Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile

Lucas Rizzo, Pierpaolo Dondio, Sarah Delany, Luca Longo

Modeling mental workload via rule-based expert system: a comparison with NASA-TLX & Workload Profile

Lucas Rizzo, Pierpaolo Dondio, Sarah Jane Delany, and Luca Longo

School of Computing, Dublin Institute of Technology, Dublin, Ireland
luca.longo@dit.ie

Abstract. In the last few decades several fields have made use of the construct of human mental workload (MWL) for system and task design as well as for assessing human performance. Despite this interest, MWL remains a nebulous concept with multiple definitions and measurement techniques. State-of-the-art models of MWL are usually ad-hoc, considering different pools of pieces of evidence aggregated with different inference strategies. In this paper the aim is to deploy a rule-based expert system as a more structured approach to model and infer MWL. This expert system is built upon a knowledge-base of an expert and translates into computable rules. Different heuristics for aggregating these rules are proposed and they are elicited using inputs gathered in an user study involving humans performing web-based tasks. The inferential capacity of the expert system, using the proposed heuristics, is compared against the one of two ad-hoc models, commonly used in psychology: the NASA-Task Load Index and the Workload Profile assessment technique. In detail, the inferential capacity is assessed by a quantification of two properties commonly used in psychological measurement: sensitivity and validity. Results show how some of the designed heuristics can over perform the baseline instruments suggesting that MWL modelling using expert system is a promising avenue worthy of further investigation.

Keywords: Rule-based expert system, mental workload, heuristics

1 Introduction

Mental workload (MWL) is a multi-faceted phenomenon with no clear and widely accepted definition. Intuitively, it can be described as the amount of cognitive work expended to a certain task during a given period of time. However, this is a simplistic definition and other factors such as stress, time pressure and mental effort can all influence MWL [11]. The principal reason for measuring MWL is to quantify the mental cost of performing a task in order to predict operator and system performance [1]. It is an important construct, mainly used in the fields of psychology and ergonomics, mainly with application in aviation and automobile industries [5, 20] and in interface and web design [23, 16, 15]. According to

Young and Stanton, underload and overload can weaken performance [28]. However, optimal workload has a positive impact on user satisfaction, system success, productivity and safety [12]. Often the information necessary for modelling the construct of MWL is uncertain, vague and contradictory [13]. State-of-the-art measurement techniques do not take into consideration the inconsistency of data used in the modelling phase, which might lead to contradictions and loss of information. For example, if the time spent on a certain task is low it can be derived that the overall MWL is also low, however, if the effort invested in the task is extremely high, then the contrary can be inferred. The aim of this study is to investigate the use of rule-based expert systems for the modelling and inference of MWL. An expert system is a computable program designed to model the problem-solving ability of a human expert [3]. This human expert has to provide a knowledge base, then in turn is translated into computable rules. These rules are used by an inference engine aimed at inferring a numerical index of MWL. Since there is no ground truth indicating if such index is fully correct, the inferential capacity of the defined expert system needs to be investigated in order to gauge its quality. To solve this, the proposal is to adopt some of the most commonly used criteria used in psychometrics such as validity and sensitivity [4, 24, 22]. In simple terms, these criteria are aimed at assessing whether a technique is measuring the construct under investigation and whether it is capable of differentiating variations in workload. From this, the following research question can be defined: *can implementations of rule-based expert systems, compared to state-of-the-art MWL inference techniques, enhance the modelling of mental workload according to sensitivity and validity?*

The remainder of this paper is organised as follows: section 2 describes related works on MWL, its assessment techniques and provides a general view on rule-based expert systems. Section 3 presents the design of an experiment, the methodology adopted. Findings are discussed in section 4 while section 5 concludes our contribution and introduces future work.

2 Related work

2.1 Mental workload assessment techniques

As stated by several authors, there is no simple and agreed definition of mental workload [27, 6, 20]. It is thought to be multidimensional and multifaceted, resulting from the aggregation of many different factors thus difficult to be uniquely defined [1]. The basic intuition is that mental workload is the necessary amount of cognitive work for a person to accomplish a task over a period of time. Nevertheless, a large number of measures have been developed [29, 7] and practitioners have found measuring MWL to be useful [25]. Most empirical classification assessment procedures can be divided in three major categories [19]:

- *Subjective measures*: operators are required to evaluate their own MWL according to different rating scales or a set of questionnaires.

- *Performance-based measures*: these infer an index of MWL from objective notions of performance on the primary task, such as number of errors, completion time or reaction time to respond to secondary tasks.
- *Physiological measures*: these infer a value of MWL according to some physiological response from the operator such as pupillary reflex or muscle activity.

Further details for each category can be found in [17, 5]. This study makes use of two of the subjective measures of MWL that have been largely employed for the last four decades [21, 7, 24]. These are used as base-lines and are: NASA-Task Load Index (TLX) [7] and Workload Profile (WP) [24].

The NASA-TLX is a multidimensional scale, initially developed for the use in the aviation industry. Its application has been spread across several different areas, such as automobile drivers, medical profession, users of computers and military cockpits. Also, it has achieved great importance and is considered a reference point for the development of new measures and models [6]. NASA-TLX consists of six sub scales: mental demand, physical demand, temporal demand, frustration, effort and performance (Table 4, in the Appendix, questions 1-5 plus physical demand). The computation of an overall MWL index is made through a weighted average of these six dimensions d_i quantified using a questionnaire. The weights w_i are provided by the operator according to a comparison of each possible pair of the six dimensions, for example "which contributed more for the MWL: mental demand or effort?", "which contributed more for the MWL: performance or frustration?", giving a total of 15 preferences. The number of times each dimension is chosen defines its weight (equation 1).

The Workload Profile is another MWL assessment technique based on the Multiple Resource Theory (MRT) [26]. In contrast to the NASA-TLX, it is built upon 8 dimensions: perceptual/central processing, response processing, spatial processing, verbal processing, visual processing, auditory processing, manual responses and speech responses (Table 4, question 6-13). The operator is asked to rate the proportion of attentional resources, in the range 0 to 1, for each dimension, then summed. For comparison purpose, this sum is averaged (eq. 2).

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

According to [22] WP is preferred to NASA-TLX if the goal is to compare the MWL of two or more tasks with different levels of difficulty, while NASA-TLX is preferred if the goal is to predict the performance of a particular individual in a single task. Several criteria have been proposed for the selection and development of measurement techniques [19]. In this study the focus is on two of them:

- *validity*: to determine whether the MWL measurement instrument is actually measuring MWL. Two variations of validity are usually employed in psychology: concurrent and convergent. The former aims at determining to

what extent a technique can explain objective performance measures, such as task execution time. The second indicates whether different MWL techniques correlate to each other [24]. In literature, concurrent and convergent validity are calculated adopting statistical correlation coefficients [22, 12].
- *sensitivity*: the capability of a technique to discriminate significant variations in MWL and changes in resource demand or task difficulty [19]. Formally, sensitivity has been assessed in two different ways: multiple regression [24] and ANOVA [22, 12]. The aim was to identify statistically significant differences of the MWL indexes associated to each task under examination.

2.2 Mental workload and rule-based expert system

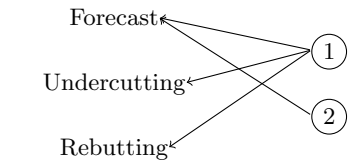
An expert system is a computer program created in order to emulate an expert in a given field [3]. The goal is to imitate the experts capability of solving different tasks in its area. Unlike usual procedural algorithms, an expert system normally has two modules: a *knowledge base* and an *inference engine*. The knowledge base is provided by the expert and translated into a set of rules, which will be utilised by an inference engine. A typical rule is of the form "*IF ... THEN ...*" and the engine will elicit and aggregate all the rules in order to infer a conclusion. In [9], a literature review of many areas in which expert systems have been applied is provided, while [8, 18] are examples of works in the more general field of knowledge representation. To the best of our knowledge, the only study that attempted to model MWL employing inference rules by Longo [10]. Here, modelling MWL has been proposed as a defeasible reasoning process, which is a kind of reasoning built upon inference rules that are defeasible. Defeasible reasoning does not produce a final representation of MWL, but rather a dynamic representation that might change in the light of new evidence and rules. Following this approach, rule-based expert systems might be suitable complements because of their capacity to imitate the problem-solving ability of an expert and facilitate the justification of the inferred conclusion.

3 Design and methodology

In order to answer the research question an experiment is designed as it follows:

1. acquisition of a knowledge base (KB) related to MWL from an expert;
2. KB translation into different types of rule (forecast, undercutting, rebutting)
3. construction of models ($e_1 - e_4$, $fr_1 - fr_4$) based on two variations of KB, each employing different types of rules and heuristics ($H_1, ..., H_4$);
4. comparison of the inferential capacity of each model against selected baseline instruments (NASA-TLX and WP) according to validity and sensitivity:
   - validity is measured to investigate if the implemented rule-based expert system is capable of inferring MWL as well as the baseline instruments.
   - sensitivity is measured to determine the quality of the inference made by the implemented expert system.

Table 1. Experiments set up: types of rules employed by two variations of the same knowledge base (left) and name of each model, variation used, heuristic adopted (right).



Model	KB variation		Heuristics			
	1	2	h_1	h_2	h_3	h_4
e_1	✓		✓			
e_2	✓			✓		
e_3	✓				✓	
e_4	✓					✓
fr_1		✓	✓			
fr_2		✓		✓		
fr_3		✓			✓	
fr_4		✓				✓

3.1 Knowledge base (KB)

Research studies performed by Longo et al. have developed a knowledge base for the inference of MWL in the field of human computer interaction [11, 12, 16]. The goal was to investigate the impact of structural changes of web interfaces on the imposed mental workload on end-users after interacting with them. The knowledge base developed comprises by 21 attributes (Table 4), containing a set of features believed to be useful for modelling MWL, each of them quantified, through a subjective question, in the range [0, 100] $\in \mathbb{R}$. The MWL has four possible levels, as per definition 1.

Definition 1 *(Mental workload level) Four MWL levels are defined: underload (U), fitting$^-$ (F^-), fitting$^+$ (F^+) and overload (O).*

The set of rules built from the knowledge-base of the expert [11] can be seen in the Appendix and a formal definition follows.

Definition 2 *(Rules) Three types of rules are defined.*

- *Forecast rule (FR): takes a value α of an attribute X and infers a MWL level β if α is in a predefined range $[x_1, x_2]$ with $x_1, x_2 \in \mathbb{N}$ and $x_2 > x_1$.*

$$FR: \textbf{\textit{IF}}\ \alpha \in [x_1, x_2]\ \textbf{\textit{THEN}}\ \beta$$

- *Undercutting rule (UR): takes one or more attributes values, $\alpha_1, \cdots, \alpha_n$, and undercuts what is inferred by a forecast rule Y if $\alpha_1 \in [x_1^1, x_2^1], \cdots, \alpha_n \in [x_1^n, x_2^n]$. In this case it is said that rule Y is discarded, d(Y), and will not be considered for future inferences of MWL.*

$$UR: \textbf{\textit{IF}}\ \alpha_1 \in [x_1^1, x_2^1]\ \textbf{\textit{and}}\ \cdots\ \textbf{\textit{and}}\ \alpha_n \in [x_1^n, x_2^n]\ \textbf{\textit{THEN}}\ d(Y)$$

- *Rebutting rule (RR): is a relationship between two forecast rules, Y_1 and Y_2, that can not coexist.*

$$RR: \textbf{\textit{IF}}\ Y_1\ \textbf{\textit{and}}\ Y_2\ \textbf{\textit{THEN}}\ d(Y_1)\ \textbf{\textit{and}}\ d(Y_2)$$

Example 1 *An example of possible rules are:*
- Forecast rules
*EF1: [**IF** effort* $\in [0, 32]$ ***THEN** U]* *EF4: [**IF** effort $\in$ [67, 100] **THEN** O]*
*MD1: [**IF** mental demand $\in$ [0, 32] **THEN** U]*
*PK1: [**IF** past knowledge* $\in [0, 32]$ ***THEN** O]*
- Undercutting rule
*DS1: [**IF** task difficulty $\in$ [67, 100] **and** skills $\in$ [67, 100] **THEN** d(EF4)]*
*- Rebutting rule - r5: [**IF** PK1 **and** EF1 **THEN** d(PK1) **and** d(EF1)]*

3.2 Inference engine

Having defined the set of rules, the next step for inferring MWL is to implement an inference engine. Our inference engine starts with the activation of rules in the set of FR. These will be called *activated rules*. This activation is based on the inputs provided by the user. Afterwards, rules from the set of UR and RR might discard activated rules, solving some part of the contradictory information. This step is not compulsory. The implementation of rule-based expert systems without UR and RR is also provided. Activated rules that are not discarded are called *surviving rules*. After defining the set of surviving rules, there still might be some inconsistent inferences. Surviving rules will likely be inferring different MWL levels, even with the application of UR and RR. The expert system, therefore, must be able to aggregate the surviving rules and produce a final inference of MWL. Next an example follows:

Example 2 *Following rules from Example 1 and given a numerical input it is possible to define the set of activated rules and the set of surviving rules.*

- ***Inputs**: [effort = 80, past knowledge = 15, task difficulty = 90, mental demand = 20, skills = 70, temporal demand = 10]*
- ***Rules**:* Activated*: [EF4, PK1, MD1, TD1, DS1]* Discarded*: [EF4]* Surviving*: [PK1, MD1, TD1]*

Example 2 illustrates a set of surviving rules inferring underload MWL (MD1, TD1) and overload MWL (PK1) at the same time. At this stage, a typical set of conflict resolutions strategies for expert systems include: deciding a priority for each rule, firing all possible lines of reasoning or choosing the first rule addressed. However, none of these strategies is applicable in our experiment, since there is no preference among rules, order of evaluation or possibility to compute more than one output. The knowledge base does not provide sufficient information for performing this computation and because of that four heuristics are defined to accomplish the aggregation of the surviving rules. The strategies are developed in order to extract different pieces of information from the surviving rules, which are aggregated or not in different fashions. The final MWL will be a value in the range [0, 100] $\in \mathbb{R}$. Before presenting such heuristics it is necessary to define the value of a surviving rule (definition 3).

Definition 3 *(Surviving rule value) The value of a surviving rule $r \in FR$, with input $0 \leq \alpha \leq 100$ related to attribute X, is given by the function*

$$f(r) = \begin{cases} \alpha, & if\ X \propto MWL \\ 100 - \alpha, & if\ X \propto \frac{1}{MWL} \end{cases}$$

with $X \propto MWL$ a direct relationship, $X \propto \frac{1}{MWL}$ an inverse relationship[1].

Given Definition 3 the following heuristics are designed:

- $\boldsymbol{h_1}$: the average of the surviving rules of the MWL level with the largest cardinality of surviving rules. In case of two or more levels with equal cardinality, it computes the mean of the averages. The idea is to give importance to the largest point of view (largest set of surviving rules) to infer MWL.
- $\boldsymbol{h_2}$: the highest average value of the surviving rules for each MWL level. This is a pessimistic point of view, and infers the highest MWL according to the different sets of surviving rules of each MWL level.
- $\boldsymbol{h_3}$: average value of all surviving rules. This is to give equal importance to all surviving rules, regardless of which level of MWL they were supporting.
- $\boldsymbol{h_4}$: average of average of surviving rules of each MWL level. This is to give equal importance to all sets of MWL levels.

Example 3 *Following Example 2, the value of the surviving rules is given by $f(PK1) = 85$, $f(MD1) = 20$ and $f(TD1) = 10$. Finally, the overall MWL computed by each heuristic is: h_1: $\frac{20+10}{2} = 15$, h_2: $max(85, \frac{20+10}{2}) = 85$, h_3: $\frac{20+10+85}{3} = 38.3$ and h_4: $\frac{\frac{20+10}{2}+85}{2} = 50$.*

4 Data collection, elicitation of models and evaluation

Nine information seeking web-based tasks of varying difficulty and demand (Table 3), were performed by participants over three websites: Google, Wikipedia and Youtube. Two alterations of the interface of each web-site were proposed, having overall (9x2=18) configurations. 40 volunteers performed 9 tasks (on a random alteration) and after each, they answered each question of Table 4 using a paper-based scale in the range [0..100] $\in \aleph$, partitioned in 3 regions delimited at 33 and 66. Due to loss of data or partial completion of questionnaires, 406 instances were valid. Collected answers, for each instance, were used to elicit the rules of each model (section 3), aggregated with their heuristic, that in turn, produced an index of MWL, in the scale [0..100] $\in \Re$. The outputs formed a distribution of MWL indexes, one for each model, and these were compared against the ones of the baseline models according to validity and sensitivity (fig. 1).

[1] Only the attributes past knowledge, skills and performance of Table 4 have an inverse relationship with MWL (the higher the answer the lower the MWL level) while the others have a direct relationship.

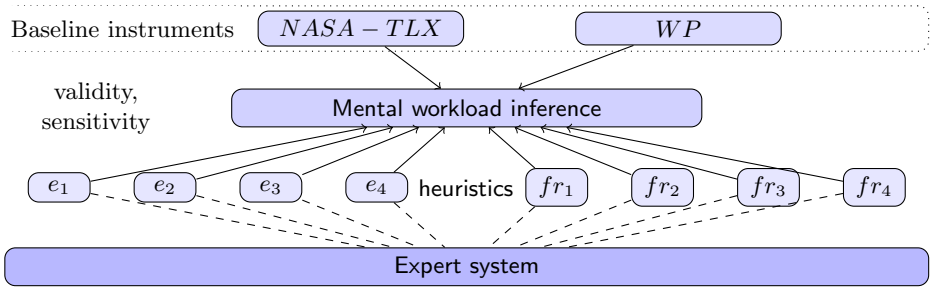


Fig. 1. Evaluation strategy schema

4.1 Validity

In line with other studies [22, 12], validity was assessed using correlation coefficients. In order to select the most suitable statistic, a test of the normality of the distributions of the MWL indexes, produced by each model, was performed using the Shapiro-Wilk test. This test did not achieve a significance greater than 0.05 for most of the models, underlying the non normality of data. As a consequence, the Spearman's rank-order correlation was selected.

Convergent validity: aimed at determining to what extent a model correlate with other model of MWL. As it can be seen from figure 2, the baseline instruments (NASA-TLX and WP) achieved a correlation of .538 (dashed reference line) with each other. When correlated with NASA-TLX, e_3 and fr_3 obtained a higher correlation than this. These two models both apply the heuristic h_3, which is the average of all surviving rules, a similar computational method used by the baseline instruments. Just in two other cases (e_1, fr_1) a good correlation (close to the reference line) with WP was obtained. These 2 models implement heuristic h_1, which is the average of the surviving rules of the MWL level (set of rules) with the largest cardinality. The above 4 cases demonstrate how models can be built using rule-based expert system showing similar validity than other baseline MWL assessment instruments believed to shape the construct of MWL.

Concurrent validity: aimed at determining the extent to which a model correlate with task completion time (objective performance measure)[2]. From figure 3, it is possible to note that even the baseline instruments do not have a high correlation with task completion time. The first dashed line represents the correlation of 0.178 between NASA-TLX and Time while the second represents the correlation of 0.119 between WP and Time. Similarly to convergent validity, the models applying heuristic h_3 (e_3, and fr_3) plus the model e_2 were the ones that better correlated with task completion time, figure 3, over performing the NASA-TLX. Almost all the models over performed also the WP baseline. These findings suggest that computational models of MWL can be built as rule-based expert systems, and these are capable of enhancing the concurrent validity of the assessments when compared with state-of-the-art models.

[2] Due to measurement errors, only 281 instances have an associated time.

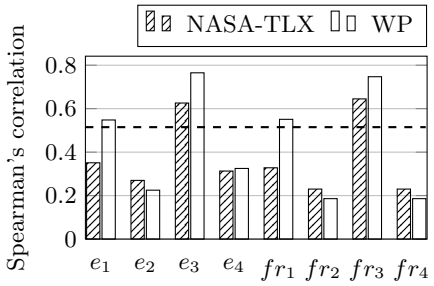


Fig. 2. Convergent validity: $p < 0.05$.

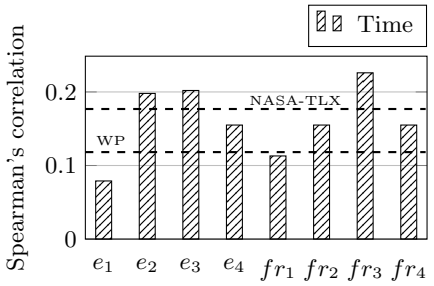


Fig. 3. Concurrent validity: $p < 0.05$.

4.2 Sensitivity

In line with other studies [22, 12], sensitivity was assessed by analysis of variance. In particular, the non-parametric Kruskal-Wallis H test was performed over the MWL distributions generated by each model, and this was selected because some of the assumptions behind the equivalent of one-way ANOVA were not met. Only model e_4 was not capable of rejecting the null hypothesis of same distribution of MWL indexes across tasks ($p < 0.01$). This means that, for the other models, statistical significant differences exist. The Kruskal-Wallis H test, however, does not tell exactly which pairs of tasks are different from each other. As a consequence, post hoc analysis was performed and the Games-Howell test was chosen because of unequal variances of the distributions under analysis. Table 2 depicts how many pairs of tasks each model was capable of differentiating at different significance levels ($p < 0.05$ and $p < 0.01$). As is can be observed, models applying heuristic h_3 (fr_3 and e_3) outperformed the WP but underperformed the NASA-TLX. This result is a confirmation that sensitive mental workload rule-based expert systems can be successfully built and compete with existing benchmarks in the field.

Table 2. Sensitivity of MWL models with Games-Howell post hoc analysis. The maximum pairwise comparisons of 9 tasks is $\binom{9}{2} = 36$).

Model	$p < 0.05$	$p < 0.01$
NASA-TLX	18	12
WP	9	4
e_1	2	1
e_2	5	3
e_3	13	10
e_4	0	0

Model	$p < 0.05$	$p < 0.01$
NASA-TLX	18	12
WP	9	4
fr_1	2	0
fr_2	4	1
fr_3	17	10
fr_4	4	1

4.3 Summary of findings

Quantifications of the validity and the sensitivity of developed models suggest that rule-based expert systems can be successfully built for mental workload modelling and assessment because their inferential capacity lies between the inferential capacity of two state-of-the-art assessment instruments, namely the Nasa Task Load Index and the Workload profile. However, here it is argued that these systems are more appealing and dynamic than selected state-of-the-art approaches. Firstly, they use rules built with terms that are closer to the way humans reason and that imitate experts problem-solving ability. Secondly, they embed heuristics for aggregating rules in a more dynamic way, with a better capacity of handling uncertainty and conflicting pieces of information compared to fixed formulas of state-of-the-art models. Thirdly, they allow the comparison of knowledge-bases and beliefs of different MWL designers thus increasing the understanding of the construct of Mental Workload itself.

5 Conclusion and future work

This research presents a new way of modelling and assessing the construct of Mental Workload (MWL) by means of rule-based expert systems. A knowledge base of a MWL designer was elicited and translated into computational rules of various typology. Different heuristics for aggregating these rules were designed aimed at inferring MWL as a numerical index. Inferred indexes were systematically compared with those generated by two state-of-the-art MWL assessment techniques: the NASA Task Load Index and the Workload Profile. This comparison included the quantification of two properties of each distribution of MWL indexes, namely sensitivity and validity, commonly employed in the literature. Findings suggest that rule-based expert systems are promising not only because they can approximate the inferential capacity of selected state-of-the-art MWL assessment techniques. They also offer a flexible approach for translating different knowledge-bases and beliefs of MWL designers into computational rules supporting the creation of models that can be replicated, extended and falsified, thus enhancing the understanding of the construct of mental workload itself. Future works will be focused on the replication of the approach adopted in this study using other knowledge bases elicited from other MWL experts. Additionally, this approach will be extended incorporating fuzzy representation of rules and acceptability semantics, borrowed from argumentation theory [2, 14], with the aim of improving conflict resolution of rules and building models expected to have an even higher sensitivity and validity.

Acknowledgments

Lucas Middeldorf Rizzo would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for his Science Without Borders scholarship, proc n. 232822/2014-0.

References

1. Cain, B.: A review of the mental workload literature. Tech. rep., Defence Research and Development Canada Toronto, Human System Integration Section (2007)
2. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence 77(2), 321–357 (1995)
3. Durkin, J., Durkin, J.: Expert systems: design and development. Prentice Hall PTR (1998)
4. Eggemeier, F.T.: Properties of workload assessment techniques. Advances in Psychology 52, 41–62 (1988)
5. Gartner, W.B., Murphy, M.R.: Pilot workload and fatigue: A critical survey of concepts and assessment techniques. National Aeronautics Space performance (1976)
6. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50, pp. 904–908. Sage Publications (2006)
7. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. Advances in psychology 52, 139–183 (1988)
8. Hatzilygeroudis, I., Prentzas, J.: Integrating (rules, neural networks) and cases for knowledge representation and reasoning in expert systems. Expert Systems with Applications 27(1), 63–75 (2004)
9. Liao, S.H.: Expert system methodologies and applicationsa decade review from 1995 to 2004. Expert systems with applications 28(1), 93–103 (2005)
10. Longo, L.: Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In: User Modeling, Adaptation, and Personalization, pp. 369–373. Springer (2012)
11. Longo, L.: Formalising Human Mental Workload as a Defeasible Computational Concept. Ph.D. thesis, Trinity College Dublin (2014)
12. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behaviour and Information Technology 34(8), 758–786 (2015)
13. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Intelligent Information and Database Systems, Part II. pp. 65–74 (2010)
14. Longo, L., Dondio, P.: Defeasible reasoning and argument-based medical systems: an informal overview. In: 27th International Symposium on Computer-Based Medical Systems, New York, USA. pp. 376–381. IEEE (2014)
15. Longo, L., Dondio, P.: On the relationship between perception of usability and subjective mental workload of web interfaces. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, Volume I. pp. 345–352 (2015)
16. Longo, L., Rusconi, F., Noce, L., Barrett, S.: The importance of human mental workload in web-design. In: 8th International Conference on Web Information Systems and Technologies. pp. 403–409 (April 2012)
17. Meshkati, N., Hancock, P.A., Rahimi, M., Dawes, S.M.: Techniques in mental workload assessment. In: Wilson, J.R., Corlett, E.N. (eds.) Evaluation of Human Work: A Practical Ergonomics Methodology, pp. 749–782. Taylor & Francis (1995)
18. Mitra, R.S., Basu, A.: Knowledge representation in mickey: An expert system for designing microprocessor-based systems. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 27(4), 467–479 (1997)

19. O'Donnell, R.D., Eggemeier, F.T.: Workload assessment methodology. In: Boff, K.R., Kaufman, L., Thomas, J.P. (eds.) Handbook of perception and human performance, vol. 2, chap. 42, pp. 1–49. John Wiley & Sons (1986)
20. Paxion, J., Galy, E., Berthelon, C.: Mental workload and driving. Frontiers in psychology 5,1344 (2014)
21. Reid, G.B., Nygren, T.E.: The subjective workload assessment technique: A scaling procedure for measuring mental workload. Advances in psychology 52, 185–218 (1988)
22. Rubio, S., Díaz, E., Martín, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. Applied Psychology 53(1), 61–86 (2004)
23. Tracy, J.P., Albers, M.J.: Measuring cognitive load to test the usability of web sites. In: Annual Conference-society for technical communication. vol. 53, pp. 256–260 (2006)
24. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
25. Tsang, P.S., Wilson, G.F.: Mental workload measurement and analysis. In: Salvendy, G. (ed.) Handbook of human factors and ergonomics, chap. 13, pp. 417–449. Wiley & Sons, 2nd edn. (1997)
26. Wickens, C.D.: Processing resources and attention. Multiple-task performance pp. 3–34 (1991)
27. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics 58(1), 1–17 (2015)
28. Young, M.S., Stanton, N.A.: Attention and automation: new perspectives on mental underload and performance. Theoretical Issues in Ergonomics Science 3(2), 178–194 (2002)
29. Young, M., Stanton, N.: Mental workload: theory, measurement, and application. In: International encyclopedia of ergonomics and human factors, vol. 1, pp. 507–509. London: Taylor and Francis (2001)

Appendix

Knowledge base

For the attribute mental demand the forecast rules are:

MD1: [**IF** mental demand ∈ [0, 32] **THEN** U]
MD2: [**IF** mental demand ∈ [33, 49] **THEN** F^-]
MD3: [**IF** mental demand ∈ [50, 66] **THEN** F^+]
MD4: [**IF** mental demand ∈ [67, 100] **THEN** O]

The same principle applies to the attributes temporal demand, physical demand, solving and deciding, selection of response, task and space, verbal material, visual resources, auditory resources, manual response, speech response, effort, parallelism, and context bias, forming 52 other rules. For psychological stress, motivation, past knowledge, skills and performance the forecast rules are:

PS1: [**IF** psychol. stress ∈ [0, 32] **THEN** U]
PS2: [**IF** psychol stress ∈ [67, 100] **THEN** O]
MV1: [**IF** motivation ∈ [0, 32] **THEN** U]
PK1: [**IF** past knowledge ∈ [0, 32] **THEN** O]
PK2: [**IF** past knowledge ∈ [67, 100] **THEN** U]
SK1: [**IF** skills ∈ [0, 32] **THEN** O]
SK2: [**IF** skills ∈ [67, 100] **THEN** U
PF1: [**IF** performance ∈ [0, 32] **THEN** O]
PF2: [**IF** performance ∈ [33, 49] **THEN** F^+]
PF3: [**IF** performance ∈ [50, 66] **THEN** F^-]
PF4: [**IF** performance ∈ [67, 100] **THEN** U]

The undercutting rules and rebutting rules are:

AD1a: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [0, 32] **THEN** d(PF4)]

AD1b: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [0, 32] **THEN** d(PF3)]
AD1c: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [0, 32] **THEN** d(PF2)]
AD2a: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [67, 100] **THEN** d(PF4)]
AD2b: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [67, 100] **THEN** d(PF3)]
AD2c: [**IF** arousal ∈ [0, 32] **and** task difficulty ∈ [67, 100] **THEN** d(PF2)]
AD3a: [**IF** arousal ∈ [33, 49] **and** task difficulty ∈ [0, 32] **THEN** d(PF1)]
AD3b: [**IF** arousal ∈ [33, 49] **and** task difficulty ∈ [0, 32] **THEN** d(PF4)]
AD4a: [**IF** arousal ∈ [33, 49] **and** task difficulty ∈ [67, 100] **THEN** d(PF1)]
AD4b: [**IF** arousal ∈ [33, 49] **and** task difficulty ∈ [67, 100] **THEN** d(PF3)]
AD4c: [**IF** arousal ∈ [33, 49] **and** task difficulty ∈ [67, 100] **THEN** d(PF4)]
AD4d: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [67, 100] **THEN** d(PF1)]
AD4e: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [67, 100] **THEN** d(PF3)]
AD4f: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [67, 100] **THEN** d(PF4)]
AD5a: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [0, 32] **THEN** d(PF1)]
AD5b: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [0, 32] **THEN** d(PF2)]
AD5c: [**IF** arousal ∈ [50, 66] **and** task difficulty ∈ [0, 32] **THEN** d(PF3)]
AD5d: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [0, 32] **THEN** d(PF1)]
AD5e: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [0, 32] **THEN** d(PF2)]
AD5f: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [0, 32] **THEN** d(PF3)]
AD6a: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [67, 100] **THEN** d(PF2)]
AD6b: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [67, 100] **THEN** d(PF3)]
AD6c: [**IF** arousal ∈ [67, 100] **and** task difficulty ∈ [67, 100] **THEN** d(PF4)]
MV2: [**IF** motivation ∈ [0, 32] **THEN** d(EF3)] - MV3: [**IF** motivation ∈ [0, 32] **THEN** d(EF4)]
MV4: [**IF** motivation ∈ [67, 100] **THEN** d(EF1)] - MV5: [**IF** motivation ∈ [67, 100] **THEN** d(EF2)]
DS1: [**IF** task difficulty ∈ [67, 100] **and** skills ∈ [67, 100] **THEN** d(EF4)]
DS2: [**IF** task difficulty ∈ [67, 100] **and** skills ∈ [67, 100] **and** effort ∈ [0, 32] **THEN** d(PF1)]
DS3: [**IF** task difficulty ∈ [67, 100] **and** skills ∈ [67, 100] **and** effort ∈ [33, 49] **THEN** d(PF1)]
DS4: [**IF** task difficulty ∈ [67, 100] **and** skills ∈ [67, 100] **and** effort ∈ [50, 66] **THEN** d(PF1)]
r1: [**IF** MD1 **and** SD4 **THEN** d(MD1), d(SD4)] - r2: [**IF** MD4 **and** SD1 **THEN** d(MD4), d(SD1)]
r3: [**IF** PK1 **and** SK4 **THEN** d(PK1), d(SK4)] - r4: [**IF** PK4 **and** SK1 **THEN** d(PK4), d(SK1)]
r5: [**IF** PK1 **and** EF4 **THEN** d(PK1, d(EF1)] - r6: [**IF** PK2 **and** EF4 **THEN** d(PK2), d(EF4)]
r7: [**IF** SK1 **and** EF1 **THEN** d(SK1), d(EF1)] - r8: [**IF** SK4 **and** EF4 **THEN** d(SK4), d(EF4)]
r9: [**IF** CB4 **and** PS1 **THEN** d(CB4), d(PS1)]

Tasks and Questionnaire

Table 3. List of experimental tasks

Task	Description	Task condition	Web-site
T_1	Find out how many people live in Sidney	Simple search	Wikipedia
T_2	Read `simple.wikipedia.org/wiki/Grammar`	No goals, no time pressure	Wikipedia
T_3	Find out the difference (in years) between the year of the foundation of the Apple Computer Inc. and the year of the 14^{th} FIFA world cup	Dual-task and mental arithmetical calculations	Google
T_4	Find out the difference (in years) between the foundation of the Microsoft Corp. & the year of the 23^{rd} Olympic games	Dual-task and mental arithmetical calculations	Google
T_5	Find out the year of birth of the 1^{st} wife of the founder of playboy	Single task + time pressure (2-min limit). Each 30 secs user is warned of time left	Google
T_6	Find out the name of the man (interpreted by Johnny Deep) in the video `www.youtube.com/watch?v=FfTPS-TFQ_c`	Constant demand on visual and auditory modalities. Participant can replay the video if required	Youtube
T_7	a) Play the song `www.youtube.com/watch?v=Rb5G1eRIj6c`. 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`	Demand on visual modality and inference on auditory modality. The song is extremely irritating	Wikipedia
T_8	Find out how many times Stewie jumps in the video `www.youtube.com/watch?v=TSe9gbdkQ8s`	Demand on visual resource + external interference: user is distracted twice & can replay video	Youtube
T_9	Find out the age of the blue fish in the video `www.youtube.com/watch?v=H4BNbHBcnDI`	Demand on visual and auditory modality, plus time-pressure: 150-sec limit. User can replay the video. There is no answer.	Youtube

Table 4. Experimental study questionnaire [11]

Dimension	**Question**
Mental demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy (low mental demand) or complex (high mental demand)?
Temporal demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely (low temporal demand) or rapid and frantic (high temporal demand)?
Effort	How much conscious mental effort or concentration was required? Was the task almost automatic (low effort) or it required total attention (high effort)?
Performance	How successful do you think you were in accomplishing the goal of the task? How satisfied were you with your performance in accomplishing the goal?
Frustration	How secure, gratified, content, relaxed and complacent (low psychological stress) versus insecure, discouraged, irritated, stressed and annoyed (high psychological stress) did you feel during the task?
Selection of response	How much attention was required for selecting the proper response channel and its execution? (manual - keyboard/mouse, or speech - voice)
Task and space	How much attention was required for spatial processing (spatially pay attention around you)?
Verbal material	How much attention was required for verbal material (eg. reading or processing linguistic material or listening to verbal conversations)?
Visual resources	How much attention was required for executing the task based on the information visually received (through eyes)?
Auditory resources	How much attention was required for executing the task based on the information auditorily received (ears)?
Manual Response	How much attention was required for manually respond to the task (eg. keyboard/mouse usage)?
Speech response	How much attention was required for producing the speech response(eg. engaging in a conversation or talk or answering questions)?
Context bias	How often interruptions on the task occurred? Were distractions (mobile, questions, noise, etc.) not important (low context bias) or did they influence your task (high context bias)?
Past knowledge	How much experience do you have in performing the task or similar tasks on the same website?
Skill	Did your skills have no influence (low) or did they help to execute the task (high)?
Solving and deciding	How much attention was required for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognizing and identifying objects)?
Motivation	Were you motivated to complete the task?
Parallelism	Did you perform just this task (low parallelism) or were you doing other parallel tasks (high parallelism) (eg. multiple tabs/windows/programs)?
Arousal	Were you aroused during the task? Were you sleepy, tired (low arousal) or fully awake and activated (high arousal)??
Task difficult	Task difficult was given by the formula: $\text{Task}_{\text{difficult}} = \frac{1}{8}$((solving/deciding) + (auditory resources) + (manual response) + (speech response) + (response) + (task/space) + (verbal material) + (visual resources))
Physical demand	The physical demand was considered 0 for all instances