2017

Representing and Inferring Mental Workload via Defeasible Reasoning: a Comparison with the NASA Task Load Index and the Workload Profile

Lucas Middeldorf Rizzo
d15123771@mydit.ie

Luca Longo
Dublin Institute of Technology

Representing and inferring mental workload via defeasible reasoning: a comparison with the NASA Task Load Index and the Workload Profile

Lucas Rizzo[1], Luca Longo[1,2*]

[1] School of Computing, Dublin Institute of Technology
[2] The ADAPT global centre of excellence for digital content and media innovation
[*] luca.longo@dit.ie

Abstract. The NASA Task Load Index ($NASA-TLX$) and the Workload Profile (WP) are likely the most employed instruments for subjective mental workload (MWL) measurement. Numerous areas have made use of these methods for assessing human performance and thusly improving the design of systems and tasks. Unfortunately, MWL is still a vague concept, with different definitions and no universal measure. This research investigates the use of defeasible reasoning to represent and assess MWL. Reasoning is defeasible when a conclusion, supported by a set of premises, can be retracted in the light of new information. In this empirical study, this type of reasoning is considered for modelling MWL, given the intrinsic uncertainty involved in assessing it. In particular, it is shown how the $NASA-TLX$ and the WP can be translated into defeasible structures whose inferences can achieve similar validity of the original instruments, even when less information is available. It is also discussed how these structures can have a higher extensibility and how their inferences are more self-explanatory than the ones produced by the $NASA-TLX$ and WP.

1 Introduction

Mental Workload (MWL) is a nebulous concept with no precise and broadly accepted characterization. An oversimplified explanation might define it as the amount of required cognitive work devoted in a specific task with limited execution time. However, other elements such as frustration, physical demand and stress might also impact general mental workload. Several fields of application have adopted MWL, such as psychology, ergonomics and human-computer interaction [23,7,8,27]. The aim of measuring MWL is to predict operator and system performance [3]. Optimal workload has several beneficial effects such as user satisfaction, productivity and safety [10]. Despite these benefits, modeling MWL is a fragmented task. Often the information necessary for modelling it is uncertain, vague and contradictory [10]. State-of-the-art MWL measurement techniques do not take into consideration the inconsistency of the data used in modelling it, which might lead to contradictions and loss of information. For example, on one hand, if the time spent on a certain task is low, it is reasonable to assume that the overall MWL exerted on that task is also low. On the other hand, if the effort invested in the task is extremely high, then the contrary (high MWL) can be inferred. This type

of reasoning is referred, in formal logics, as defeasible reasoning. A defeasible concept is built upon a set of interactive pieces of evidence (the reasons) that can become defeated by additional reasons [10]. The assumption here is that this kind of reasoning can be successfully applied for modelling MWL, since it is formed by several pieces of evidence that can be retracted in the light of new evidence. A computational implementation of defeasible reasoning is provided by Argumentation Theory (AT) [4]. AT is aimed at constructing arguments, determining conflicts between them and deciding which ones are eventually acceptable. The research question under investigation is: *can a multi-layer argument-based framework, built upon argumentation theory, compared to state-of-the-art MWL inference techniques, enhance the modelling of mental workload according to validity?* An empirical user study has been conducted in a third-level teaching environment with students from a post-graduate level in Ireland. At the end of a selection of teaching sessions, their MWL has been assessed using two well-known subjective mental workload assessment techniques: the NASA Task Load Index and the Workload Profile. These were used as baselines to evaluate a novel MWL model built upon defeasible reasoning and computationally implemented with AT. The inferential capacity of this model was subsequently compared against the one of the baselines in terms of validity.

The remainder of this paper is organised as follows: section 2 presents related work on MWL, its assessment techniques and criteria for their evaluation. Subsection 2.3 discusses MWL and how it can be seen as a defeasible phenomenon. The design of the experiment and the methodologies adopted are explained in section 3. Section 4 presents results while section 5 concludes the study and indicates possible future work.

2 Related work

MWL is intrinsically complex and multifaceted [20] with no broadly accepted definition. According to Cain [3] it could be intuitively defined as the total amount of cognitive load necessary to perform a specific task under a limited extent of time. Measuring MWL is fundamental in predicting human performance and thusly advising the design of interfaces, information-based procedures and technologies [13]. There are distinctive methods that have been proposed for measuring MWL [5,21,16]. This paper adopts the class of *subjective measures*. hhis class depends on the investigation of the subjective feedback produced by humans interacting with an underlying task and system. The feedback through surveys or questionnaires is often post-task. The most notable are the $NASA{-}TLX$ [7], the Workload profile [24], and the Subjective Workload Assessment Technique [20]. Other classes of measurement include *task performance measures* and *physiological measures*. The first is regularly alluded to as primary and/or secondary tasks measures and it concentrates on the estimation of the objective performance accomplished by humans during the execution of an underlying assignment. Reaction time to secondary tasks, number of errors, actions performed during the primary task and task completion time are examples of performance measures; the second class is based upon the investigation of physiological indicators and responses of the human

body. A few cases incorporate EEG (electroencephalogram), eye tracking and heart rate measures.

2.1 Criteria for development of MWL measurement methods

Several criteria have been proposed for the selection and development of measurement techniques [17]. Since the goal of this study is to investigate the ability of defeasible reasoning, through AT, to represent and assess MWL, the focus is on one specific criteria namely *validity*. In general, it is used to determine whether the measurement instrument is actually measuring MWL. Two forms are usually employed in Psychology.

- *face validity*: it is the extent to which a certain test is subjectively perceived by subjects as covering the concept it purports to measure. In other words, if the workload subjectively reported appears to be valid to participants of the experiment.
- *convergent validity*: it demonstrates the extend to which different MWL techniques correlate to each other [24].

In literature, face and convergent validity are generally calculated adopting statistical correlation coefficients [22].

2.2 The NASA-Task Load Index and the Workload Profile

This study makes use of two subjective measures of mental workload that have been broadly utilized in many research studies in the most recent decades: the NASA-Task Load Index ($NASA{-}TLX$) [7] and the the Workload Profile (WP) [24]. The first was initially created in the field of aviation [7] and has been utilized over an extensive range of applications, including transportation, cognitive psychology and human-computer interaction [9]. It is a combination of six factors believed to influence mental workload: mental, temporal and physical demand, frustration, effort and performance (table 8 in the appendices). Each factor is evaluated with a subjective judgment combined with a weight w calculated via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial coefficient of the 6 factors $\binom{6}{2} = 15$), *'which of the two contributed the most to mental workload during the task'*, such as 'Frustration or Effort?', 'Performance or Temporal Demand?', and so forth. The weights w are the number of preferences, for each dimension, in the 15 answer set (the number of times that each dimension was chosen). For this situation, the range is from 0 (not significant) to 5 (more significant than any other attribute). Ultimately, the final MWL score is calculated as a weighted average, considering the subjective rating of each attribute d_i (for the 6 dimensions) and the correspondent weights w_i (eq. 1). Another MWL assessment technique is the Workload Profile which is based on the Multiple Resource Theory (MRT) [26]. In contrast to the $NASA - TLX$, it is built upon 8 dimensions: context bias, speech response, manual response, auditory resources, visual resources, task and space, verbal material and selection of response (table 9). The user is requested to rate the extent of attentional resources, in the range 0 to 1, for each dimension which are in turn summed (eq. 2).

$$\text{TLX}_{\text{MWL}} = \Big(\sum_{i=1}^{6} d_i \times w_i \Big) \frac{1}{15} \quad (1) \qquad \text{WP}_{\text{MWL}} = \sum_{i=1}^{8} d_i \quad (2)$$

2.3 Mental Workload as a defeasible phenomenon

Different studies suggest that mental workload is supposedly formed by inconsistent pieces of evidence supporting contradictory levels of MWL which are retractable in the light of new information [10,12]. For example, taking into account some of the $NASA-TLX$ attributes (table 8) the following arguments could be shaped:

- *If user has reported high effort then perceived MWL is believed to be high*
- *If user has reported low mental demand then perceived MWL is believed to be low*

Since there are different levels of MWL which can be inferred (low, high), a workload designer might be uncertain about the best assessment, unless there is some sort of preference helping the reasoning process. However, preferentiality might not be the most appropriate technique to settle the dispute. For instance, high effort could be acknowledged as inconsistent in a scenario in which the user has also reported low mental demand. State-of-the-art MWL inference techniques do not provide the possibility to consider such cases. Inconsistent pieces of evidence are aggregated together which might lead to contradictions and loss of information. Here it is argued that defeasible reasoning might have an important role in improving the representation and inference of MWL when compared to the $NASA-TLX$ and the WP instruments. AT is a computational approach to model defeasible reasoning, including activities such as formalisation of arguments, counterarguments, preferences and conflict resolution strategies [11]. These activities, emerged in the literature, can be clustered in a 5 layer schema [12] as depicted in figure 1.

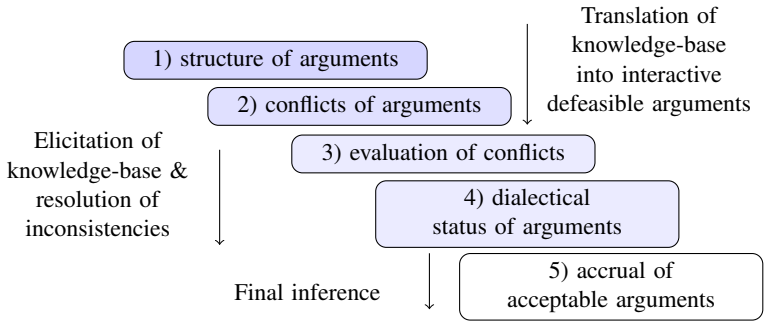


Fig. 1: Five layers upon which argumentation systems are generally built

3 Design

An empirical experiment has been designed, following the 5-layer schema of figure 1. The goal is to re-use the information and knowledge available in the original $NASA-TLX$ and WP instruments, and translate them into defeasible structures.

Layer 1 - Definition of the structure of arguments. The knowledge-base of a designer in relation to MWL can be initially represented as a set of *forecast* arguments like:

$$Argument : premises \rightarrow conclusion$$

This is a structure composed by a set of premises related to a given workload attribute and a conclusion derivable by applying an inference rule $\rightarrow$. A set of premises represents a set of reasons to believe that mental workload is likely to fall within a certain region (example low or high). In this research study, these regions are treated as conclusions of reasoning arguments and their ranges have been established as below:

- U: underload $[0..32] \in \Re$
- F^-: fitting lower load $[33..49] \in \Re$
- F^+: fitting upper load $[50..66] \in \Re$
- O: overload $[67..100] \in \Re$

Some arguments proposed to model mental workload along with their activation ranges are listed in table 1. In this case only the information available in the original NASA Task Load Index instrument is taken. Note that this is just a possible list of arguments but other definitions are possible. It is important to highlight that, as designers, we don't feel physical load should be considered as a dimension in the inference, thus it is not taken into consideration.

Table 1: Forecast arguments of experimental knowledge-base

MD1: [mental demand $\in$ [0, 32] $\rightarrow U$]	PF2: [performance $\in$ [49, 33] $\rightarrow F^+$]
MD2: [mental demand $\in$ [33, 49] $\rightarrow F^-$]	PF3: [performance $\in$ [66, 50] $\rightarrow F^-$]
MD3: [mental demand $\in$ [50, 66] $\rightarrow F^+$]	PF4: [performance $\in$ [100, 67] $\rightarrow U$]
MD4: [mental demand $\in$ [67, 100] $\rightarrow O$]	EF1: [effort $\in$ [0, 32] $\rightarrow U$]
TD1: [temporal demand $\in$ [0, 32] $\rightarrow U$]	EF2: [effort $\in$ [33, 49] $\rightarrow F^-$]
TD2: [temporal demand $\in$ [33, 49] $\rightarrow F^-$]	EF3: [effort $\in$ [50, 66] $\rightarrow F^+$]
TD3: [temporal demand $\in$ [50, 66] $\rightarrow F^+$]	EF4: [effort $\in$ [67, 100] $\rightarrow O$]
TD4: [temporal demand $\in$ [67, 100] $\rightarrow O$]	FR1: [frustration $\in$ [0, 32] $\rightarrow U$]
PF1: [performance $\in$ [32, 0] $\rightarrow O$]	FR2: [frustration $\in$ [67, 100] $\rightarrow O$]

Layer 2 - Definition of the conflicts of arguments: The previous monological structure (forecast arguments, layer 1, section 3), aimed at internally representing an argument, is complemented by dialogical structures, which are focused on the relationships among arguments. A dialogical structure investigates the issue of invalid arguments that appear to be valid (fallacious arguments). According to [15], this type of argument can be referred to as *mitigating argument*. It is an undermining inference $\Rightarrow$ that links a set or premises to an argument B, negating its validity.

$$Argument : premises \Rightarrow B$$

The notion of mitigating argument allows a designer to model possible conflicts between arguments. *Conflict*, often known as *attack* or *counterargument*, is an important notion in defeasible reasoning. Here, two types of conflicts are defined: rebutting and undercutting. A *rebutting attack* occurs when a forecast argument negates the conclusions of another argument. A rebuttal attack is symmetrical so it holds that if an argument A rebuts B ($\Leftrightarrow$), then also B rebuts A. This type of attack models special scenarios that are believed to be logically improbable. Table 2 lists the rebutting attacks proposed in this study, that occur between the forecast arguments defined in table 1.

An *undercutting attack* occurs when the target argument uses a defeasible (tentative) inference rule, thus it can be attacked on its inference by arguing that there is a special case that does not allow the application of the defeasible inference rule [18,19]. In

Table 2: Rebutting attacks

R1: (MD1 $\Leftrightarrow$ FR2)	R3: (TD1 $\Leftrightarrow$ FR2)	R5: (FR1 $\Leftrightarrow$ TD4)	R7: (EF1 $\Leftrightarrow$ MD4)
R2: (MD1 $\Leftrightarrow$ EF4)	R4: (FR1 $\Leftrightarrow$ MD4)	R6: (FR1 $\Leftrightarrow$ EF4)	R8: (EF1 $\Leftrightarrow$ FR2)

contrast to rebutting, 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. Table 3 lists undercutting attacks developed using the forecast arguments in table 1.

Table 3: Undercutting attacks

U1: [perform. $\in$ [67, 100] $\Rightarrow$ FR2]	U2: [perform. $\in$ [0, 32] $\Rightarrow$ FR1]
U3a: [effort $\in$ [67, 100] **&** perform. $\in$ [0, 32] $\Rightarrow$ MD1]	U4a: [effort $\in$ [0, 32] **&** perform. $\in$ [67, 100] $\Rightarrow$ MD4]
U3b: [effort $\in$ [67, 100] **&** perform. $\in$ [0, 32] $\Rightarrow$ TD1]	U4b: [effort $\in$ [0, 32] **&** perform. $\in$ [67, 100] $\Rightarrow$ TD4]

The set of arguments, forecast and mitigating (nodes) as well as the set of attacks, rebutting and undercutting (links) can be seen as a graph, now on referred to as *argumentation framework*. This represents a knowledge-base of a designer that can be now elicited for assessing mental workload. Fig. 2 illustrates an argumentation framework using the arguments in tables 1, 2 and 3.

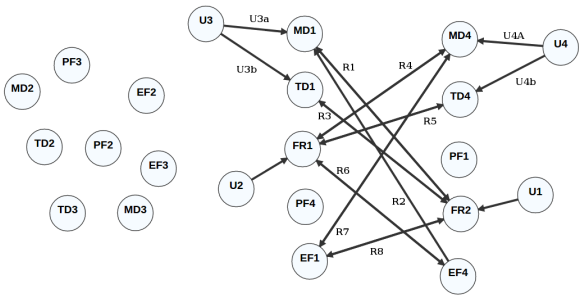


Fig. 2: Argumentation framework: graphical representation of a knowledge-base using the $NASA-TLX$ attributes

Layer 3 - evaluation of the conflicts of arguments: Once the knowledge-base of a designer is formally translated into an *argumentation framework*, it can be now elicited with inputs provided by human subjects. These inputs activate some of the arguments in the argumentation framework, discarding others. For example, if a human subject rated the mental demand question of table 8 with a value of 80, then argument $MD4$ is activated, while arguments $MD1, MD2$ and $MD3$ are discarded. Based on the inputs gathered from the original $NASA-TLX$ questionnaire (table 8), a *sub-argumentation framework* emerges, that can be evaluated against inconsistencies and used to compute the dialectical status of each argument. It is important to highlight that in this study all types of attacks assume a form of a binary relation between two arguments. Once

the arguments of an attack are activated, the attack is automatically considered valid. However, in other domains it is possible that the evaluation of attacks are made through the preferentiality of arguments or strength of arguments, or through the preferentiality of attacks or strength of an attack relation [6].

Layer 4 - definition of the dialectical status of arguments: In order to investigate the potential inconsistencies that might emerge from the interaction of activated arguments (sub-argumentation framework emerged in layer 3), Dung-style acceptability semantics are applied [4]. The underlying idea is that, given a set of arguments, where some of them defeat (attack) others, a decision is to be taken to determine which arguments can ultimately be accepted. Merely looking at an argument's defeaters to determine the acceptability status of an argument is not enough: it is also important to determine whether the defeaters are defeated themselves. An argument B *defeats* argument A if and only if B is a reason against A. If the internal structure of arguments and the reasons why they defeat each other are not considered, an *abstract argumentation framework* (AAF) emerges [4]. An AAF is a pair $< Arg, attacks >$ where:

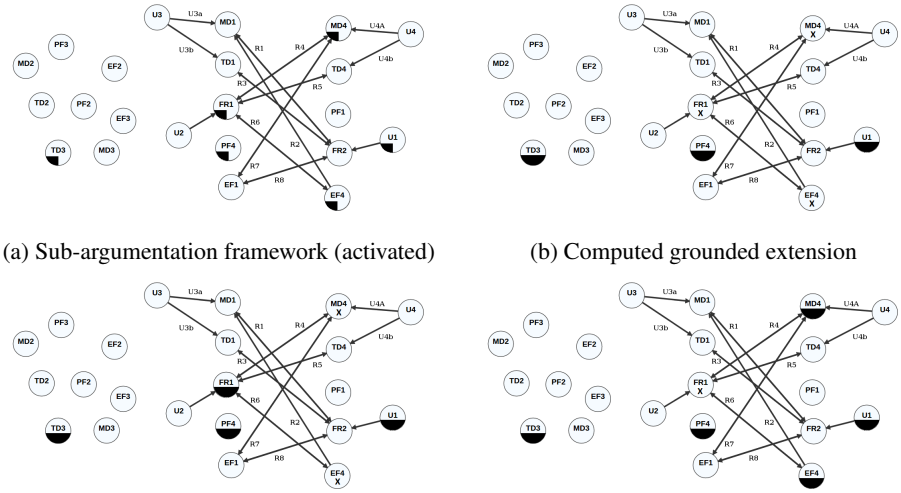
- Arg is a finite set of (abstract) *arguments*,
- $attacks \subseteq Arg \times Arg$ is binary relation over Arg.

Given sets $X, Y \subseteq Arg$ of arguments, X *attacks* Y if and only if there exists $x \in X$ and $y \in Y$ such that $(x, y) \in attacks$. The question is which arguments should ultimately be accepted and a formal criterion that determines it is needed. In the literature, this criterion is known as *semantics*: given an AAF, it specifies zero or more sets of acceptable arguments, called *extensions*. Various argument-based semantics have been proposed [1], but here the focus is on the preferred and grounded semantics proposed in [4]. A set $X \subseteq Arg$ of argument is

- *admissible* iff X does not attack itself and X attacks every set of arguments Y such that Y attacks X;
- *complete* iff X is admissible and X contains all arguments it *defends*, where X *defends* x if and only if X attacks all attacks agains x;
- *grounded* iff X is minimally complete (with respect to $\subseteq$);
- *preferred* iff X is maximally admissible (respect to $\subseteq$)

Grounded semantics always produce an unique extension, while preferred semantics can produce one or more extensions (conflict free set of arguments). In case just one extension is produced, this coincides with the grounded extension. However, in case multiple extensions are computed, a quantification of the credibility of each extension is needed. Here it is argued that the cardinality of an extension is an important factor: intuitively, an extension with a higher cardinality can be seen as more credible than extensions with lower cardinality as it contains more pieces of evidence that are consistent with each other. Figure 2 illustrates examples of grounded and preferred extensions emerged from layer 4 for a user who has answered the questions of the original $NASA - TLX$ (table 8).

Layer 5 - Accrual of acceptable arguments and computation of MWL: Given a set of extensions, the final step is to infer an index of mental workload. As defined before,



(a) Sub-argumentation framework (activated)

(b) Computed grounded extension

(c) Computed preferred extension 1

(d) Computed preferred extension 2

Fig. 3: Example of different extensions generated by the same input. White nodes have no status, quarter filled nodes are activated ,half filled nodes are accepted and X-nodes are rejected. Hypothetical set of input values: Mental demand: 77, Temporal demand: 53, Effort: 74, Performance: 81, Frustration: 23, Physical demand: not considered.

two typologies of arguments have been formalised: forecast and mitigating. A preferred or grounded extension can contain both types. However, just forecast arguments support a conclusion (workload level) that can be considered for the final inference of a single index of mental workload. The value v of a forecast argument is essentially a linear relationship from the range of the argument's premise to the range of the argument's conclusion[3]. Finally, mitigating arguments already played their role (through their attacks against other arguments), contributing to the computation of the acceptable extensions. Therefore, for each forecast argument in an extension, a value is computed and the final MWL index is given by the aggregation of these figures. This aggregation might be done in different ways, according to the designer's choice. For instance, one possible way is to calculate the average, however, other domains might require different calculations like the median, the weighted average or the sum. This study investigates the use of different accrual strategies: the average, the weighted average and the median of values v. For the weighted average, weights are the number of times the attribute on the premise of a forecast argument was chosen in the $NASA-TLX$ pairwise comparison process.

3.1 Knowledge bases

While figure 2 presents a possible translation of the $NASA-TLX$ into an argumentation framework, other translations can be designed. A second argumentation framework

[3] For instance if argument MD1 is activated, with reported mental demand equal 32, then $v_{MD1} = 32$, while if PF1 is activated, with reported performance equal 0, then $v_{PF1} = 100$.

using the $NASA-TLX$ attributes is suggested. In this case, it takes into consideration the paired comparison procedure, as part of the original $NASA-TLX$ instrument (section 2.2), to produce 14 new undercutting attacks listed in table 4 and depicted in figure 4.

Table 4: Undercutting attacks

U5: [FrustrationOrEffort = effort **&** effort ∈ [0, 32] **&** frustration ∈ [67, 100] ⇒ FR2]
U6: [FrustrationOrEffort = frustration **&** effort ∈ [0, 32] **&** frustration ∈ [67, 100] ⇒ EF1]
U7: [MentalOrEffort = effort **&** effort ∈ [0, 32] **&** mental demand ∈ [67, 100] ⇒ MD4]
U8: [MentalOrEffort = mental demand **&** effort ∈ [0, 32] **&** mental demand ∈ [67, 100] ⇒ EF1]
U9: [FrustrationOrEffort = effort **&** effort ∈ [67, 100] **&** frustration ∈ [0, 32] ⇒ FR1]
U10: [FrustrationOrEffort = frustration **&** effort ∈ [67, 100] **&** frustration ∈ [0, 32] ⇒ EF4]
U11: [TempOrFrustration = frustration **&** temporal ∈ [67, 100] **&** frustration ∈ [0, 32] ⇒ TD4]
U12: [TempOrFrustration = temporal **&** temporal ∈ [67, 100] **&** frustration ∈ [0, 32] ⇒ FR1]
U13: [FrustrationOrMental = frustration **&** frustration ∈ [0, 32] **&** mental ∈ [67, 100] ⇒ MD4]
U14: [FrustrationOrMental = mental **&** frustration ∈ [0, 32] **&** mental ∈ [67, 100] ⇒ FR1]
U15: [TempOrFrustration = frustration **&** temporal ∈ [0, 32] **&** frustration ∈ [67, 100] ⇒ TD1]
U16: [TempOrFrustration = temporal **&** temporal ∈ [0, 32] **&** frustration ∈ [67, 100] ⇒ FR2]
U17: [FrustrationOrMental = frustration **&** frustration ∈ [67, 100] **&** mental ∈ [0, 32] ⇒ MD1]
U18: [FrustrationOrMental = mental **&** frustration ∈ [67, 100] **&** mental ∈ [0, 32] ⇒ FR2]

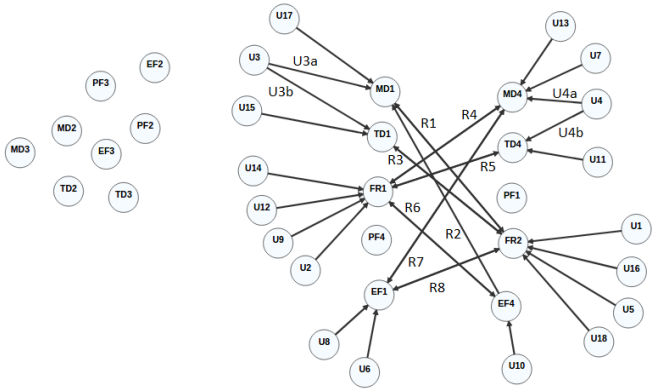


Fig. 4: Argumentation framework: graphical representation of a knowledge-base using the $NASA-TLX$ attributes and the information from the paired comparison procedure.

Finally, a third argumentation framework is also proposed using only the WP attributes (table 9). The nature of these attributes is more self-sufficient (in contrast with the $NASA-TLX$ attributes). In this case only forecast arguments are designed (table 5 and graphical representation in figure 5). The reason to propose this knowledge-base is to demonstrate that the 5-layer schema described in section 3, can be used even in the absence of conflicting arguments. While one of the biggest advantages of the system is lost, it remains the possibility to update the knowledge-base with new attributes and arguments or carry out the accrual of accepted arguments in different fashions.

Table 5: Forecast arguments using WP attributes. Speech response is used as an example while the same principle applies to the attributes manual response (MR), solving and deciding (SD), auditory resources (AR), visual resources (VR), selection of response (SR), task and space (TS) and verbal material (VM).

SP1: [speech response $\in$ [0, 32] $\rightarrow U$]	SP3: [speech response $\in$ [50, 66] $\rightarrow F^{+}$]
SP2: [speech response $\in$ [33, 49] $\rightarrow F^{-}$]	SP4: [speech response $\in$ [67, 100] $\rightarrow O$]

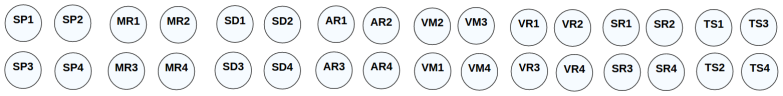


Fig. 5: Argumentation framework: graphical representation of a knowledge-base using WP attributes and forecast arguments of table 5. Arguments have no conflicts.

3.2 Experiment set up and dataset

Different knowledge-bases, computational semantics and accrual strategies were employed in this study, making use of the 5-layer schema of section 3 . Four different configurations were created, as listed in table 6.

Using the $NASA-TLX$ and the WP subjective scales, the mental workload experienced by master and PhD students in different teaching sessions was measured and their attributes used to elicit models described in table 6. Three different topics of the module "research methods" in the school of computing, at Dublin Institute of Technology, were evaluated in different semesters during the period 2015-2017. These topics were: 'Science', 'The Scientific method', 'Research planning' and 'Literature review'. Three delivery methods were used across different teaching sessions:

- A) a one-way traditional lecture delivered by means of slides.
- B) a one-way multimedia lecture delivered by means of videos.
- C) a one-way multimedia lecture (B) followed by a collaborative activity where students had printed handouts of viewed content (A).

The number of students of each task can be seen in table 7. The average time for the 'Science' lectures for each delivery method were: (A) 61 minutes, (B) 18 minutes (exact time) and (C) 58 minutes. Following that the average time for 'The Scientific method', 'Research planning' and 'Literature review' lectures were respectively: (A) 46, 54 and 64 minutes, (B) 28, 11 and 19 minutes (exact times), (C) 50, 79, and 77 minutes. The students were from 16 different nationalities and their age was in the range [22, 74] (average 33.7 and standard deviation of 7.3 years). Students were asked to fill in questionnaires associated to the two mental workload assessment instruments: the Nasa Task Load Index ($NASA-TLX$) and the Workload Profile (WP). Each questionnaire had also a scale in the range $[0..100] \in \aleph$ in which the student had to inform what, in his opinion, was the MWL imposed by the task (table 10 in the appendices). The dataset had missing values for individual columns of the pairwise variables in the $NASA-TLX$ measurement. Data imputation was used to estimate the missing values. The imputation method used logistic regression, in line with other studies [2,25].

Table 6: Set up of each model investigated. For each step the respective layer L is indicated as in section 3. Since attack relations are binary layer 3 was not listed.

Model	**Attributes** (*L1*)	**AAF** (*L2*)	**Semantics** (*L4*)		**Accrual** (*L5*)
			Grounded	**Preferred**	
$N1$	*TLX* (no preferences)	figure 2	✓		Average
$N2$	*TLX* (no preferences)	figure 2		✓	Average
$N3$	*TLX*	figure 4	✓		Weighted average
$W1$	*WP*	figure 5	✓		Median

Table 7: Number of students who answered the $NASA-TLX$, WP and self report questionnaires. In total there were 12 tasks divided by content and knowledge transmission approach.

Topic	***NASA-TLX*** **Delivery method**			***WP*** **Delivery method**		
	A	**B**	**C**	**A**	**B**	**C**
Science	14	13	9	17	13	7
Scientific method	10	18	10	13	18	8
Research planning	11	22	6	9	22	3
Literature review	10	13	9	11	11	7

4 Results and Findings

Collected answers from the questionnaire of the two MWL instruments (tables 8 and 9) were used to elicit the argumentation frameworks of each designed model (table 6). The distributions of the MWL scores produced by defeasible models were correlated against the ones produced by the $NASA-TLX$ and the WP to test their convergent validity (definition in section 2.1). Additionally, the MWL indexes produced by the designed defeasible models and the two baseline models were correlated against the self-reported MWL scores (table 10), to test face validity (definition in section 2.1). Self-report scores are referred to as *SR-MWL*. In order to select the most appropriate correlation statistic, a test of the normality of the MWL indexes distributions generated by all models and self-reported scores, was performed using the Shapiro-Wilk test. The significance was reported as following: $N1-3$, WP and $NASA-TLX > 0.05$, $W1$ and $SR-MWL < 0.01$. This indicates that Spearman should be applied for correlations of $W1$ and $SR-MWL$ scores, while Pearson should be applied for the others. Results are depicted in figure 6.

4.1 Face validity

In line with other studies [22] face validity was assessed using correlation coefficients. From figure 6 it is possible to observe that the baseline instruments ($NASA-TLX$, WP), presented a moderated correlation with the $SR-MWL$ scores, 0.46 and 0.412 respectively. The results for the designed defeasible models ($N1-N3$, $W1$) are similar, indicating that the inferential capacity of these is approximately the same of the state-of-the-art MWL measurement techniques. Finally, it is important to highlight the small

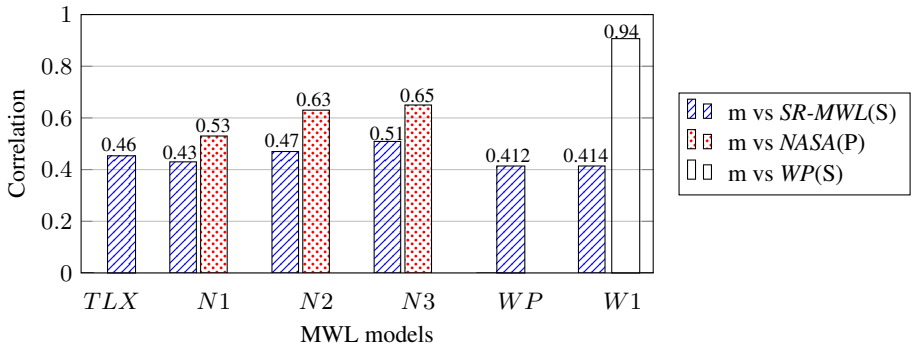


Fig. 6: Face validity and convergent validity of the defeasible mental workload models. $p < 0.05$. $S = Spearman$, $P = Pearson$, $m = model$.

improvement of model $N3$ when compared to $N1$ and $N2$. It demonstrates the positive effect of including additional information provided by the paired comparison procedure, as done in layers 2 (new undercutting attacks) and 5 (weights to arguments).

4.2 Convergent validity

According to definition in section 2.1, convergent validity is aimed at demonstrating whether a model correlates with other models of MWL. From figure 6 it is possible to observe that the defeasible models $N1-3$ achieved a moderate to high correlation with $NASA-TLX$ (correlation coefficient between 0.53 and 0.65), with the highest one achieved by $N3$. Since $N3$ make use of additional information provided by the paired comparison procedure a higher correlation with $NASA-TLX$ was expected. As for WP and $W1$ it is possible to observe an extremely high correlation (0.94 coefficient), indicating that the only difference between the two techniques (in the accrual stage, layer 5) did not had significant impact on the MWL assessments.

4.3 Summary of findings

Analysis of face validity indicates that the generated inferences, by defeasible models ($N1-N3$), are in line with the subjective interpretation of MWL by respondents. Differences between models $N1$ and $N2$ point out preferred semantics as more relevant for the definition of the dialectical status of arguments in contrast with the grounded semantics. Preferred extensions take a credulous view on which arguments can be accepted while grounded extensions provide a skeptical view. This difference illustrates how the credulous view is more in line with the perceived MWL by participants of the study, according to the correlation coefficients obtained for face validity. Defeasible models also achieved a moderate-high correlation for convergent validity. It demonstrates that defeasible reasoning was capable of assessing MWL even with the same or less attributes than state-of-the-art MWL subjective measurement techniques. For instance, models $N1$ and $N2$ make use of only 5 attributes, while $NASA-TLX$ makes

use of 6 attributes plus 15 preferences among attributes. Furthermore, baseline instruments are static and difficult to be updated. The $NASA-TLX$ pair-wise comparison indeed is a basic form of reasoning, giving importance to each considered attribute, but it has problems in the case another dimension, believed to be useful for assessing MWL, needs to be added. As for WP, it is a simple sum, so no form of reasoning is involved. Certainly, it can be easily updated adding dimensions and summing all of them. However, its defeasible counterpart can offer a reasoning chain that can be better described, since it uses the language of arguments. This self-explanatory capacity is in line with some of the appealing properties of AT as suggested in the literature [14]. Eventually, the proposed 5-layer schema (section 3) allows the comparison of different knowledge-bases and it seems more appealing and dynamic when compared to fixed formulas used within the $NASA-TLX$ and the WP models.

5 Conclusion and future work

This research investigated the use of defeasible reasoning to represent and assess mental workload (MWL). Two well-known subjective MWL assessment techniques, namely the NASA Task Load Index and the Workload Profile, were selected as baseline instruments. A formal 5-layer schema, as emerged in literature, has been used for reasoning defeasibly. In details, defeasible models of MWL has been constructed fully or partially using the information carried in the baseline instruments. A user study has been conducted in an educational context to assess the mental workload imposed by different teaching methods on students. Questionnaires were used to gather data from students, which served as the input for the baseline and the defeasible models. In turn, the inferences produced by these models were methodically compared against each other and against a self reported MWL score provided by students. This comparison was aimed at investigating the face validity and the convergent validity of generated inferences. Findings suggest that defeasible reasoning is a promising avenue for the assessment of MWL. It offers a similar or better inferential capacity than the baselines models, even in the presence of partial information. It also allows the comparison of different knowledge-bases of MWL designers, supporting MWL research. The 5-layer schema support MWL practitioners to represent their knowledge and to perform inference under uncertainty by using a language closer to the way they usually reason. Future work will be focused on the replication of the approach adopted in this empirical study by employing other knowledge-bases provided by other MWL designers. Different practical domains will be considered in order to gauge the capability of discriminating significant variations in MWL.

Acknowledgments

Lucas Middeldorf Rizzo acknowledges CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for his Science Without Borders scholarship n.232822/2014.0.

References

1. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. The Knowledge Engineering Review 26(4), 365–410 (November 2011)
2. Brand, J.: Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Ph.D. thesis, University Rotterdam (1999)
3. Cain, B.: A review of the mental workload literature. Tech. rep., Defence research and development Toronto (Canada) (2007)
4. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence 77(2), 321–358 (1995)
5. Eggemeier, F.T.: Properties of workload assessment techniques. Advances in Psychology 52, 41–62 (1988)
6. García, D.C.M.A.J., Simari, G.R.: Strong and weak forms of abstract argument defense. Computational Models of Argument: Proceedings of COMMA 2008 172, 216 (2008)
7. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology 52(C), 139–183 (1988)
8. Longo, L.: Human-Computer Interaction and Human Mental Workload: Assessing Cognitive Engagement in the World Wide Web, pp. 402–405. Springer (2011)
9. Longo, L.: Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In: UMAP, pp. 369–373. Springer (2012)
10. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behaviour and Information Technology 34(8), 758–786 (2015)
11. Longo, L.: Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). pp. 106–111 (June 2016)
12. Longo, L., Dondio, P.: Defeasible reasoning and argument-based systems in medical fields: An informal overview. In: Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on. pp. 376–381. IEEE (2014)
13. Longo, L., Dondio, P.: On the relationship between perception of usability and subjective mental workload of web interfaces. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM Int. Conf. on. vol. 1, pp. 345–352. IEEE (2015)
14. Longo, L., Hederman, L.: Argumentation Theory for Decision Support in Health-Care: A Comparison with Machine Learning, pp. 168–180. Springer, Cham (2013)
15. Matt, P.A., Morgem, M., Toni, F.: Combining statistics and arguments to compute trust. In: 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada. vol. 1, pp. 209–216. ACM (May 2010)
16. Moustafa, K., Luz, S., Longo, L.: Assessment of Mental Workload: A Comparison of Machine Learning Methods and Subjective Assessment Techniques, pp. 30–50. Springer (2017)
17. O'Donnell, R., Eggemeier, F.: Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas. John Wiley and Sons, Inc (1986)
18. Pollock, J.L.: Knowledge and justification. Princeton University press, Princeton (1974)
19. Pollock, J.L.: Defeasible reasoning. Cognitive Science 11(4), 481–518 (1987)
20. Reid, G.B., Nygren, T.E.: The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. Adv. Psychol. 52(C), 185–218 (1988)
21. Rizzo, L., Dondio, P., Delany, S.J., Longo, L.: Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile, pp. 215–229. Springer International Publishing, Cham (2016)

22. Rubio, S., Díaz, E., Martín, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. Applied Psychology 53(1), 61–86 (2004)
23. Tracy, J.P., Albers, M.J., Stc: Measuring cognitive load to test the usability of web sites. STC's 53rd Annu. Conf. Proc. 2005 pp. 256–260 (2006)
24. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
25. White, I.R., Daniel, R., Royston, P.: Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational Statistics and Data Analysis 54(10), 2267 – 2275 (2010)
26. Wickens, C.D.: Processing resources and attention. Multiple-task performance pp. 3–34 (1991)
27. Young, M.S., Stanton, N.A.: Mental workload: theory, measurement, and application. In: Karwowski, W. (ed.) International encyclopedia of ergonomics and human factors, vol. 1, pp. 818–821. Taylor & Francis, 2nd edn. (2006)

Appendix

Questionnaires

Table 8: The questionnaire of 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 9: The questionnaire of the Workload Profile

Dimension	**Question**
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)?
Solving and deciding	How much attention was required for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognizing and identifying objects)?

Table 10: The questionnaire for MWL self assessment

Dimension	**Question**
Self report	What is in your opinion the mental workload imposed by the performed task?