2015-01-30

A defeasible reasoning framework for human mental workload representation and assessment

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Luca Longo (2015) A defeasible reasoning framework for human mental workload representation and assessment, Behaviour & Information Technology, 34:8, 758-786, DOI: 10.1080/0144929X.2015.1015166

Behaviour & Information Technology, 2015
http://dx.doi.org/10.1080/0144929X.2015.1015166

Taylor & Francis
Taylor & Francis Group

A defeasible reasoning framework for human mental workload representation and assessment

Luca Longo*

School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland; School of Computing, Dublin Institute of Technology, Dublin, Ireland

(*Received 4 January 2014; accepted 30 January 2015*)

Human mental workload (MWL) has gained importance in the last few decades as an important design concept. It is a multifaceted complex construct mainly applied in cognitive sciences and has been defined in many different ways. Although measuring MWL has potential advantages in interaction and interface design, its formalisation as an operational and computational construct has not sufficiently been addressed. This research contributes to the body of knowledge by providing an extensible framework built upon defeasible reasoning, and implemented with argumentation theory (AT), in which MWL can be better defined, measured, analysed, explained and applied in different human–computer interactive contexts. User studies have demonstrated how a particular instance of this framework outperformed state-of-the-art subjective MWL assessment techniques in terms of sensitivity, diagnosticity and validity. This in turn encourages further application of defeasible AT for enhancing the representation of MWL and improving the quality of its assessment.

Keywords: human mental workload; defeasible reasoning; argumentation theory; knowledge-representation

1. Introduction

Human mental workload (MWL) is gaining momentum as an important design concept in human–computer interaction (HCI) and is important in considering the interaction of people with computers and other technological devices. It has been extensively documented that both mental overload and underload can negatively affect performance (Xie and Salvendy 2000). At a low level of MWL, people may often experience annoyance and frustration when processing information. On the other hand, a high level can also be both problematic and even dangerous, as it leads to confusion, decreases performance in information processing and increases the chances of errors and mistakes. 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 so that design alternatives can be evaluated (Hart 2006). A wide range of ad hoc definitions of MWL can be found in the literature. 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. However, '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' (Hart and Staveland 1988). Although MLW has been extensively applied in the human factors community for the theoretical advantages it provides in interaction and interface design, its formalisation as an operational and computational construct has not sufficiently been addressed. Many researchers agree that there are too many ad hoc computational models and definitions in the literature and their use by MWL designers has been subjective, limiting both their application in different contexts and the ease with which they may be compared. This study is part of a larger research project (Longo 2011, 2012, 2014) and it introduces a novel computational framework for representing and assessing MWL based on defeasible reasoning (DR). The starting analysis is an investigation of the nature of MWL as a defeasible phenomenon – a concept built upon a set of reasons that can be defeated by providing additional reasons. The word 'defeasible' is inherited from DR and is 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. 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 DR activity (Baroni, Guida, and Mussi 1997). Formally, state-of-the-art models of DR are implemented using argumentation theory (AT). This systematically studies how reasons, built upon the notions of argument and logical consequence, can

*Email: longo.luca@gmail.com

be sustained or discarded in a reasoning process, and the validity of the conclusions reached. AT has been chosen as the knowledge representation tool for MWL because of its capacity to deal with inconsistent and incomplete knowledge that can be captured more intuitively by employing the notion of arguments (Longo, Kane, and Hederman 2012a). AT captures expertise in an organised fashion and is a promising tool for handling the uncertainty and vagueness associated with the representation of MWL. AT can lead to explanatory reasoning and is a paradigm that allows the representation of a construct to be retracted and updated with additional knowledge (Longo and Dondio 2014). Eventually, a knowledge base built upon AT can be elicited without requiring a complete data set, in contrast with machine and other learning-based techniques. This is particularly useful for unstructured knowledge, where quantitative evidence has not yet been gathered (Longo and Hederman 2013). Since MWL may be seen as a defeasible phenomenon, AT may have a positive impact on its representation and assessment. MWL can be captured, analysed and measured in ways that increase its understanding, allowing it to be used for practical activities. The research question being investigated is: *can the representation of MWL and the quality of its assessment be improved by defeasible AT?*

In order to investigate this research problem, the paper is organised as follows. Related works are introduced describing measurement techniques, definitions and computational strategies employed by state-of-the-art models of MWL. A summary of the core tenets for representing and assessing MWL follows, highlighting the limitations of the current solutions which have been reviewed; this will provide the motivation for the research question being investigated. The key notions behind DR and the reasons why MWL can be seen as a defeasible phenomenon are provided. A brief illustration is given of the main components employed by AT and used for implementing DR in practice; in turn, a formal framework based upon these components is designed. The proposed framework is subsequently used to build representational instances of mental workload. These are then evaluated through user studies by comparing the properties of sensitivity, diagnosticity and validity against two well-known subjective MWL assessment techniques (the NASA Task Load Index and the Workload Profile, WP) in the context of human–web interaction (HWI). A summary of the main contributions concludes this paper, along with some proposals for future research.

2. Related work

The concept of MWL has a long history in the fields of ergonomics and psychology, with several applications in the aviation (Hart and Staveland 1988; Hart 2006) and automotive industries (De Waard 1996). Although it has been studied for the last four decades, no clear definition of MWL has emerged that has a general validity and that is universally accepted (Cain 2007). The main reason for assessing MWL is to measure the mental cost of performing a certain task with the goal of predicting operator and system performance (Cain 2007). MWL is an important design criterion: at an early system design phase, not only can the system/interface be optimised to take workload into consideration, but MWL can also guide designers in making appropriate structural changes (Xie and Salvendy 2000). Modern technologies such as web applications have become increasingly complex (Longo et al. 2012b), with increments in the degree of MWL imposed on operators (Gwizdka 2009, 2010). The assumption in design approaches is that as the difficulty of a task increases, perhaps due to interface complexity, MWL also increases and performance usually decreases (Cain 2007). In turn, errors are more frequent, there are longer response times, and fewer tasks are completed per time unit (Huey and Wickens 1993). On the other hand, when task difficulty is negligible, systems can impose a low MWL on operators: this should be avoided as it leads to difficulties in maintaining attention and increasing reaction time (Cain 2007). In the following sections it is shown how MWL can be measured and which techniques have so far been employed to aggregate heterogeneous measures towards an index of workload. This review of current solutions is aimed at identifying both reasons why a more generally applicable solution has not yet been developed, and the key characteristics of MWL representation and assessment.

2.1. *Measures of MWL*

The measurement of MWL is a vast and heterogeneous topic as the 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 multidimensional property that characterises MWL (Longo and Barrett 2010b). Several reviews attempted to collate the enormous amount of knowledge behind measurement procedures. According to Gopher and Donchin (1986), measurements can be divided into subjective measures, performance measures, arousal measures, specific measures and psychophysiological measures. Young and Stanton (2006) proposed three broader classes of measures: primary and secondary task measures, physiological measures and subjective measures. This is also supported by O'Donnell and Eggemeier (1986) and Wickens and Hollands (1999). Tsang and Vidulich (2006) proposed four categories: performance, subjective, physiological measures and multiple measures of workload. Xie and Salvendy (2000) introduced a further classification based on empirical and analytical methods.

In general, the measurement techniques which have emerged in the research can be classified into three broad categories (Young and Stanton 2004; Tsang 2006; Tsang and Vidulich 2006; Wilson and Eggemeier 2006; Cain 2007):

- *self-assessment measures*, including self-report measures and subjective rating scales;
- *performance measures*, which consider both primary and secondary task measures;
- *physiological measures*, which are derived from the physiology of the operator.

The class of *self-report measures* is often referred to as subjective measures. This category is obtained from the direct estimation of task difficulty from subjects and it relies on the subjective perceived experience of the interaction operator–system. Subjective measures have always appealed to many workload practitioners and researchers because it is strongly believed that only the person concerned with the task can provide an accurate and precise judgement with respect to the MWL experienced. Various dimensions and attributes of MWL are considered in self-report measures. These include effort, performance, as well as individual differences such as the emotional state, attitude and motivation of the operator (De Waard 1996). The class of subjective measures includes multidimensional approaches such as the NASA's task load index (NASATLX) (Hart and Staveland 1988), the subjective workload assessment technique (SWAT) (Reid and Nygren 1988), the WP (Tsang and Velazquez 1996) as well as unidimensional approaches such as the Cooper–Harper scale (Cooper and Harper 1969), the rating scale mental effort (Zijlstra 1993), the subjective workload dominance technique (Vidulich and Ward Frederic 1991) and the Bedford scale (Roscoe and Ellis 1990).

The class of *performance measures* assumes that mental workload practitioners and, more generally, system designers are typically concerned with the performance of their systems and technologies. The assumption is that the MWL of an operator when interacting with a system acquires importance only if it influences system performance. As a consequence, it is believed that this class of techniques is the most valuable options for designers (Tsang and Vidulich 2006). According to different reviews (Cain 2007; O'Donnell and Eggemeier 1986; Wickens and Hollands 1999; Young and Stanton 2004; Wilson and Eggemeier 2006), performance measures can be classified into two sub-categories: primary task and secondary task measures. In primary task methods, the performance of the operator is monitored and analysed according to changes in primary task demands. Examples of common measurement parameters are response and reaction time, accuracy and error rate, speed and signal detection performance, estimation time and tapping regularity (Tsang and Vidulich 2006). In secondary task assessment procedures, there are two tasks involved and the performance of the secondary task may not have practical importance, but rather may serve to load or to measure the MWL of the operator performing the primary task (O'Donnell and Eggemeier 1986).

The class of *physiological measures* includes bodily responses derived from the operator's physiology, and relies on the assumption that they correlate with MWL. They are aimed at interpreting psychological processes by analysing their effect on the state of the body, rather than measuring task performance or perceptual subjective ratings. The principal reason for adopting physiological measures is because they do not require an overt response by the operator and they can be collected continuously, within an interval of time, representing an objective way of measuring the operator state (O'Donnell and Eggemeier 1986).

2.1.1. *Advantages and disadvantages of measurement techniques*

Each typology of measurement technique has its own advantages and disadvantages and is suitable for different contexts to different extents. Several criteria exist and have been proposed as guidelines for selecting and developing techniques (O'Donnell 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 in that it must be capable of indicating the sources that cause variations in workload and of quantifying the contributions by type or resource demand;
- *intrusiveness*: the methodology should not be intrusive or interfere with the performance of the task of the operator, becoming a source of workload itself (this property is referred to as *obtrusiveness* by Wickens and Hollands (1999, Chapter 11));
- *requirements*: the methodology should require the minimum possible equipment to avoid influencing the operator's performance. Muckler and Seven (1992) refer to this as *resource requirements*;
- *acceptability*: the method should have a high level of operator acceptance, showing at least face validity,[1] without being onerous. Muckler and Seven (1992) refer to this as *relative simplicity*;
- *selectivity*: the method should be selectively sensitive to differences in resource demand and not to changes in factors unrelated to MWL (Wickens and Hollands 1999, Chapter 11);
- *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 (Wickens and Hollands 1999, Chapter 11). Wierwille and F. (1993) and Muckler and Seven (1992) refer to this as *transferability* and *sufficient reliability*, respectively, highlighting the importance of the capability of a technique to be used in different applications.

Subjective measures are in general easy to administer and analyse. They provide an index of perceived strain and multidimensional measures can determine the source of MWL. However, the main drawback is that they can only be administered post-task, thus influencing the reliability for long tasks. In addition, meta-cognitive limitations can diminish the accuracy of reporting and it is difficult to perform comparisons among raters on an absolute scale. However, they appear to be the most appropriate types of measurement for assessing MWL because they have demonstrated high levels of 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 MWL. They are capable of discriminating individual differences in resource competition. However, the main limitation is that they cannot distinguish performance of multiple tasks that are executed simultaneously by an operator. If taken in isolation, they do not represent reliable measures, though if used in conjunction with other measures, such as subjective ratings, they can be useful. Secondary task measures 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. However, they are only sensitive to large changes in MWL and they might be highly intrusive, influencing the behaviours of users while interacting with the primary task.

Physiological measures are extremely good at monitoring data at a continuous interval, thus having high measurement sensitivity. They do not interfere with the performance in the primary task. However, the main drawback is that they can be easily confounded by external interference. Moreover, they require equipment and tools that are often physically obtrusive and the analysis of data is complex, requiring well-trained experts.

2.2. Aggregation strategies and computational aspects

As has been seen, several MWL measures exist, showing how non-trivial the measurement problem is. For unidimensional procedures, intuitively, the only dimension does not need to be aggregated with any other dimension. However, for multidimensional procedures, there are issues of how to uniformly represent those attributes believed to influence MWL and how to aggregate them towards a representative meaningful index that can be employed in practice. In the NASATLX (Hart 2006), for example, subjective ratings are expressed as natural numbers within the range 0–100, while in the SWAT (Reid and Nygren 1988), they are expressed as natural numbers within the discrete range 1–3. These ranges and scales are commonly adopted, but they are not the only choices for expressing a subjective judgement of a rater. For example, Moray (Neville et al. 1988) has proposed the use of fuzzy sets,[2] borrowed from Fuzzy set theory (Zadeh 1965) as a means to express judgements in a qualitative way but at the same time formalising the use of verbal judgements. In Longo and Barrett (2010a,b), the authors attempted to propose an ad hoc formalisation of various attributes believed to influence MWL. Some of these were modelled as natural numbers, others ranged from negative to positive numbers and others were designed as a taxonomy of sub-factors organised as a unidirectional tree, where leaf nodes represent subjective judgements and internal nodes indicate aggregation clusters. Clearly, different scales, non-uniform attributes and different aggregation strategies are difficult to share and employ across different MWL assessment techniques. This issue can be observed by describing the computational techniques adopted by some subjective multidimensional workload assessment procedures, as in the following sections.

2.2.1. Simple aggregation

In the WP assessment procedure (Tsang and Velazquez 1996), the accounted workload dimensions are based upon the multiple resource theory proposed in Wickens and Hollands (1999) and Wickens (2008). Each dimension is quantified through subjective rates (question) and subjects, after task completion, are required to rate the proportion of attentional resources used for performing a given task with a value in the continuous range 0–1. 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. The questions behind the WP model are the ones in Appendix A5 (6–13) and the aggregation strategy employed is relatively simple as it only sums these 8 rates d provided by a subject:

$$\mathrm{MWL_{WP}} : [0\ldots8] \in \Re, \quad \mathrm{MWL_{WP}} = \sum_{i=1}^{8} d_i.$$

Although this aggregation method is intuitive and simple, it implies that each dimension has the same strength in affecting overall MWL. Additionally, it does not consider external factors affecting the execution of the task, the state of the operator and his/her previous knowledge of the task being executed.

2.2.2. Weighted aggregation and preferences

In the NASATLX instrument (Hart 2006), the combination of the factors believed to influence MWL is not based on

a simple sum, rather on a weighted average. Each factor is quantified with a subjective judgement (questions 1–5 in Appendix A5 plus a further question related to physical demand) whose weight is computed via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial coefficient) of the six factors employed by the procedure, '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 (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 MWL score is computed as a weighed average, considering the subjective rating of each attribute d_i (for the 6 dimensions) and the correspondent weights w_i:

$$\mathrm{MWL}_{\mathrm{NASATLX}} : [0 \ldots 1] \in \Re,$$

$$\mathrm{MWL}_{\mathrm{NASATLX}} = \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15}.$$

The procedure is probably the most adopted because of its ease of application. However, the main issue associated with this aggregation approach is that, in the case a new dimension has to be added, the paired comparison procedure will become more tedious, as requiring more judgements by subjects. With only 9 or 10 dimensions, the comparisons required are respectively 36 and 45, which can be too cumbersome for an operator to be performed. This issue has been acknowledged by various authors who have proposed a modified version of the NASATLX (Thomas 1991).

2.2.3. *Ranking-based and correlation-based aggregation*

In the SWAT (Reid and Nygren 1988), three workload attributes (time, effort and stress) are modelled using discrete numbers in the range 1–3. Each number has an associated description. 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 completing the card sort procedure is to generate data that are used to produce 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 to determine the degree of agreement among the participants (raters), for a certain experiment on a given task. In this step, Kendall's coefficient of concordance (W) is employed, a non-parametric statistic 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 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, and 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 individual differences have to be accounted. Thus, for instance, when $W < 0.75$, homogeneous subgroup scales can be developed. In the original SWAT (Reid and Nygren 1988), the authors have developed six hypothetical orderings, based on the relative importance of each attribute. The subsequent step consists in the application of the Spearman correlation coefficient between the sorting provided by the subject and the hypothetical ordering. This is aimed at deciding which of the six subgroups is more suitable, meaning which group a subject belongs to. 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.

$$\mathrm{MWL}_{\mathrm{SWAT}} : [0 \ldots 100] \in \aleph.$$

As it is possible to note, SWAT relies on a very cumbersome and tedious procedure for subjects to obtain the workload ratings. Although it has been demonstrated that it has high diagnosticity and content validity (Rubio et al. 2004; Vidulich and Tsang 1986), the procedure is not straightforward to understand even by different MWL designers. Eventually, it is rather an ad hoc model that cannot be easily expanded with additional dimensions believed to affect MWL.

2.2.4. Ad hoc *aggregations and frameworks*

Hancock and Chignell (1988) employed the construct of MWL as a means for investigating the capability of operators interacting with machine through interfaces. Their theoretical formulation of MWL includes the notions of skill of operators, the time pressure they are exposed to and

the effort exerted for the execution of the task. Being psychology their main research field, the authors were inspired by the proposal of a computational model that included a power function to represent and assess MWL, formalism widely applied for fitting psychological data. Their approximation of overall workload may be described by the following formula:

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

where $\mathrm{MWL_{HC}}$ 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 skill degree. The issues associated with this formulation of workload are various. Firstly, as also agreed by the authors (Hancock and Chignell 1988), the use of the function does not solve the problem of workload assessment as 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 hard to be expanded if further factors are considered. Eventually, it does not account for the potential interactions that might occur between workload factors and their theoretical relationships. Other ad hoc solutions have been presented in Longo and Barrett (2010a,b). However, the aggregation strategies described so far are sufficient for having an almost complete panoramic view of all the possible techniques for MWL representation and assessment.

3. Discussion on modelling human MWL

The general consensus is that any single definition and assessment procedure is not capable of providing full information and entirely describing MWL (Xie and Salvendy 2000). As a construct, MWL is certainly complex and multidimensional and its assessment is not straightforward. Several definitions have been proposed by various researchers from different backgrounds and influences, and they seem to be intuitively appealing. 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, beliefs and choices, or simply driven by intuition. To further complicate matters, each assessment technique aggregates attributes differently, employing different scales, weights or ad hoc computational techniques. Workload attributes can be static or dynamic, reflecting MWL over a period of time or at a single moment. In addition, these attributes might be related or at least may not always be totally independent of each other. These relationships can be theoretical, such as that between the attributes 'demand' and 'performance' (O'Donnell and Eggemeier 1986), or empirically demonstrated, such as the U-shaped relationship between 'arousal' and 'performance' (Yerkes and Dodson 1908). However, none of the current assessment techniques includes a strategy of handling these theoretical relationships and the inconsistencies that might emerge from their interaction. According to Annett (2002), the validity of individual attributes, and more generally complex constructs, lies especially in their relationships with the other attributes of interest in the context of a specific situation. The suggestion is 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 behavioural or physiological, subjective or objective as well as the expression of intentions or opinions. A measure is rarely valid in isolation, but rather it gains validity 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. However, if the interaction of attributes is acknowledged by a given assessment technique, a method for resolving inconsistencies that might arise from their interaction is needed. To facilitate the understanding of MWL and the issues associated with its representation and assessment, we present a summary of the core tenets found in the literature.

- *Multidimensionality*: MWL is believed to be a multidimensional construct influenced by many factors with both static and dynamic properties (Hart and Staveland 1988; Cain 2007). Factors can relate to one of the limited expendable resources of human processing capacity and can be unbalanced during task execution, remaining unaffected, or becoming overloaded or underloaded (Wickens and Hollands 1999; Wickens 2008; Tsang and Velazquez 1996).
- *Hypotheticality*: MWL is believed to be a hypothetical construct. It cannot be detected directly, but only through the measurement/aggregation of some other factors believed to correlate highly with it (Gopher and Donchin 1986).
- *Context-awareness*: MWL is a context-aware construct that can be applied in single- or multi-task environments. A task might be affected by external factors or by other concurrent tasks (Eggemeier et al. 1991; Wilson and Eggemeier 1991; Xie and Salvendy 2000).
- *User-specificity*: MWL is a user-specific construct, influenced not only by factors such as previous task knowledge, skills and experience but also by a person's state, intentions and the effort he/she devotes to it (Hart and Staveland 1988).
- *Task-specificity*: MWL is task specific, influenced by factors such as task demands in terms of required cognitive resources, and objective or perceived task difficulty (Hart and Staveland 1988; Tsang and Velazquez 1996; Wickens 2008).

- *Relationality*: the dimensions considered within a MWL assessment technique might be related, monotonically or not, mitigating or enhancing others' strength (Cain 2007).
- *Preferentiality*: a factor considered within a MWL assessment technique might be preferred, thus having a greater influence on overall MWL (Reid and Nygren 1988; Hart 2006).
- *Subjectivity*: assessments of MWL are characterised by a degree of subjectivity. This refers to the design choices adopted for the development of an assessment procedure and the consideration of which factors to account for and how to aggregate them.
- *Uncertainty*: representing MWL is intrinsically uncertain. The selection and quantification of the factors believed to influence it are non-trivial problems. Selection is problematic because of disagreement among researchers on how to define MWL and how to measure it (Cain 2007). The latter refers to the accuracy of the measurement of each factor. In particular, in the case of subjective measures, quantifications are often made under uncertainty.
- *Partiality*: the quantification of the factors accounted for in a MWL assessment technique may be partial or incomplete. This mainly refers to objective measures (e.g. physiological) that can be incompletely gathered by the devices/sensors used. Additionally, factors might be correctly measured in a laboratory but only partially measured in a practical setting, thus invalidating the theoretical model in the event that it strictly requires them (Kramer, Sirevaag, and Braune 1987).
- *Computational aggregation*: the factors believed to influence MWL might be aggregated towards a single index employable for design purposes. Computational technique can be based on a simple sum Tsang and Velazquez (1996) or on a weighted average of factors (Hart and Staveland 1988; Hart 2006). Others can use ranking or correlation-based aggregation (Reid and Nygren 1988) or ad hoc, not-extensible formulas (Hancock and Chignell 1988).

The aforementioned tenets represent the starting points for the design of a new defeasible framework for MWL representation and assessment. *In this study it is argued that, in order to facilitate an understanding of the construct of MWL and its application, an extensible/open framework, able to handle several workload factors and their interrelationships and capable of resolving the potential inconsistencies that can derive from their interaction, is needed. This solution should account for the uncertainty that characterises the definition of each factor and provide clear aggregation semantics to merge these factors meaningfully*. This new perspective on MWL representation and assessment might breathe new life into this fascinating area of research.

4. MWL as a defeasible phenomenon

According to state-of-the-art research studies in the field, MWL is a complex multidimensional construct built upon a network of pieces of evidence. This network can vary according to the knowledge base of a workload designer elicited in a practical context. It is composed of those factors and their hypothetical or demonstrated relationships believed to be useful for assessing the MWL of a user performing a given task in a given context. Different MWL factors might support different and contradictory levels of MWL, creating inconsistent scenarios. To clarify these difficulties, let us consider an illustrative reasoning process that a designer might follow to assess the MWL imposed by a web-based interface on a skilled user after interacting with it and performing a given task.

The 'mental demand' of the task perceived by the user was poor, thus low MWL can be inferred. If this is the only evidence available, the majority of MWL designers would likely infer the same conclusion. However, if it is also known that 'interruptions' occurred during task execution, then the previous conclusion could be retracted inferring higher MWL. Yet, if it is also known that the user was highly skilled with respect to the task, additional evidence is now available and a lower degree of MWL could be inferred, retracting again the previous conclusion. Eventually, if the overall task 'performance' was perceived being poor, an inconsistency now arises and the conclusion could be retracted again to a higher degree of MWL because low performance is believed to be a sign of high workload. Although the task was not demanding and the user was skilled, external distractions might have played a significant role in increasing the completion time, minimising performance. The designer might eventually infer a relatively high degree of MWL because 'time' and 'distractions' are preferred over 'skill' and 'task complexity'.

From the aforementioned illustrative (and arguable) reasoning process, it is plausible to assume the following:

- *Assumption* 1: MWL is a complex construct built upon a network of pieces of evidence with different strength.
- *Assumption* 2: accounting the relationships of these pieces of evidence and resolving the inconsistencies arising from their interaction are essential in modelling MWL.

In formal logics, these assumptions are the key components of a *defeasible concept*: a concept built upon a set of interactive pieces of evidence, the reasons, that can become defeated by additional reasons. The term 'defeasible' comes from the multi-disciplinary fields of defeasible

reasoning (DR) aimed at studying the way humans reason under uncertainty and with contradictory and incomplete knowledge (Pollock 1987). 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 NMR because of the technical property (non-monotonicity) of the logical formalisms that are aimed at modelling DR activity (Baroni, Guida, and Mussi 1997). A computational implementation of DR is provided by AT, a new important multi-disciplinary topic in artificial intelligence that incorporates element of philosophy, psychology and sociology and that studies how people reason and express their arguments. It systematically investigates how arguments can be built, sustained or discarded in a DR process and the validity of the conclusions reached through resolutions of potential inconsistencies. AT has been proved useful for implementing DR activities and modelling complex constructs (Toni 2010). Argumentation systems are based upon the notion of argument and around an associated notion of logical consequence. This notion is monotonic: new information cannot invalidate existing arguments as constructed, but can only be responsible for the generation of new counterarguments. Argumentation systems are typically constructed upon an *underlying logical* language and are generally built on four layers (Longo and Dondio 2014):

(1) internal definition of arguments (monological structure),
(2) definition of an argumentation framework introducing conflicts (relations) among arguments (dialogical structure),
(3) validation of defeat (valid) relations among arguments (activation/elicitation of an argumentation framework),
(4) definition of the arguments' dialectical status (acceptability semantics for computation of arguments justifications).

The definition of an *argument* means internally assign a structure to it (Toulmin 1958). The focus is 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 (Bentahar, Moulin, and Bélanger 2010). The definition of *conflicts*, often replaced by the terms *attacks* or *counterarguments*, is aimed at connecting arguments in a dialogical structure. Dialogical models have driven argument-based approaches to be referred to as *DR 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 (Dung 1995). The validation of *defeats relations* refers to the identification of those conflicts that seem to be valid and credible. Eventually, the definition of the *dialectical status* is aimed at assigning a justification status to each argument. A fifth layer can be added and a final conclusion, claim or decision can be drawn by accruing arguments according to their status. In line to the multilayered schema, the following section is aimed at designing a defeasible framework for MWL representation and assessment. This framework incorporates the aforementioned fifth layer to produce a usable numerical index of MWL.

5. Design

In the following subsection, a framework for MWL representation and assessment is designed according to the multilayered schema of the previous section. Each layer is described in detail providing illustrative examples.

5.1. *Layer 1 – Definition of the monological structure of arguments*

The knowledge base of a designer in relation to MWL can be initially represented as a set of natural language propositions as in the following examples:

(1) *'the mental demand required by a task is linearly related to MWL: the higher the demand, the higher the MWL'*,
(2) *'given a low degree of performance there is a reason to believe the MWL exerted by a user on a given task is high'*,
(3) *'although the task can be highly mentally demanding, if the user devoted low effort and has a high degree of skills, there is a reason to believe MWL is low'*.

Each of the aforementioned proposition can be seen as an argument: a structure composed of a set of premises, each related to a given workload attribute (e.g. mental demand) and a conclusion derivable by applying an inference rule →.

Argument: premises → conclusion

Informally, possible translations of the earlier propositions into structured arguments might be

(1) a: Low mental demand → Underload,
b: Medium mental demand → Fitting load
c: High mental demand → Overload
(2) d: Low performance → Overload
(3) e: Low effort and high skill → Underload

Each premise can be seen as a piece of knowledge that alone or jointly to other premises is tentatively linked to a degree of MWL believed to be appropriate. In other words,

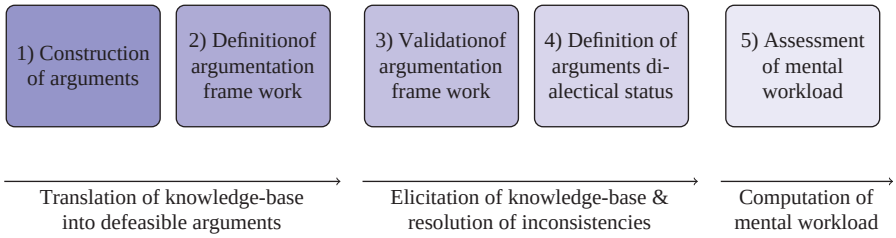

Figure 1. Multilayered argument-based framework for human MWL.

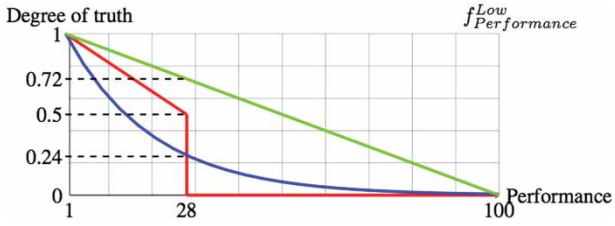

Figure 2. Possible membership functions for the fuzzy set 'Performance' and its fuzzy subset 'Low'.

a set of premises represents a set of reasons to believe that MWL is likely to fall within a certain region (underload, fitting load or overload). In order to computationally represent the vague linguistic terms associated with each premise of an argument (e.g. low or high) here the proposal is to use membership functions and the notion of *degree of truth* (Zadeh 1965).

DEFINITION 1 (Membership function) *For any set X, a membership function on X is any*

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

Membership functions on X represent fuzzy subsets of X. For an element x of X, the value f (x) is called the '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

$$\mathrm{MF}_X = \{f \mid f : X \rightarrow [0, 1] \in \Re\}.$$

Possible membership functions for describing the premise of the argument 'd' are shown in Figure 2. Each function, once elicited, produces a degree of truth as per Definition 1.

An argument might contain multiple premises; thus multiple degrees of truth can be produced. It turns out that in order to obtain a representative degree of truth from a set of premises of an argument, an aggregation strategy is needed. Here, the proposal is to average them. This representative value has to be tentatively linked to a conclusion, which means an estimation of MWL. In the literature, the overall spectrum of MWL is often separated by two *red lines*, theoretical thresholds that indicate when MWL is too low or too high (Wierwille and Eggemeier 1993; Colle and Reid 2005). Red lines (RL) can be set by a designer, according to his knowledge applied in a given context or they could be automatically learnt. They define three not overlapping regions of MWL: underload (U), fitting (optimal) load (F) and overload (O). Each region has a precise influence on user performance, attention and reaction time as highlighted in Figure 3.

In this study, the overall spectrum of MWL is proposed to be a number in the continuous range $[0\ldots100] \in \Re$; thus, the two RL are numbers lying within this range.

DEFINITION 2 (Redlines) RL^U_F, $\mathrm{RL}^F_O : [1\ldots100] \in \aleph$, *with* $0 < \mathrm{RL}^U_F < 50 < \mathrm{RL}^F_O < 100$.

With a method (average) for computing a representative degree of truth of the premises of an argument, now the goal is to formally model its conclusion. The proposal here is to use a function that, given a representative degree of truth of an argument's premises, infers one unique index of MWL. In formal terms, this function has to be a *strict monotonic function*.[3] Although it would be intuitive to design strict monotonic functions for the regions of underload and overload, this is not the case for the region of fitting workload. In fact, a designer might propose a triangular or a parabolic (symmetric) function with the top-peak

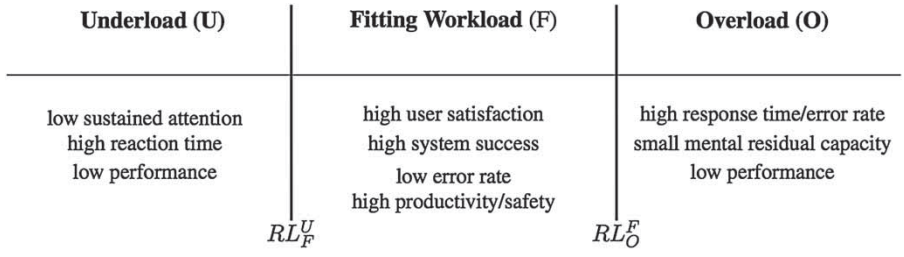

Figure 3. Disadvantages associated with low/high MWL and advantages of optimal workload.

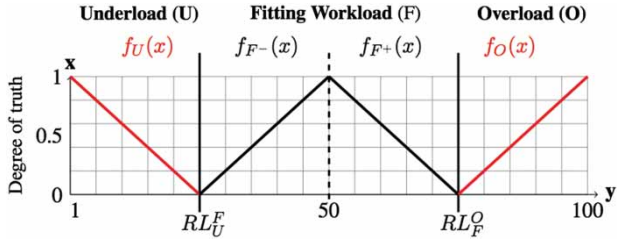

Figure 4. Workload spectrum separated into four regions by 4 illustrative dichotomies and by 2 RL.

around 50 and with two bottom-peaks in proximity of the two RL. However, this proposal violates the definition of a strict monotonic function because given one input, two outputs are possible. In order to solve this issue, it is reasonable to split the fitting workload region (F) into two sub-partitions (fitting workload lower F^- and upper F^+). Thus, the emerging configuration requires a MWL designer to define four functions, now on referred to as *workload dichotomies*[4] as in Figure 4.

DEFINITION 3 (Workload dichotomies) *A workload dichotomy is a strict monotonic function:*

$$f : [0\ldots1] \in \Re \rightarrow [0\ldots100] \in \Re$$

such that $\forall x,y$ *with* $x \leq y$, *then* $f(x) < f(y)$. *Four dichotomies are defined using* RL*:*

- $f_U : [0\ldots1] \rightarrow [0\ldots \mathrm{RL}_U^F)$
- $f_{F^-} : [0\ldots1] \rightarrow [\mathrm{RL}_U^F\ldots50)$
- $f_{F^+} : [0\ldots1] \rightarrow [50\ldots \mathrm{RL}_F^O]$
- $f_O : [0\ldots1] \rightarrow (\mathrm{RL}_F^O\ldots100]$

A possible definition of the workload dichotomies is shown in Figure 4.

Using membership functions for premises and dichotomies for conclusions, the argument has now a complete structure and can be formally defined.

DEFINITION 4 (Argument) *An argument* Ar *is a tentative inference* $\rightarrow$ *that links one or more premises* P_i *to a conclusion* C

$$\mathrm{Ar} : P_1, \ldots, P_n \rightarrow C,$$

where each premise P_i *is a membership function and the conclusion* C *is a workload dichotomy* f_C.

Eventually, to assess an index of MWL of an argument, the proposal is to use the argument's degree of truth (average of degrees of truth of its premises) as the input of the workload dichotomy associated with its conclusion.

DEFINITION 5 (Argument's degree of truth) *Given an argument* Ar, *its degree of truth* $\mathrm{Ar}_{\mathrm{deg}}$ *coincides with the average of the degrees of truth of its premises*

$$\mathrm{Ar}_{\mathrm{deg}} : \frac{1}{n} \sum_{P_i \in P_i,\ldots,P_n} P_i.$$

The average is an intuitive aggregation strategy that can be used for the unification of more premises; however, other approaches might be considered in the future, such as the fuzzy AND/OR unifications.

DEFINITION 6 (Argument's MWL) *Given an argument* Ar, *its degree of truth* $\mathrm{Ar}_{\mathrm{deg}}$ *and the associated workload dichotomy* f_C, *the MWL inferred from* Ar *is*

$$\mathrm{Ar}_{\mathrm{MWL}} = f_C(\mathrm{Ar}_{\mathrm{deg}}).$$

In order to bring clarity to the representation of the internal structure of arguments, and the MWL assessed by each of them, consider Example 1.

Example 1

- Ar: Low effort $\wedge$ high skill $\rightarrow$ underload
- Workload attributes: Effort (E), Skill (S)
- Possible inputs: $E = 25$, $S = 85$
- Premises: $f_E^{\mathrm{low}} \in \mathrm{MF}_E \wedge f_S^{\mathrm{high}} \in \mathrm{MF}_S$
- Conclusion: f_U
- $\mathrm{AR}_{\mathrm{deg}} = (f_E^{\mathrm{low}}(25) + f_S^{\mathrm{high}}(85))/2$
- $\mathrm{AR}_{\mathrm{MWL}} = f_U(\mathrm{AR}_{\mathrm{deg}})$

5.2. *Layer 2 – Definition of the dialogical structure of arguments*

Monological models, aimed at internally represent an argument are 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). According to a previous similar study (Matt, Morgem, and Toni 2010), arguments might be classified as

- *forecast* when in favour or against a certain claim (workload dichotomy), but justification is not infallible;
- *mitigating* when defeating forecast or other mitigating arguments, undermining their justification.

Forecast arguments are tentative defeasible inferences: they can be seen as justified claims concerning the expected or the anticipated behaviour of the target (MWL index). They represent hints or clues given by a designer under uncertainty and not mathematical proofs. The definition of forecast argument coincides with Definition 4. Mitigating arguments are used to express uncertainties concerning the validity of other arguments.

DEFINITION 7 (Mitigating argument) *A mitigating argument A is an undermining inference $\Rightarrow$ that links a set or premises to an argument B, negating its validity. $A : P_1, \ldots, P_n \Rightarrow B$ where each premise P_i is a membership function and the conclusion is another argument either forecast or mitigating: $B \in \mathrm{Ar}^F \cup \mathrm{Ar}^M$.*

Notation: Now on, the sets of forecast and mitigating arguments defined by a designer are, respectively, denoted as AR^F and AR^M.

The notion of mitigating argument allows a designer to model possible conflicts between arguments. *Conflict*, often replaced by the terms *attack* or *counterargument*, is an important notion for DR. Three types of conflicts have emerged in the literature (Prakken 2011): undermining, rebutting and undercutting.

A *rebutting attack* occurs when a forecast argument negates the conclusions of another argument.[5] A rebuttal attack is symmetrical, so it holds that if an argument A rebuts B, then also B rebuts A.

DEFINITION 8 (Rebutting attack) *Given two distinct forecast arguments A, $B \in \mathrm{AR}^F$ with $A : P_1, \ldots, P_n \rightarrow c_1$, $B : P_1, \ldots, P_j \rightarrow c_2$, A is a rebuttal of B, denoted as (A, B) if c_1 logically contradicts c_2.*

Property 1 A rebuttal attack is symmetrical so it holds that iff (A, B), then $\exists(B, A)$

An *undermining attack* occurs when an argument is attacked on one of its premises, by another argument having a conclusion that negates that premise.[6]

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 (Pollock 1974, 1987). In contrast to rebutting, an undercutting attack does not negate the conclusion of its target argument, rather it argues that the target's conclusions are not supported by its premises and, as a consequence, cannot be drawn. For simplicity, just undercutting attacks are used in this research study.

DEFINITION 9 (Undercutting attack) *Given a mitigating argument $A \in AR^M$ that challenges some or all of the information used to construct a forecast or another mitigating argument $B \in AR^F \cup AR^M$, A undercuts B and it is indicated as (A, B) when A claims there is a special case that does not allow the application of the inference rule of B.*

In the definitions of rebutting, undercutting and undermining, the attacker and the attacked arguments must be distinct. This excludes situations of self-defeating. Here, it is assumed that a workload designer does not deal with self-defeating propositions. 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 elicited for assessing MWL.

5.3. *Layer 3 – Activation of argumentation framework*

Once the knowledge base of a designer is formally translated into an *argumentation framework*, it can be now elicited with objective inputs. These inputs activate designed arguments with certain degrees of truth, making them more or less credible. In turn, also the credibility of the designed attacks is affected, thus few questions raise: What are the arguments that are credible enough? what are the proper attacks? When is an attack from a less credible attacker to a more credible attacked argument still valid? In other words, two key issues emerge: how to consider an argument strong enough to be part of an argumentation framework and (2) how to consider an attack powerful enough in order to succeed. These issues are well known in the arena of computational models of arguments that use the notion of strength both for arguments and attacks and their resolution is far from being trivial. In detail, in relation to attacks, we argue that if an attacker's degree of truth is higher than the degree of truth of the attacked argument, there is no doubt the attack can be considered a proper one. Even if the difference in their degrees of truth is minimal, the attack still makes sense because it is conceptualised by a designer. However, if an attacker has a lower degree of truth than the attacked argument, the issue is when to consider it a proper attack and when to disregard it.

The proposal here is to use the degree of truth of an argument, as in Definition 5, for the two problems, jointly with two *reluctancy thresholds*. These thresholds respectively indicate how reluctant a designer would be to disregard: (1) an argument (and all its outgoing/incoming attacks) and (2) an attack (rebutting/undercutting/undermining). The application of these reluctancy thresholds defines the *set of activated arguments* and the *set of activated attacks*. These thresholds, applicable jointly with the notion of degree of truth, have been designed for a finer-grained level of investigation of the elicitation of a knowledge base in relation to MWL. A similar proposal has been presented in another study (Dunne et al. 2011). However, future works will be more focused on these thresholds.

DEFINITION 10 (Argument reluctancy threshold) *The argument reluctancy threshold* $\mathrm{Rel}_{\mathrm{Arg}} : [0 \ldots 1] \in \Re$ *indicates the minimum degree of truth an argument must have to be activated and included in an argumentation framework.*[7]

DEFINITION 11 (Set of activated arguments) *Given a set Args of designed arguments and the argument reluctancy*

threshold Rel_{Arg}, *the set of activated arguments is*

$$\text{Arg}_{\text{act}} = \begin{cases} a | a \in \text{Args} \wedge (1 \geq a_{\text{deg}} \geq \text{Rel}_{\text{Arg}}) \\ \quad \text{if} \quad a \in \text{Ar}_F, \\ a | a \in \text{Args} \wedge (1 \geq a_{\text{deg}}, b_{\text{deg}} \geq \text{Rel}_{\text{Arg}}) \\ \quad \text{if} \quad a : P_1, \ldots, P_n \rightarrow b \in \text{Ar}^M. \end{cases}$$

In simpler words, for forecast arguments, their degree of truth has to be equal or higher than the value specified by the argument reluctancy threshold, while for mitigating arguments, the degree of truth of both their premises and the attacked argument has to be equal or higher than the argument reluctancy threshold.

DEFINITION 12 (Attack reluctancy threshold) *The attack reluctancy threshold* $\text{Rel}_{\text{Att}} : [0 \ldots 1] \in \Re$ *indicates the reluctancy to tolerate an attack from a less to a more credible argument.*[8]

DEFINITION 13 (Set of activated attacks) *Given the set* Arg_{act} *of activated arguments, a set* Atts *of attack relations, the attack reluctancy threshold* Rel_{Att} *and Abs the absolute function. the set of activated attacks is defined as* $\text{Attack}_{\text{act}} : \{(a, b) | (a, b) \in \text{Atts} \wedge \ a, b \in \text{Arg}_{\text{act}} \ \wedge (a_{\text{deg}} \geq b_{\text{deg}} \ \vee \ 0 \leq \text{Abs}(a_{\text{deg}} - b_{\text{deg}}) < 1 - \text{Rel}_{\text{Att}})\}$.

The set of activated arguments and the set of activated attacks define a new argumentation framework which is equal or smaller than the framework emerged at the end of layer 2.

5.4. *Layer 4 – Execution of acceptability semantics*

In order to investigate the potential inconsistencies that might emerge from the interaction of activated arguments (through the activated attacks), Dung-style acceptability semantics are applied (Dung 1995). 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* emerges (Dung 1995).

An *abstract argumentation framework* (*AAF*) is a pair ⟨Arg, attacks⟩ where

- Arg is a finite set of (abstract) *arguments*,
- attacks ⊆ Arg × 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. In Figure 5, *A* is defeated by *B*, and apparently *A* should not be accepted since it has a counterargument. However, *B* is itself defeated by *C* that is not defeated 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 situation, it is said that *C reinstates A*. Due to this issue of *reinstatement*, a formal criterion that determines which arguments of an AAF can be accepted 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 (Baroni, Caminada, and Giacomin 2011), but here the focus is on the preferred semantics proposed in Dung (1995). A set $X \subseteq$ Arg of argument is

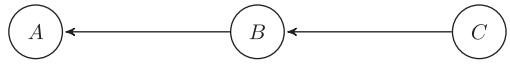

Figure 5. Argument reinstatement.

- *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 $\subseteq$).[9]

Preferred semantics can produce one or more extensions (set of arguments). In the case just one extension is produced, this coincides with the grounded extension. However, in the 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 (extension: a conflict-free set of arguments). However, considering just the cardinality might be reductive in the case, for instance, an extension with several arguments has a combined degree of truth lower than an extension containing fewer arguments. For these reasons, the proposal is to adopt the cardinality of an extension jointly with the degree of truth of its arguments to quantify its credibility.

DEFINITION 14 (Acceptable extension credibility) *Given the set of activated arguments* Arg_{act}, *an acceptable extension E, as computed by the Dung-preferred acceptability semantics and Card the cardinality function, E's credibility*

is defined as

$$E_{\text{cred}} : \frac{\text{Card}(E)}{\text{Card}(\text{Arg}_{\text{act}})} + \frac{1}{\text{Card}(E)} \sum_{\text{Arg} \in E} \text{Arg}_{\text{deg}}.$$

5.5. *Layer 5 – Computation of MWL*

Given a set of extensions, as computed by the preferred acceptability semantics, and their quantified credibility, the final step is aimed at inferring a final MWL index. The most credible extension could be used as the most credible and reasonable point of view for assessing MWL. One preferred extension should emerge as the most credible, but in the case multiple equally credible extensions are computed by an acceptability semantics (preferred in this study), all are considered to assess a final crisp index of MWL. As defined before, two typologies of arguments can exist within an extension: forecast and mitigating. However, just forecast arguments support a conclusion (workload dichotomy) that can be considered to infer a single MWL index. Mitigating arguments already played their role (through their attacks against other arguments), contributing to the computation of the acceptable extensions.

DEFINITION 15 (Overall MWL index) *Given a set AE containing the n computed preferred acceptable extensions, the set* SE *containing the most credible extension/s* $\text{SE} = \{A \mid A \in AE \wedge A_{\text{cred}} = \max(E^1_{\text{cred}}, \ldots, E^n_{\text{cred}})$ *with* $E^1, \ldots, E^n \in AE\}$, *the overall index of* $\text{MWL} : [0 \ldots 100] \in \Re$ *is*

$$\text{MWL} = \sum_{A \in \text{SE}} \left(\frac{\sum_{\text{arg} \in A} \text{arg}_c(\text{arg}_{\text{deg}})}{\text{Card}(A)} \right) \frac{1}{\text{Card}(\text{SE})},$$

with arg_c *being the workload dichotomy of a forecast argument and* Card *the cardinality function.*

The final index of MWL is assessed using the degree of truth of each argument, in the acceptable preferred extension, as the input of the same argument's workload dichotomy. Each output represents a partial MWL assessment that is then averaged with the other values computed for the other arguments in the same extension. In case multiple equally stronger extensions exist, the aforementioned process is repeated for each of them and computed values are averaged again.

6. Experiments and evaluation

In order to evaluate the designed argument-based framework for MWL, the knowledge base of the author of this study (Appendix 1) has been translated into arguments and attack relations as per layers 1 and 2 of the multi-layer schema of Figure 1. The emerging argumentation framework (Appendix 2, now on referred to as MWL_{def} or $\text{MWL}^{\text{NI}}_{\text{def}}$ when interaction of arguments is not accounted – no attacks) has been elicited performing a user study. Computed MWL indexes have been compared against the indexes computed with two well-known subjective MWL assessment techniques:

- the NASATLX (Hart 2006), developed by the Human Performance Group at NASA.
- the WP (Tsang and Velazquez 1996), based upon the multiple resource theory (Wickens 2008).

In particular, three of the properties of Section 2.1.1 have been compared: the degree of sensitivity, diagnosticity and validity. As previously mentioned, these properties are well known in the literature of MWL and they have been proposed as guidelines for evaluating MWL assessment techniques (O'Donnell and Eggemeier 1986; Rubio et al. 2004; Tsang and Velazquez 1996; Zhang and Luximon 2005). Table 1 underlines each property and the formal statistical tool adopted to test it.

6.1. *Participants and procedure*

A sample of 40 people fluent in English volunteered to participate in the study. They were divided into 2 groups of 20 each. Ages ranges from 20 to 35 years; there were 20 females and 20 males (Total – Avg.: 28.6, Std. 3.98; Group A – Avg. 28.35, Std.: 4.22; Group B – Avg: 28.85, Std.: 3.70), all with a daily Internet usage of at least 2 hours. Subjects were instructed about the study and were required to sign a consent form. Participants were required to execute a set of 11 information-seeking web-based tasks (Table A4 in the appendix) as naturally as they could, over 2 or 3 sessions of approximately $\frac{45}{70}$ minutes each, on different non-consecutive days. Tasks differed in terms of difficulty, time pressure, time-limits, and interruptions. Two groups were created because the tasks were executed on web-based interfaces, sometimes altered at run-time and sometimes not (as in Table A3). This was done because at the end of the study a statistical analysis of the MWL imposed by the original and altered interfaces was performed. However, the outcomes of this analysis are not presented in this paper. Subjects in group A were different to the subjects in group B. Participants could not interact with examiners during the tasks. Although the 11 tasks were the same across the two groups, they were performed on two different web interfaces. 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 experiment, a computerised questionnaire (Table A5 in the appendix) was administered immediately after task completion.[10] In addition, a pair-wise comparison of the questions required by the NASATLX instrument was performed.[11] Each question had to be answered with a value within the range 0–100, by moving a slider on a web page. The default value was 50 and the range was divided into three parts of equal size, guided by two separation lines, generating 3 regions (low, medium and high). Each answer represents

Table 1. Properties for evaluating different MWL assessment techniques and associated statistical tests.

Property/method	Description/goal
Sensitivity	The reliability to detect changes in resource demand, task difficulty, user features and environmental influence
ANOVA + *Post Hoc*	To find out to what extent the indices of MWL varied as a function of objective changes and manipulation of tasks
Diagnosticity	The capacity to quantify the contributions to MWL by the type, resource demand or the human operator capabilities
Multinomial logistic regression	To determine to what extent MWL attributes allow discrimination between tasks
Validity	The capacity to measure MWL
Pearson/Spearman	*Convergent validity*: to determine to what extent the model measured what is supposed to be measured
	Concurrent validity: to determine to what extent the model is able to explain objective performance measure

an objective input (numerical value) and can be employed to compute the degree of truth of one or more arguments.

6.2. Results

6.2.1. Sensitivity

In order to test the *sensitivity* of MWL_{def}, a one-way analysis of variance (ANOVA) is adopted to determine whether there are any significant differences between the means of the 11 independent tasks . The assumptions behind the procedure are met: continuity of dependent variables, their independency, the absence of outliers, their normality and the homogeneity of variance. In details, homogeneity of variance was verified using Levene's test that was positive for the WP and the MWL_{def} but not for the NASATLX instrument. In the last case, a Welch F-test is added to the ANOVA procedure and the Games–Howell *post hoc* tests were carried out instead of the Tukey *post hoc* tests. In the other cases (WP, MWL_{def}) the classical ANOVA procedure was adopted, and the Tukey *post hoc* test was conducted as all the assumptions were met. In general, the ratio of between-groups (tasks) and within-groups (participants) was higher with the NASATLX (Group A: $F(10, 206) = 13.467$, Group B: $F(10,81.065) = 10.316$) and the instance MWL_{def} (Group A: $F(10, 207) = 12.146$, Group B: $F(10,205) = 9.895$), underlying higher variance. WP was the assessment instrument with the lowest variance (Group A: $F(10,209) = 5.182$, Group B: $F(10,204) = 5.649$). From the summary of detected statistically significant differences with *post hoc* tests (Table 2), WP was the lowest in sensitivity, detecting half of the statistically significant differences spotted by the other instruments. For group A, the defeasible instance MWL_{def} behaved very analogously, demonstrating similar sensitivity with the NASATLX 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 NASATLX underlying a higher degree of robustness and being more stable in detecting differences among tasks in the groups.

Table 2. Detected statistically significant differences.

	Group A		Group B	
Model $\|\alpha$	0.05	0.01	0.05	0.01
NASATLX	22	13	14	10
WP	9	5	8	6
MWL_{Def}	21	18	18	13

In summary, according to the number of detected statistically significant differences in Table 2, out of all the possible detectable differences ($110 - 55$ for each group), *the instance* MWL_{def} showed 39.9% and 36.3% of sensitivity more than the WP and 5.45% and 14.5% of sensitivity more than the NASATLX instrument, respectively, at significance levels of 0.05 and 0.01.

6.2.2. Diagnosticity

In order to test the *diagnosticity* of MWL_{def}, *stepwise multinomial logistic regression*[12] has been used to investigate the differences between tasks on the basis of the MWL 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 (MWL attributes) and it extends logistic regression as it compares multiple groups (tasks) through a combination of binary logistic regressions. The goal was to determine the impact of multiple independent MWL attributes to predict the membership of one or other of the 22 tasks (11 for group A and 11 for group B). The assumptions behind multinomial logistic regression were met: minimum sample size of 10 (Peduzzi et al. 1996) and absence of multicollinearity of the independent variables.

Table 3 shows the model fitting information: every Sig. value for every set of attributes, considered in each MWL instrument, is less than the level of significance ($< .05$). The null hypothesis of no difference (Chi-square value) between the model without independent variables (intercept only) and the model with the independent

Table 3. Model fitting information.

	22 tasks – 11 Group A, 11 Group B			
	Fitting criteria	Likelihood ratio tests		
Model	−2 Log likelihood	Chi-Square	df	Sig.
Intercept only	2720.117			
NASATLX (5 MWL attributes)				
Final	2258.907	461.210	105	< 0.001
WP (8 MWL attributes)				
Final	1773.885	946.233	168	< 0.001
MWL_{def} (19 MWL attributes)				
Final	1188.568	1531.550	357	< 0.001

Table 4. Accuracies of regression models.

95% CI	Prediction accuracy (%)
NASATLX attributes	19.1
WP attributes	32.3
MWL_{def} and MWL_{def}^{NI} attributes	53.2

variable (final) is rejected in every test. This underlines the existence of a relationship between the MWL 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 MWL 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 4 summaries the classification accuracy rates computed by each logistic regression model. All of these rates are above 5.56%, satisfying the criteria for classification accuracy.

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 MWL attribute to the overall classification. The interpretation for an independent MWL 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 MWL 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 Appendix 5. From Table A6, it is possible to note how all the attributes considered in the NASATLX 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 A7. 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 A8 shows the likelihood ratio tests. Here, the attributes all have a significance value less than 0.05, but the *mental demand* and *intention* are not included, as they are not considered significant to classify tasks by the stepwise multinomial logistic regression procedure.

The information associated with 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 MWL attribute for classifying tasks. Appendix 6 lists the step summary tables for each multinomial logistic regression procedure of each MWL assessment instrument. From these tables, 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), as well as the contributions that each attribute had to the model's goodness of fit. In the case of the attributes accounted in the original NASATLX, *temporal demand*, *effort* and *performance* were the most significant contributors as their addition, at each step, reduced the chi-square significantly. Table A9 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 attribute *performance* to 2337.516. *Psychological stress* and *mental demand*, although they were valid contributors, had a less powerful role in reducing the chi-square. Regarding the attributes accounted in the WP instrument, all had a significant effect

in reducing the chi-square. From Table A10, 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 goodness of fit. Eventually, all the attributes accounted in the two instances (MWL_{def} and MWL_{def}^{NI}) of the defeasible framework 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 A11, the most impactful contributor was *auditory resources*, followed by *parallelism*, *temporal demand* and *effort*. The attributes with lowest influence to the goodness of fit were *arousal* and *central processing*.

In summary, the attributes accounted in MWL_{def} showed a greater diagnosticity compared to the one achieved by the attributes of the NASATLX and WP instruments, in terms of capacity of classifying each case in the right category (one of the executed tasks). Considering the set of executed tasks, listed in Table A4 , MWL_{def} had an accuracy rate 34.1% higher than that of the NASATLX instrument and 20.9% higher than that of the WP instrument, confirming its prospective in assessing subjective MWL.

6.2.3. *Validity*

In order to test the *validity* of MWL_{def}, the intercorrelation of the scores computed by the other two MWL instruments (NASATLX and WP) and the correlation of each of them against objective performance measure have been computed. The former is referred to as *convergent validity*, while the latter as *concurrent validity*, both assessed using Pearson's correlation coefficients and Spearman's

Table 5. Convergent validity of the MWL scores and concurrent validity against time – Pearson's coefficients.

		Pearson				
		NASATLX	WP	MWL_{def}^{NI}	MWL_{def}	Time
NASATLX	Correlation	1	0.584	0.562	0.778	0.315
	Sig.		0.000	0.000	0.000	0.000
	Cases		440	440	440	352
WP	Correlation		1	0.654	0.859	0.264
	Sig.			0.000	0.000	0.000
	Cases			440	440	352
MWL_{def}^{NI}	Correlation			1	0.713	0.272
	Sig.				0.000	0.000
	Cases				440	352
MWL_{def}	Correlation				1	0.381
	Sig.					0.000
	Cases					352

Note: All the coefficients are statistically significant ($p < .000$).

Table 6. Convergent validity of the MWL scores and concurrent validity against time – Spearman's coefficients.

		Pearson				
		NASATLX	WP	MWL_{def}^{NI}	MWL_{def}	Time
NASATLX	Correlation	1	0.571	0.579	0.780	0.335
	Sig.		0.000	0.000	0.000	0.000
	Cases		440	440	440	352
WP	Correlation		1	0.658	0.854	0.259
	Sig.			0.000	0.000	0.000
	Cases			440	440	352
MWL_{def}^{NI}	Correlation			1	0.738	0.250
	Sig.				0.000	0.000
	Cases				440	352
MWL_{def}	Correlation				1	0.346
	Sig.					0.000
	Cases					352

Note: All the coefficients are statistically significant ($p < .000$).

Table 7. Convergent validity of the MWL scores and concurrent validity against time – No time-limit tasks – Pearson coefficients.

		Pearson				
		NASATLX	WP	MWL_{def}^{NI}	MWL_{def}	Time
NASATLX	Correlation	1	0.590	0.597	0.763	0.384
	Sig.		0.000	0.000	0.000	0.000
	Cases		320	320	320	248
WP	Correlation		1	0.679	0.856	0.305
	Sig.			0.000	0.000	0.000
	Cases			320	320	248
MWL_{def}^{NI}	Correlation			1	0.752	0.344
	Sig.				0.000	0.000
	Cases				320	248
MWL_{def}	Correlation				1	0.447
	Sig.					0.000
	Cases					248

Note: All the coefficients are statistically significant ($p < .000$).

Table 8. Convergent validity of the MWL scores and concurrent validity against time – no time-limit tasks – Spearman's coefficients.

		Spearman				
		NASATLX	WP	MWL_{def}^{NI}	MWL_{def}	Time
NASATLX	Correlation	1	0.571	0.623	0.761	0.369
	Sig.		0.000	0.000	0.000	0.000
	Cases		320	320	320	248
WP	Correlation		1	0.681	0.853	0.286
	Sig.			0.000	0.000	0.000
	Cases			320	320	248
MWL_{def}^{NI}	Correlation			1	0.779	0.333
	Sig.				0.000	0.000
	Cases				320	248
MWL_{def}	Correlation				1	0.392
	Sig.					0.000
	Cases					248

Note: All the coefficients are statistically significant ($p < .000$).

rank correlation coefficients.[13] In this comparison, also the instance MWL_{def}^{NI} has been included. The performance measure adopted for convergent validity is the objective *task completion time* of participants.[14] Tables 5 and 6 refer to all the tasks used in the experiments, while Tables 7 and 8 present the correlations of those tasks with no imposed time limit.

The *convergent validity* of the ML instruments is high, with the instance MWL_{def} highly correlating with the NASATLX and WP both according to Pearson's and Spearman's correlation coefficients (Pearson: 0.778, 0.859 with all tasks, 0.763, 0.856 without time-limit tasks – Spearman: 0.780, 0.854 with all tasks. 0.761, 0.853 without time-limit tasks). The NASATLX and the WP showed a moderate positive correlation (Pearson: 0.584 with all tasks, 0.590, without time-limit tasks – Spearman: 0.571 with all tasks, 0.571 without time-limit tasks). The instance MWL_{def}^{NI} of the framework with no interaction of arguments only moderately correlated to NASATLX WP (Pearson: 0.562, 0.654 with all tasks, 0.597, 0.679 without time-limit tasks – Spearman: 0.579, 0.658 with all tasks. 0.623, 0.681 without time-limit tasks) having less convergent validity than its counterpart with interactions among arguments (MWL_{def}).

Regarding the *concurrent validity*, the instance MWL_{def} of the defeasible framework correlated better with time, showing a moderate positive correlation (Pearson: 0.381 with all tasks, 0.447, without time-limit tasks – Spearman: 0.346 with all tasks. 0.392 without time-limit tasks) than the NASATLX (Pearson: 0.315 with all tasks, 0.384, without time-limit tasks – Spearman: 0.335 with all tasks. 0.369 without time-limit tasks), the WP instrument (Pearson: 0.264 with all tasks, 0.305, without time-limit tasks – Spearman: 0.259 with all tasks. 0.286 without time-limit tasks) and the instance MWL_{def}^{NI} of the defeasible framework with no interaction of arguments (Pearson:

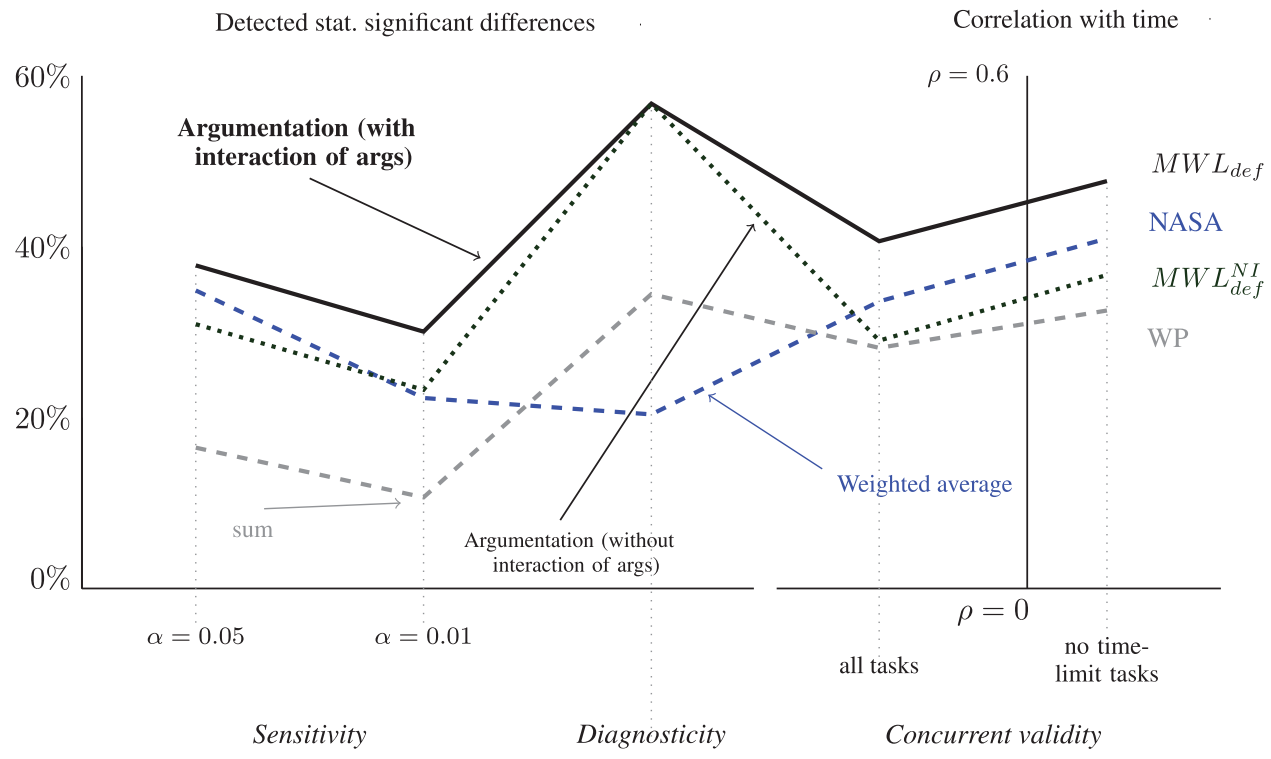

Figure 6. Sensitivity, diagnosticity, validity – comparisons.

0.272 with all tasks, 0.344, without time-limit tasks – Spearman: 0.250 with all tasks. 0.333 without time-limit tasks). All the correlation coefficients are statistically significant.

In summary, the instance MWL_{def}, as hypothesised, showed a high convergent validity with the NASATLX and the WP instruments, but it had a better concurrent validity with the objective time than the other two instruments and its counterpart MWL^{NI}_{def} with no interaction of arguments.

7. Discussion

Enhancement in the quality of MWL assessment has been proved by building two new instances of the defeasible framework (MWL_{def}, MWL^{NI}_{def}), with and without interaction of arguments, respectively. In line with other studies aimed at comparing the psychometric properties of different MWL assessment instruments (Rubio et al. 2004), these two instances have been compared against the NASATLX (Hart 2006) and the WP (Tsang and Velazquez 1996) instruments with respect to sensitivity, diagnosticity and validity.

Figure 6 summarises the outcomes of this comparison, as achieved in previous sections. Using a confidence interval of 95%, the instance MWL_{def} was superior to its counterpart MWL^{NI}_{def} (with no interaction), similar to the NASATLX and clearly superior to the WP in terms of *sensitivity*. These outcomes were also confirmed by increasing the confidence interval to 99%, where, however, the NASATLX lost capacity in detecting statistically significant differences among tasks.

Regarding *diagnosticity*, the pieces of knowledge used in the argumentation frameworks of MWL_{def} and MWL^{NI}_{def} enabled better discrimination between tasks than the other two instruments. This suggests that an open framework able to incorporate different MWL attributes allows multiple tests and can provide information on their capacity in discriminating tasks.

The *convergent validity* showed the capacity of the two constructed instances (MWL_{def} and MWL^{NI}_{def}) of the defeasible framework to measure MWL effectively, as their correlations against the NASATLX and the WP instruments were strong and positive (on average $\rho > 0.57$).

However, *the concurrent validity* of the MWL_{def}, against the objective time for task completion, was superior to the other three procedures (both considering all tasks and just those tasks with no time-limit). This achievement not only suggests that MWL_{def} (as constructed) can explain the objective time better than the other procedures, but it also underlines how studying and reasoning upon the construct of MWL in a defeasible way can enhance its assessment effectively. In addition, the fact that the instance of the framework (MWL_{def}) was clearly superior to its counterpart (MWL^{NI}_{def}), with no interactions of arguments, highlights the real benefits achievable by incorporating the relationships among the pieces of evidence during the formalisation of a knowledge base. This confirms the role of AT in MWL representation and assessment, enabling and promoting further research.

8. Conclusions

The main contribution of this research is the presentation of a methodology, developed as a formal framework, to represent MWL as a defeasible computational concept and to assess it as a numerical usable index. This research contributes to the body of knowledge by providing an extensible framework, built upon AT, in which MWL can be better measured, analysed and explained. The framework allows the translation of a knowledge base of a MWL designer into interactive defeasible arguments. These are in the form of premises-conclusion, incorporating the notion of logical consequence. In turn they are activated through quantification of their premises using the notion of degree of truth. The emerging activated argumentation framework is subsequently evaluated by applying acceptability semantics for the resolution of the potential inconsistencies that might emerge from interaction of arguments. These are algorithms that partition the argument set in conflict-free extensions of arguments. A strategy for selecting the most credible extension is then introduced from which a final index of MWL can be assessed. The proposed framework has been firstly employed in practice for the translation of the author's knowledge base into a set of interactive arguments. The resulting instance of the framework has been elicited through a users' study involving 40 participants who were asked to execute a set of 11 information-seeking web-based tasks with different conditions. Afterwards, participants were asked to fill in a questionnaire, providing numerical inputs for the designed instance. A comparative evaluation showed how this particular instance was superior in terms of sensitivity, diagnosticity and validity to the NASA Task Load Index and the WP, these being among the most popular current subjective MWL assessment techniques. The results underlined the positive impact of the defeasible framework on MWL representation and assessment, encouraging further research.

As the first study of this kind, this study proposes a new reasoning framework for tackling the problem of MWL representation. This has been introduced as a theoretical solution with the aim of breathing new life into research on MWL. Future work will focus on the implementation of a graphical interface for automating the use of the theoretical framework. This is aimed at increasing its acceptability by MWL designers, guiding them towards the translation of their knowledge bases using more familiar terms and less formal reasoning notions. The goal is to supply most of the logic out of the box, allowing designers to extend it and tweak input parameters.

On the theoretical side, alternative methods for selecting the most credible extension and for the accrual of its internal arguments using degrees of truth will be investigated. This will include the test of different acceptability semantics, as proposed in the literature on formal AT. Eventually, an analysis of the impact of the variation of the reluctancy thresholds on the elicitation of knowledge bases will follow aimed at evaluating whether the theoretical framework could be simplified.

Eventually, regarding the evaluation of the capability of the framework to assess MWL, a more effective sensitivity analysis will 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 MWL (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.

Finally, the ultimate goal of the solution described in this paper is to allow different MWL designers with different backgrounds, knowledge and beliefs to create different instances of the framework, with different arguments, workload attributes and relationships towards a better modelling and understanding of the construct of MWL.

Conflict of interest disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. Face validity refers to what a concept superficially appears to measure, mainly testing whether it looks valid. It is in contrast with content validity – a more strict property that requires the use of recognised tests or subject experts for evaluating whether the items evaluated assess defined content. This includes statistical tests which are in general more rigorous than methodologies applied in face validity tests.
2. 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 that it can either belong or not belong to the set. 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 0–1.
3. A strict monotonic function has the property that for two different inputs, being the former 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.
4. Dichotomies are both *jointly exclusive* and *mutually exclusive*.
5. 'Tweety flies because it is a bird' can be negated by 'Tweety does not fly because it is a penguin'.
6. 'Tweety flies because it is a bird' can be attacked by another argument 'Tweety is not a bird'.
7. 1 mean just those arguments with full degree of truth are activated (indeed too restrictive). 0 indicates no reluctancy at all: each designed argument will be activated, regardless of its degree of truth.
8. 0 indicates null reluctance: any designed attack is considered valid. Intermediate values indicate partial reluctancy: with 0.6 the designer is willing to tolerate an attack if the difference in the argument's degrees of truth is less than or equal to $1 - 0.6 = 0.4$.
9. In Figure 5 there is just one complete extension, $\{A, C\}$, which is conflict-free and defends exactly itself. It can be seen as a subjective and internally coherent point of view. The grounded extension is $\{A, C\}$. The admissible sets are C, A, C. B and A are not admissible as they do not defend themselves, respectively, against C and B. One preferred extension exits: A, C.
10. This included 20 questions. Six were associated with the NASA Task Load Index original instrument (Hart 2006). Eight were associated with the WP procedure (Tsang and Velazquez 1996). The remaining six were designed to model some other aspects of MWL. The order of these three blocks was random.

11. This procedure aims to create an individual weighting of the 5 sub-scales (physical demand was not taken into account) by letting the subjects compare them pair-wise, based on their perceived importance. The user is required to choose which measurement is more relevant to the workload. The number of times each is chosen is the weighted score. This is multiplied by the scale score for each dimension and then divided by 10 to get a workload score $[0...100] \in \Re$ (Hart 2006).
12. Differently from other studies that employed discriminant analysis to assess diagnosticity (Rubio et al. 2004; Tsang and Velazquez 1996), multinomial logistic regression was adopted because it does not impose all the assumptions required by the discriminant analysis that were not all met in this study.
13. Spearman's rank correlation coefficient is a non-parametric measure of statistical dependence between two variables. Likewise, Pearson's correlation coefficient tells how the relationship between two variables can be described using a monotonic function, but upon the ranked variables.
14. Some cases do not have time due to measurement error.

References

Annett, J. 2002. "Subjective Rating Scales in Ergonomics: A Reply." *Ergonomics* 45 (14): 1042–1046.

Bailey, B. P., and J. A. Konstan. 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.

Baroni, P., M. Caminada, and M. Giacomin. 2011. "An Introduction to Argumentation Semantics." *The Knowledge Engineering Review* 26 (4): 365–410.

Baroni, P., G. Guida, and S. Mussi. 1997. "Full Nonmonotonicity: A New Perspective in Defeasible Reasoning." ESIT 97, European Symposium on Intelligent Techniques, 58–62.

Bentahar, J., B. Moulin, and M. Bélanger. 2010. "A Taxonomy of Argumentation Models Used for Knowledge Representation." *Artificial Intelligence Review* 33 (3): 211–259.

Cain, B. 2007. *A Review of the Mental Workload Literature*. Technical Report. Toronto: Defence Research and Development Canada Toronto, Human System Integration Section.

Colle, H. A., and G. B. Reid. 2005. "Estimating a Mental Workload Redline in a Simulated Air-to-Ground Combat Mission." *The International Journal of Aviation Psychology* 15 (4): 303–319.

Cooper, G. E., and R. P. Harper. 1969. *The Use of Pilot Ratings in the Evaluation of Aircraft Handling Qualities*. Technical Report. AD689722, Report 567. Neuilly-sur-Seine: Advisory Group for Aerospace Research & Development.

De Waard, D. 1996. *The Measurement of Drivers' Mental Workload*. The Netherlands: The Traffic Research Centre VSC, University of Groningen.

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.

Dunne, P. E., A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. 2011. "Weighted Argument Systems: Basic Definitions, Algorithms, and Complexity Results." *Artificial Intelligence* 175 (2): 457–486.

Eggemeier, F. T., G. W. Wilson, A. F. Kramer, and D. L. Damos. 1991. "Workload Assessment in Multi-task Environments." In *Multiple-Task Performance*, edited by D. L. Damos, 207–216. London: Taylor & Francis.

Gopher, D., and E. Donchin. 1986. "Workload – An Examination of the Concept." In *Handbook of Perception and Human Performance*, Vol. 2, edited by K. R. Boff, L. Kaufman, and J. P. Thomas, 41/1–41/49. New York: John Wiley & Sons.

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., and M. H. Chignell. 1988. "Mental Workload Dynamics in Adaptive Interface Design." *IEEE Transactions on Systems, Man and Cybernetics* 18 (4): 647–658.

Hart, S. G. 2006. "Nasa-task Load Index (NASA-TLX); 20 Years Later." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50 (9): 904–908. doi:10.1177/154193120605000909

Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, Advances in Psychology, Vol. 52, edited by P. A. Hancock and N. Meshkati, 139–183. Amsterdam: North-Holland.

Huey, B. M., and C. D. Wickens. 1993. *Workload Transition: Implication for Individual and Team Performance*. Washington, DC: National Academy Press.

Kahneman, D. 1973. *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.

Kramer, A. F., E. J. Sirevaag, and R. Braune. 1987. "A Psychophysiological Assessment of Operator Workload During Simulated Flight Missions." *Human Factors* 29: 145–160.

Longo, L. 2011. "Human–Computer Interaction and Human Mental Workload: Assessing Cognitive Engagement in the World Wide Web." In *Human–Computer Interaction – INTERACT 2011*, Lecture Notes in Computer Science, Vol. 6949, 402–405. Berlin: Springer.

Longo, L. 2012. "Formalising Human Mental Workload as Non-monotonic Concept for Adaptive and Personalised Web-Design." In *User Modeling, Adaptation, and Personalization*, Lecture Notes in Computer Science, Vol. 7379, edited by J. Masthoff, B. Mobasher, M. Desmarais and R. Nkambou, 369–373. Berlin: Springer.

Longo, L. 2014. "Formalising Human Mental Workload as a Defeasible Computational Concept." PhD thesis, Trinity College Dublin.

Longo, L., and S. Barrett. 2010a. "Cognitive Effort for Multi-agent Systems." In *International Conference on Brain Informatics, Toronto, Canada*, Lecture Notes in Computer Science, Vol. 6334, edited by Yiyu Yao, Ron Sun, Tomaso Poggio, Jiming Liu, Ning Zhong, and Jimmy Huang, 55–66. Berlin: Springer.

Longo, L., and S. Barrett. 2010b. "A Computational Analysis of Cognitive Effort." In *Intelligent Information and Database Systems. Second International Conference, ACIIDS, Hue City, Vietnam,* Lecture Notes in Computer Science, Vol. 5991, edited by Ngoc Thanh Nguyen, Mann Thanh Le, and Jerzy Świątek, 65–74. Berlin: Springer.

Longo, L., and P. Dondio. 2014. "Defeasible Reasoning and Argument-based Medical Systems: An Informal Overview." In 27th International Symposium on Computer-Based Medical Systems. New York: IEEE, 376–381.

Longo, L., and L. Hederman. 2013. "Argumentation Theory for Decision Support in Health-care: A Comparison with Machine Learning." International Conference on Brain Informatics, Maebashi, Japan, Springer, 168–180.

Longo, L., B. Kane, and L. Hederman. 2012a. "Argumentation Theory in Health Care." 25th International Symposium on Computer-Based Medical Systems, Rome, Italy. IEEE, 1–6.

Longo, L., F. Rusconi, L. Noce, and S. Barrett. 2012b. "The Importance of Human Mental Workload in Web-design." The 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, SciTePress, April 2012, 403–409.

Matt, P. A., M. Morgem, and F. Toni. 2010. "Combining Statistics and Arguments to Compute Trust." The 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada, May 10–14, Vol. 1, 209–216.
Muckler, F. A., and S. A. Seven. 1992. "Selecting Performance Measures: 'Objective' Versus 'Subjective' Measurement." *Human Factors – Special Issue: Measurement in Human Factors* 34 (4): 441–455.
Neville, M., Paul Eisen, Laura Money, and I. B. Turksen. 1988. "Fuzzy Analysis of Skill and Rule-based Mental Workload." In *Human Mental Workload*, Advances in Psychology, Vol. 52, edited by P.A. Hancock and N. Meshkati, 289–304. Amsterdam: North-Holland.
O'Donnell, R. D., and T. F. Eggemeier. 1986. "Workload Assessment Methodology." *Handbook of Perception and Human Performance*, Vol. 2. New York: Wiley-Interscience. 4:21–4:249.
Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. "A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49 (12): 1373–1379.
Pollock, J. L. 1974. *Knowledge and Justification*. Princeton, NJ: Princeton University Press.
Pollock, J. L. 1987. "Defeasible Reasoning." *Cognitive Science* 11 (4): 481–518.
Prakken, H. 2011. "An Abstract Framework for Argumentation with Structured Arguments." *Argument and Computation* 1 (2): 93–124.
Reid, G. B., and T. E. Nygren. 1988. "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload." In *Human Mental Workload*, Chap. 8, Advances in Psychology, Vol. 52, edited by P. A. Hancock and N. Meshkati, 185–218. Amsterdam: North-Holland.
Roscoe, A. H., and G. A. Ellis. 1990. *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use.* Technical Report TR 90019, Royal Aerospace Establishment, Farnborough.
Rubio, S., E. Diaz, J. Martin, and J. M. Puente. 2004. "Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods." *Applied Psychology* 53 (1): 61–86.
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. 2010. "Argumentative Agents." Proceedings of the Multi-conference on Computer Science and Information Technology, Wisla: IEEE, 223–229.
Toulmin, S. 1958. *The Use of Argument*. Cambridge: Cambridge University Press.
Tsang, P. S. 2006. "Mental Workload." In *International Encyclopedia of Ergonomics and Human Factors*, Vol. 1, 2nd ed., edited by W. Karwowski, Chap. 166, 809–813. Boca Raton, FL: CRC Press.
Tsang, P. S., and V. L. Velazquez. 1996. "Diagnosticity and Multidimensional Subjective Workload Ratings." *Ergonomics* 39 (3): 358–381.
Tsang, P. S., and M. A. Vidulich. 2006. "Mental Workload and Situation Awareness." In *Handbook of Human Factors and Ergonomics*, 3rd ed., edited by G. Salvendy, 243–268. Hoboken, NJ: John Wiley & Sons.
Vidulich, M. A., and P. S. Tsang. 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 G. S. J. Ward Frederic. 1991. "Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment." *Human Factors Society* 33 (6): 677–691.
Wickens, C. D. 2008. "Multiple Resources and Mental Workload." *Human Factors* 50 (2): 449–454.
Wickens, C. D., and J. G. Hollands. 1999. *Engineering Psychology and Human Performance*. 3rd ed. Prentice Hall.
Wierwille, W. W., and T. F. Eggemeier. 1993. "Recommendations for Mental Workload Measurement in a Test and Evaluation Environment." *Human Factors* 35 (2): 263–281.
Wilson, G. F., and T. F. Eggemeier. 1991. "Psychophysiological Assessment of Workload in Multi-task Environments." In *Multiple-Task Performance*, edited by D. Damos, 329–360. London: Taylor & Francis.
Wilson, G. F., and T. F. Eggemeier. 2006. "Mental Workload Measurement." In *International Encyclopedia of Ergonomics and Human Factors*, Vol. 1, 2nd ed., edited by W. Karwowski, Chap. 167. Boca Raton, FL: CRC Press.
Xie, B., and G. Salvendy. 2000. "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 J. D. Dodson. 1908. "The Relation of the Strength of Stimulus to Rapidity of Habit-formation." *Journal of Comparative Neurology and Psychology* 18: 459–482.
Young, M. S., and N. A. Stanton. 2004."Mental Workload." In *Handbook of Human Factors and Ergonomics Methods*, Chap. 39, edited by N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, and H. W. Hendrick, 1–9. Boca Raton, FL: CRC Press.
Young, M. S., and N. A. Stanton. 2006. "Mental Workload: Theory, Measurement, and Application." In *International Encyclopedia of Ergonomics and Human Factors*, Vol. 1, 2nd ed., edited by W. Karwowski, 818–821. Boca Raton, FL: CRC Press.
Zadeh, L. A. 1965. "Fuzzy Sets." *Information and Control* 8 (3): 338–353.
Zhang, Y., and A. Luximon. 2005. "Subjective Mental Workload Measures." *Ergonomia* 27: 199–206.
Zijlstra, F. R. H. 1993. "Efficiency in Work Behaviour." Doctoral thesis, Delft University, The Netherlands.

Appendix 1. Knowledge base of author to be translated

- Task demands (exogenous factors) (from Reid and Nygren 1988; Hart 2006)
  (1) *Mental demand*: the higher the perceived mental demand of the task, the higher the MWL.
  (2) *Temporal demand*: the higher the perceived temporal demand of the task, the higher the MWL.
  (3) *Physical demand*: the higher the perceived physical demand of the task, the higher the MWL.
- Task features/complexity and interaction with the user (exogenous factors) (from Wickens and Hollands 1999; Wickens 2008; Tsang and Velazquez 1996)
  (4) *Solving and deciding*: the higher the attention required for decision-making, problem-solving, the higher the MWL.
  (5) *Selection of response*: the higher the attention required for selecting the proper response channel, the higher the MWL.
  (6) *Task and space*: the higher the attention required for spatially paying attention around, the higher the MWL.

(7) *Verbal material*: the higher the attention required for processing linguistic material or listening to verbal conversation or reading, the higher the MWL.
(8) *Visual resources*: the higher the attention for task execution (based on information visually received), the higher the MWL.
(9) *Auditory resources*: the higher the attention for task execution (based on information auditorily received), the higher the MWL.
(10) *Manual response*: the higher the attention required for manually responding to the task, the higher the MWL.
(11) *Speech response*: the higher the attention required for producing the speech response, the higher the MWL.

- User's state (endogenous factors)

(12) *Psychological stress* has been thought as having a direct relationship with MWL: the higher the stress felt by the user, the higher the MWL (Hart and Staveland 1988; Reid and Nygren 1988; Hart 2006). However, here the belief is that the stress perceived by the user influences MWL only when it is too low or too high. In these two cases, the operator's state is significantly affected. In the former case, MWL is at a minimum level (underload), while in the latter case, it is at a maximum level (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 or unfamiliar tasks, the relationship becomes inverse, with declines in performance as arousal increases (Yerkes and Dodson 1908).

- User intentions (endogenous factors)

(14) *Effort*: the higher the effort exerted by the user the higher the MWL (Hart and Staveland 1988; Hart 2006).
(15) *Motivation* is related to effort and performance: the higher the user's motivation to attend to the task, the higher the willingness to exert effort to improve task performance. When motivation is moderate, the belief is that it does not have a significant influence on MWL. When motivation is too low, it might have a direct relationship with MWL: the user's state is affected and workload is hypothesised to be at a minimum level.

- context/domain (exogenous factors)

(16) *Parallelism*: the higher the parallelism regarding the execution of multiple tasks, the higher the MWL. 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 that are similar to each other are harder to execute 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 (Wickens and Hollands 1999; Wickens 2008; Tsang and Velazquez 1996).
(17) *Context bias*: when bias is not too low, the higher the bias and distraction degree, the higher the MWL. When bias is too low, workload is not influenced. On the other hand, when a moderate or high degree of bias and interruptions occurs during a 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 MWL (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*: the higher the user's knowledge of the task or the context/domain, the lower the MWL. 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 MWL 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 MWL is likely to be low. Past knowledge is an important factor that contributes to developing 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*: the higher the user's skill, the lower the MWL. 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 plays a significant role in reducing MWL on task. Skill can be related to past knowledge: if a subject has already dealt with a task or similar tasks, the skill degree 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*: the higher the performance perceived by the user, the lower the MWL (Hart and Staveland 1988; Hart 2006; O'Donnell and Eggemeier 1986).

Appendix 2. Working example

A new instance of the defeasible framework is designed according to the knowledge base of the author of this paper (as summarised with the natural language proposition of Appendix 1), driven by his subjective interpretation of the literature of MWL and his beliefs. It does not aim to be fully exhaustive and the final ultimate set of pieces of evidence to consider for representing MWL, but just a subjective proposal open to criticisms that can be extended, reduced or discarded as a whole. This knowledge base serves for demonstrating how to create a set of interactive arguments according to the multilayer schema presented in Section 5 (layers 1 and 2).

A.1. Layer 1 – definition of the monological structure of arguments

For the attribute *mental demand* (1st of Appendix 1), the designed forecast arguments are:

- MD1: [low mental demand $\rightarrow U$]
- MD2: [medium lower mental demand $\rightarrow F^{-}$]

- MD3: [medium upper mental demand → F^+]
- MD4: [high mental demand → O]

The same rationale applies to the attributes 2–11, 14, 16, 17 (forming other 52 arguments).

For the attribute *psychological stress,* the arguments are:

- PS1: [low psychological stress → U]
- PS2: [high psychological stress → O]

For the attribute *arousal*, no forecast argument is designed.

For *motivation*, just one argument is designed:

- MV1: [low motivation → U]

For *past knowledge*:

- PK1: [low past knowledge → O]
- PK2: [high past knowledge → U]

For *skills*:

- SK1: [low skills → O]
- SK2: [high skills → U]

For the *performance*, which has an inverted relationship with MWL, 4 arguments are designed:

- PF1: [low performance → O]
- PF2: [medium lower performance → F^+]
- PF3: [medium upper performance → F^-]
- PF4: [high performance → U]

The mitigating arguments that might be designed considering the aforementioned knowledge base (Appendix 1) are as follows:

13 *arousal*:
- 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]

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]

A.2. Layer 2 – Definition of the dialogical structure of arguments

The rebutting attack that might be extracted from the aforementioned knowledge base (Appendix 1) is as follows:

- 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 this inconsistency:
    - r1: (MD1, SD4)
    - r2: (MD4, SD1)
  - From the knowledge base, (A.1) *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 modelling such inconsistency:
    - r3: (PK1, SK4)
    - r4: (PK4, SK1)
  - From the knowledge base, (points chkA.1, A.1) 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, 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 from the designed mitigating arguments are as 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)
- Undercutting
  - 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 represents a possible instance of the defeasible framework that is referred to as $\mathrm{MWL}_{\mathrm{def}}^{\mathrm{NI}}$ (no interactions). Extending this instance by adding the designed rebutting, undercutting and undermining attacks, in the argumentation graph, a new instance emerges, referred to as $\mathrm{MWL}_{\mathrm{def}}$. These two are treated as different instances of the defeasible framework because they are separately evaluated.

The last step for completing the definition of these two instances is the design of the membership functions for each MWL factor considered in the aforementioned knowledge base. During the completion of the questionnaire of Table A5, it has been noted that subjects could better quantify low levels rather than high levels while answering a question. In other words, subjects were able to easily quantify the null impact of a workload factor rather than the full impact, manifesting more uncertainty in indicating higher levels. These reasons lead to the definition of the following functions, generalised Bell curves or Gaussian curves commonly used in Fuzzy Logic (that could be easily automised with a GUI).

The functions adopted for the workload dichotomies are depicted in Figure A2, partitioned by the following RL:

- $RL_U^F = 33$
- $RL_O^F = 66$

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

- $REL_{Arg}^{th} = 0$. Willingness to consider all the arguments with degree of truth greater than 0.
- $REL_{Att}^{th} = 0.5$. Willingness to tolerate an attack from a less to a more credible argument just if their difference in degree of truth is less than 0.5.

Eventually, as it is possible to see in the list of attributes of Appendix 1, the attribute *arousal* is based on *task difficulty* for which no question has been designed in the questionnaire of Table A5. As a consequence, an explicit mechanism to quantify task difficulty is needed. Here, the proposal is to model it as the average of the workload attributes accounted in the WP instrument which can be quantified because an explicit question has been designed for each of them (questions 6–13 of questionnaire of Table A5).

A.3. Layer 3: reduction of argumentation graph

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

- $Reluct_{Arg}^{th} = 0$. Willingness to consider all the arguments whose degree of truth is greater than 0 (Definition 5).

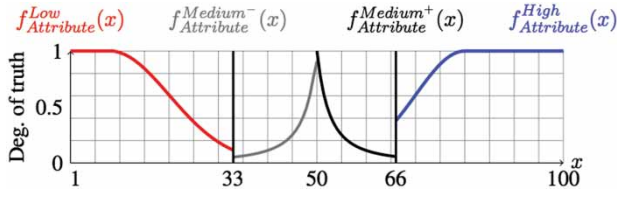

Figure A1. Membership functions associated to the premises of every argument.

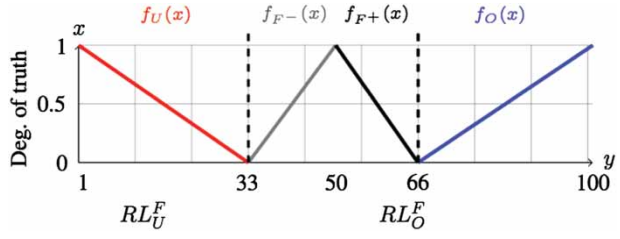

Figure A2. Function associated with the workload dichotomies partitioned by RL.

Table A1. An illustrative scenario: activated arguments and degree of truth for MWL_{def}.

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

Table A2. An illustrative scenario: activated attack relations for MWL_{def}.

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

- $Reluct_{Att}^{th} = 0.5$. 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 (Definition 12).

The instance MWL_{def} can be summarised with the following tuple using an illustrative set of inputs (answers of the questionnaire of Table A5 in the appendix):

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

- **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**: $f_{\text{pref}}(x)$ is *undefined* (no preferentiality considered)
- **MF**: the membership function for the attributes are the ones defined in Figure A1
- **RL**: {$\text{RedLine}^{\text{fitting}}_{\text{underload}} = 33$, $\text{RedLine}^{\text{overload}}_{\text{fitting}} = 66$ }
- **DMF**: workload dichotomies of Figure A2
- **ARGS**: the designed arguments built upon the attributes in ATTR are the ones listed in Section B.1
- **ATTACKS**: the designed attack relationships are the ones defined in Section A.2
- **RT**: {$\text{Reluct}^{\text{th}}_{\text{Arg}} = 0$, $\text{Reluct}^{\text{th}}_{\text{Att}} = 0.5$ }
- **INPUTS**: {70, 15, 78, 76, 12 18, 17, 14, 78, 82, 9, 13, 0, 9, 71, 64, 16, 7, 21, 30}

The values in the **INPUTS** are responsible for the activation of the argumentation graph behind the instances $\text{MWL}^{\text{NI}}_{\text{def}}$ and MWL_{def}.

The instances $\text{MWL}^{\text{NI}}_{\text{def}}$ and MWL_{def} can now be evaluated by starting with the activation of arguments and attack relations (using the values in the *INPUTS* set of the tuple). Table A1 lists which arguments are activated with the correspondent degree of truth (according to Definition 11).

Table A2 lists the activated attacks (according to Definition 13). The union of the set of activated arguments and the set of activated attacks forms the argumentation graph depicted in Figure A3 that can now be evaluated by applying Dung's *preferred semantics*, as described in Section 5.4.

A.4. Layer 4: extraction of credible extensions

Multiple extensions of arguments might be computed by the *preferred semantics*. In this case, their strength is separately computed as per Definition 14. From the reduced argumentation graph of Figure A3, two preferred extensions are computed (with the values in the *INPUTS* set of the tuple), and according to Definition 14 their strengths are as follows:

- Extension 1: 1.673
- Extension 2: 1.679

As a consequence, extension 2, although very similar to 1, is the strongest preferred extension that can be used to compute the final index of MWL, according to Definition 15.

A.5. Layer 5: assessment of MWL

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 MWL, which in this case is 16.81. It is important to recall that Definition 15 can handle multiple strongest extensions, and it accounts for the importance associated with each arguments that, however, it is undefined in the instance MWL_{def} ($f_{\text{pref}}(x) =$ *undefined*).

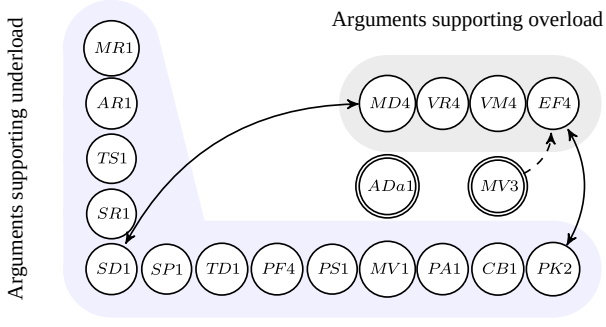

Figure A3. An illustrative scenario: activated arguments and attack relations for MWL_{def}.

- Extension 1: 1.673
- Extension 2: 1.679

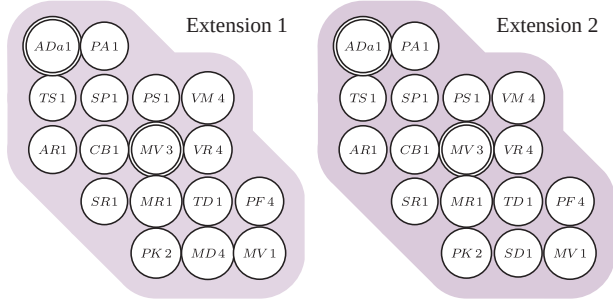

Figure A4. An illustrative scenario: computed preferred extensions for MWL_{def}.

Appendix 3. List of experimental tasks

Table A3. Interfaces used in experimental tasks by the two groups.

Task	Typology of tasks & conditions	Website	Group A interface	Group B interface
T_1	Fact finding: simple search	Wikipedia		*altered*
T_2	Browsing: not goal oriented + time limit	Wikipedia	*altered*	
T_3	Browsing: goal-oriented task	YouTube	*altered*	
T_4	Fact finding: dual task + arithmetic	Google		*altered*
T_5	Fact finding: dual task + arithmetic	Google	*altered*	
T_6	Fact finding: single task + time pressure	Google		*altered*
T_7	Fact finding: constant demand on visual + auditory resource	YouTube		*altered*
T_8	Fact finding: simultaneous demand on auditory resource + visual resource + arithmetic	YouTube + Wikipedia	*altered*	
T_9	Fact finding: single tasks on visual resource + external interference	YouTube	*altered*	
T_{10}	Fact finding: multiple concurrent tasks + time pressure	Google + Wikipedia	*altered*	
T_{11}	Fact finding: demands on auditory + visual resources + verbal processing	YouTube		*altered*

Table A4. List of experimental tasks.

Task	Description	Notes	Website
T_1	Find out how many people live in Sydney		Wikipedia
T_2	Read the content of simple.wikipedia.org/wiki/Grammar	No time imposed (user can exit at any time)	Wikipedia
T_3	Use youtube.com to play your favourite song and while listening to it, search the related lyrics	90 secs 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 14th FIFA world cup		Google
T_5	Find out the difference (in years) between the foundation of the Microsoft Corporation and the year of the 23rd Olympic games		Google
T_6	Find out the year of birth of the 1st wife of the founder of Playboy	2 mins-limit. Each 30 secs user is warned of the time left	Google
T_7	Find out the name of the man (interpreted by Johnny Depp) in the video www.youtube.com/watch?v = FfTPS-TFQ_c	Participant can replay the video if required	YouTube
T_8	(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_9	Find out how many times Stewie jumps in the video www.youtube.com/watch?v = TSe9gbdkQ8s	Participant is distracted twice & can replay video	YouTube
T_{10}	Find out (a) (using google.com) the difference (in years) between the foundation of the Alfa Romeo and the year of the 15th New York City marathon, (b) (using wikipedia.com) the capital of Namibia, (c) the two common words appearing in the titles of each referenced paper of Longo L. in en.wikipedia.org/wiki/Collaborative_search_engine	Every 30 secs user is forced to switch to subsequent task in a loop until the 3 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 secs-limit. User can replay. There is no answer	YouTube

Appendix 4. Experimental questionnaire

Table A5. Experimental study questionnaire.

No.	Dimension	Question
1	Mental demand	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking and searching)? Was the task easy (low demand) or complex (high mental demand)?
2	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 demand) or rapid and frantic (high temporal demand)?
3	Effort	How much conscious mental effort or concentration was required? Was the task almost automatic (low effort) or did require it total attention (high effort)?
4	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?
5	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?
6	Solving & deciding	How much attention was required for activities such as remembering, problem-solving, decision-making and perceiving (e.g. detecting, recognising and identifying objects)?
7	Response selection	How much attention was required for selecting the proper response channel and its execution (manual – keyboard/mouse, or speech – voice)?
8	Task and space	How much attention was required for spatial processing (spatially pay attention around you)?
9	Verbal material	How much attention was required for verbal material (e.g. reading or processing linguistic material or listening to verbal conversations)?
10	Visual resources	How much attention was required for executing the task based on the information visually received?
11	Auditory resources	How much attention was required for executing the task based on the information auditorily received?
12	Manual Response	How much attention was required for manually responding to the task (e.g. keyboard/mouse usage)?
13	Speech response	How much attention was required for producing the speech response (e.g. engaging in a conversation or talk or answering questions)?
14	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)?
15	Past knowledge	How much experience do you have in performing the task or similar tasks on the same website?
16	Skill	Did your skills have no influence (low) or did they help to execute the task (high)?
17	Motivation	Were you motivated to complete the task?
18	Parallelism	Did you perform just this task (low parallelism) or were you doing other parallel tasks (high parallelism) (e.g. multiple tabs/windows/programs)?
19	Arousal	Were you aroused during the task? Were you sleepy/tired (low arousal) or fully awake (high arousal)?

Appendix 5. Likelihood ratio tests for the multinomial logistic regression

Table A6. Likelihood ratio tests of the multinomial logistic regression with the attributes of the NASATLX.

	Model fitting criteria	Likelihood ratio tests		
Effect(s)	− 2 Log 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 A7. Likelihood ratio tests of the multinomial logistic regression with the attributes of the WP.

	Model fitting criteria	Likelihood ratio tests		
Effect(s)	− 2 Log 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 A8. Likelihood ratio tests the multinomial logistic regression with the attributes of MWL_{def} and MWL_{def}^{NI}.

	Model fitting criteria	Likelihood ratio tests		
Effect(s)	− 2 Log 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.

Appendix 6. Step summaries of the multinomial logistic procedure

Table A9. Step summary of the multinomial logistic regression with the attributes of the NASATLX.

			Model fitting criteria	Likelihood ratio tests		
Model	Action	Effect(s)	− 2 Log 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 A10. Step summary of the multinomial logistic regression with the attributes of the WP instrument.

			Model fitting criteria	Likelihood ratio tests		
Model	Action	Effect(s)	− 2 Log 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 A11. Step summary of the multinomial logistic regression with the attributes of MWL_{def} and MWL_{def}^{NI}.

			Model fitting criteria	Likelihood ratio tests		
Model	Action	Effect(s)	− 2 Log 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.