikely event that illicit activity is reported, Luca Longo will be obliged to report it to appropriate authorities. I have read this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me. I have received a copy of this agreement.

Date:
Name of participant (print): Signed:
Name of researcher (print): Luca Longo Signed:

Researcher's Contact details: llongo@cs.tcd.ie
Department of Computer Science and Statistics, Distributed Systems Group. Trinity College Dublin