2015

On the Relationship Between Perception of Usability and Subjective Mental Workload of Web Interfaces

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

Pierpaolo Dondio
Technological University Dublin, pierpaolo.dondio@tudublin.ie

Recommended Citation

Longo, L. & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, p. 345-352. doi:10.1109/WI-IAT.2015.157

On the relationship between perception of usability and subjective mental workload of web interfaces

Luca Longo, Pierpaolo Dondio
School of Computing, Dublin Institute of Technology, Dublin, Ireland
luca.longo@dit.ie

***Abstract*—Inspection methods and cheap self-reporting procedures have been significantly employed in the field of Human-Computer Interaction for assessing the usability of interfaces, systems and technologies. However, there is a tendency to overlook aspects related to the context and features of the users during the usability assessment process. This research introduces the concept of mental workload as an aid to enhance usability measurement. A user-study has been designed and executed in the context of human-web interaction. The aim was to investigate the relationship between the perception of usability of three popular web-sites, and the mental workload imposed by a set of typical tasks executed over them. Scores obtained with the System Usability Scale were compared to the mental workload scores obtained from the NASA Task Load Index and the Workload Profile assessment procedures. Findings suggest that perception of usability and mental workload are likely to be two non-overlapping constructs, and there is no clear evidence of their interaction. They measure two different aspects of human-system interaction and therefore they could be jointly employed to better describe user experience.**

Keywords*—*Usability, Mental Workload, Web-design, A/B testing

I. Introduction

In recent decades the demands of evaluating usability of interactive systems have produced several assessment procedures, which have been applied significantly in various contexts. However, there has been a tendency to overlook aspects of the context and characteristics of the users during the usability assessment process. For instance, assessing usability in testing environments is different to assessing it in operational environments. Similarly, a skilled person is likely to perceive the usability of an interface differently to someone who is inexperienced; also easy and difficult tasks might affect perception of usability differently. One of the main reasons why these aspects are often overlooked is because there is no cohesive model that considers the context of use, the features of users and the characteristics of tasks. Studies suggest that considering the context is fundamental for inferring robust and significant assessments of usability [25] [3]. Similarly, considering features and characteristics of users for enhancing the design of interactive systems is a central notion for the User Modeling community, an important discipline within Human-Computer Interaction [9] [1] [22], [21]. Another important factor that we believe is worthy of consideration in system design, is the concept of human Mental Workload (MWL) [19], [20] borrowed from the disciplines of Ergonomics and Human Factors, with roots in Psychology. This construct is often referred to as Cognitive Workload and is strictly connected to the notion of performance. Assessing mental workload is key to measuring performance. Several MWL assessment procedures have been proposed but no generally applicable measure has yet been devised. Recent studies have tried to adopt the concept of MWL to explain usability [34] [2]. Despite this interest, not much has yet been done to link these two concepts together and to investigate their relationship empirically. The aim of this research is to shed some light on how these two concepts are linked.

This paper is organised as follows: Firstly, notable definitions of usability and mental workload are provided, followed by an overview of the assessment techniques employed in the field of HCI. Related studies are also presented, highlighting how the constructs of mental workload have been employed so far, distinctly and jointly with usability. An experiment is subsequently designed in the context of human-web interaction, aimed at investigating the relationship between the perception of usability of three popular web-sites and the mental workload experienced by users after interacting with them. Results are presented and critically discussed, showing how usability and mental workload are related. A summary of this study concludes the paper describing future work and presenting recommendations.

II. Usability and mental workload

A. Definition of usability

Usability is a widely used notion and it has been defined in several ways [32]. The amount of literature covering definitions, frameworks and methodologies for assessing usability is vast [16] and it would be not possible to list everything here. The ISO (International Organisation for Standardisation) defines usability as 'The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use'. Usability, according to Nielsen's, is a method for improving ease-of-use in the design of interactive systems and technologies [26]. It is a notion that embraces other concepts such as learnability, user satisfaction and efficiency and it is often associated with the functionalities of a product rather than being merely a feature of the user interface [27].

B. Measure of usability

Often when selecting an appropriate procedure in the context of web-design and web-based systems, it is desirable to take into account the effort and expense that will be incurred in collecting and analysing data. For this reason, web-designers have tended to adopt subjective usability assessment techniques for collecting feedback from users [16]. On one hand,

these self-reporting techniques can only be administered post-task, thus influencing their reliability with regard to long tasks. Meta-cognitive limitations can also diminish the accuracy of reporting and it is difficult to perform comparisons among raters on an absolute scale. However, on the other hand, these techniques also appear to be the most sensitive and diagnostical procedures [16]. Nielsen's principles represent the most widely adopted heuristic to test the usability of an interface due to their simplicity in terms of effort and time [26]. The evaluation is done by systematically finding usability problems in an interface and judging them according to the usability principles in an iterative design process [27]. However, the heuristics mainly focus on the user interface without considering external factors such as the environment, the context of use and the cognitive state of the users.

The System Usability Scale [4] is a questionnaire developed at Digital Equipment Corporation and consists of ten questions (table IX). It is a 'quick and dirty' tool for measuring usability and has been applied and cited over thousands of articles and publications. It is a very easy scale to administer, it has proved useful for distinguishing usable and unusable systems and it can be reliably employed even on small sample sizes [38]. The Computer System Usability Questionnaire (CSUQ) developed at IBM consists of 19 questions on a seven-point Likert scale of 'strongly disagree' to 'strongly agree' [18]. It is an easy scale to administer. The Questionnaire for User Interface Satisfaction (QUIS) is a technique developed at the HCI lab at the University of Maryland and was designed to assess users' satisfaction with aspects of a computer interface [33]. It is not as easy as the previous procedures because it includes a demographic questionnaire, a measure of system satisfaction along six scales, and it has a hierarchy of measures of nine specific interface factors (screen factors, terminology and system feedback, learning factors, system capabilities, technical manuals, on-line tutorials, multimedia, teleconferencing, and software installation). Each of these relates to a user's satisfaction with that particular aspect of an interface as well as to the factors that make up that facet, on a 9-point scale. Although it is more complex than other instruments, this tool has shown high reliability across several interfaces [13]. Many other usability inspection methods, frameworks and questionnaires have been proposed in the literature [16] [38]. The information provided so far is sufficient for the remainder of this paper and the System Usability Scale is the tool that has been adopted for the experimental user study.

C. Definition of mental workload

Human Mental Workload (MWL) is an important design concept and it is fundamental for investigating the interaction of people with computers and other technological devices [19]. It has a long history in the disciplines of Ergonomics, Human Factors and Psychology, with several applications in the aviation [15], [14] and automotive industries [8]. It has been extensively documented that mental overload or underload can both negatively affect performance [43]. On one hand, 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 [14]. Unfortunately, although it has been studied for the last five decades, no clear definition of MWL has emerged that has a general validity and that is universally accepted. MWL can be intuitively defined as the amount of cognitive work necessary for a person to complete a task over 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' [15]. The main reason for assessing MWL is to measure the cognitive cost of performing a task for operator/system performance prediction [5].

D. Measures of mental workload

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 multi-dimensional property that characterises mental workload [21]. Several reviews attempted to collate the enormous amount of knowledge behind measurement procedures. According to [10] measurements can be divided into subjective, performance, arousal, specific measures and psychophysiological measures. [43] 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 [44] [35] [42] [5] [45]: a) *self-assessment measures* or self-report measures and subjective rating scales; b) *performance measures* which consider primary and secondary task measures; c) *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 relies on the subjective perceived experience of the interaction operator-system. They have always appealed workload practitioners because it is strongly believed that only the person concerned with the task can provide an accurate judgement with respect to the MWL experienced. This category consider various factors believed to influence MWL: effort, performance, individual differences such as the emotional state, attitude and motivation of the operator [8]. The class of *performance measures* assumes that the MWL of an operator, interacting with a system, acquires importance only if it influences system performance. In turn, it is believed that this class of techniques is the most valuable options for designers [37] [31]. 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. Their main advantage is that 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 [28]. However, they require specific equipment and well trained operators. Self-assessment measures have been adopted for the user study planned in this research mainly for their ease of use.

III. Related work

This section mainly focuses on related work in the field of mental workload with HCI applicability rather than on usability applications. The reason is that the former topic is relatively new in the design of web interfaces, while the latter have already been investigated extensively for many years [16] [38]. In the context of web-design, MWL is believed to be 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 [43]. Modern technologies and web applications have become increasingly complex [24] with increments in the degree of MWL imposed on operators [11] [12]. The assumption in design approaches is that, on one hand, as the difficulty of a task increases, perhaps due to interface complexity, MWL also increases and performance usually decreases [5]. In turn, errors are more frequent, there are longer response times, and fewer tasks are completed per time unit [17]. 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 [5]. Subjective measures of MWL include multi-dimensional approaches such as NASA's Task Load Index ($NASATLX$) [15], the Subjective Workload Assessment Technique [29], the Workload Profile [36] as well as uni-dimensional measures such as the Copper-Harper scale [7], the Rating Scale Mental Effort [47], the Subjective Workload Dominance Technique [39] and the Bedford scale [30]. These subjective techniques have low implementation requirements along with low intrusiveness and high subject acceptability. This has led to them being used in new research in which the construct of MWL has been adopted for evaluating alternative interfaces [19]. For instance, the $NASATLX$ has been used for evaluating user interfaces in health-care [23]. Similarly, the Workload Profile [36], the NASATLX and the Subjective Workload Assessment Technique [29] have been compared in a user study to evaluate different web interfaces [24]. Tracy and Albers adopted three different techniques for measuring mental workload applied to web-site design: $NASATLX$, the Sternberg Memory Test and a tapping test [34] [2]. They proposed a technique to individuate sub-areas of a web-site, in which end-users manifested a higher mental workload during interaction. In turn, this allowed designers to modify those critical regions for enhancing their interface. [46] noted how roles can be useful in interface design and proposed a role-based method to measure MWL. This can be applied in the field of HCI for dynamically adjusting human workload levels to enhance performance in interaction.

IV. Design of experiments

A user study has been designed to investigate the relationship between perception of usability and the perception of mental workload. The following self-reporting assessment instruments have been adopted:

- the System Usability Scale (SUS) [4]
- the Nasa Task Load Index ($NASATLX$), developed by the Human Performance Group at NASA [14].
- the Workload Profile (WP) [36], based on the multiple resource theory [41], [40]

A. The NASA Task Load Index

The NASA Task Load Index instrument [14] belongs to the category of performance measures and it is a combination of six factors believed to influence mental workload: mental, temporal and physical demand, stress, effort and performance. Each factors is quantified with a subjective judgement (questions of table X) coupled with a weight computed via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial coefficient) of the 6 factors, *'which of the two contributed the most to mental workload during the task'*, such as 'Mental or Physical Demand?', 'Physical Demand or Performance?', and so forth.

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

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.

$$NASATLX : [0..1] \in \Re \quad NASATLX = \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15}$$

For comparison purpose, the value is converted in $[0..100] \in \Re$.

B. The Workload Profile

The Workload Profile (WP) assessment procedure [36], is built upon the multiple resource theory proposed in [41], [40]. In this theory, individuals are seen as having different capacities of 'resources':

- *stage of information processing*: perceptual/central processing and response selection and execution
- *code of information processing*: spatial/verbal
- *input*: visual and auditory processing
- *output*: manual and speech output

Each dimension is quantified through subjective rates (questions of table XI) 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 to 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 aggregation strategy is a simple sum of the 8 rates d:

$$WP : [0..8] \in \Re \qquad WP = \sum_{i=1}^{8} d_i$$

This aggregation method is intuitive but contrasts the $NASATLX$ because it does not consider the state of the operator, perceived performance and effort devoted. For comparison purposes, the WP value is divided by 8 and converted in the scale $[0..100] \in \Re$.

C. The System Usability Scale

As described in section II-B, the original SUS is a subjective usability assessment instrument that uses a Likert scale, bounded in the range 1 to 5 [4]. Individual scores are not meaningful on their own. For odd questions (SUS_i with $i = \{1|3|5|7|9\}$), the score contribution is the scale position (SUS_i) minus 1. For even questions (SUS_i with $i = \{2|4|6|8|10\}$), the contribution is 5 minus the scale position. For comparison purposes, the SUS value is converted in the range $[1..100] \in \Re$. Formally:

$$SUS : [0..1] \in \Re \quad i_1 = \{1,3,5,7,9\}\ i_2 = \{2,4,6,8,10\}$$

$$SUS = \frac{1}{10} \cdot [\sum_{i_1}(SUS_i) + \sum_{i_2}(100 - SUS_i)]$$

D. 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. Subjects in group A were different to the subjects in group B. Ages ranges from 20 to 35 years; there were 20 females and 20 males evenly distributed across the 2 groups (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 9 information-seeking web-based tasks (table VII in appendix) as naturally as they could, over 2 or 3 sessions of approximately 45/70 minutes each, on different non-consecutive days. Tasks differed in terms of difficulty, time-pressure, time-limits, interference, interruptions and demands on different modalities (as in table VII). Two groups were created because the tasks were executed on web-based interfaces, sometimes altered at run-time, through CSS and HTML manipulation, and sometimes not (as in table VIII). Manipulation was implemented to investigate how the perception of usability between the two groups interacts with subjective assessment of mental workload of users. Participants could not interact with instructors during the tasks. The order of the tasks administered over the sessions was the same for all the participants. In each experiment, a computerised version of the questions of the $NASATLX$ (table X), the WP (table XI) and the SUS (table IX) instruments was administered immediately after task completion. In addition a pair-wise comparison of the questions required by the $NASATLX$ instrument was performed[1].

V. Results

Tables I and II list the descriptive statistics of the mental workload and usability scores while figure 1 depicts the means of the scores of each task. From an initial analysis of figure 1, though a correlation might be spotted between the mental workload scores ($NASATLX$ vs. WP), there is no clear correlation between the mental workload scores ($NASATLX$, WP) and the usability scores (SUS). This is statistically confirmed in table III by the Pearson and the Spearman correlation coefficients obtained over the full dataset (360 cases). The two MWL assessment procedures are fairly correlated, and this was expected as they try to measure the same construct: mental workload. However, perception of usability, as assessed by the SUS technique, does not seem to have any correlation with mental workload assessments.

TABLE I. Mental workload and usability scores - Group A

	NASATLX		WP		SUS	
Task	avg	std	avg	std	avg	std
1	23.53	14.03	26.57	14.85	77	19.49
2	40.91	16.64	28.27	14.73	73.24	16.92
3	42.52	13.91	35.64	15.47	82.44	14.27
4	42.72	13.8	34.83	14.91	46.9	17.56
5	50.1	13.7	33.13	14.06	82.11	15.39
6	38.57	14.69	44.19	13.36	82.66	13.81
7	47.83	20.01	37.84	18.02	59.62	17.97
8	55.33	14.45	43.5	16.81	80.28	14.53
9	69.88	15.62	48.78	13.13	76.98	17.57

TABLE II. Mental workload and usability scores - Group B

	NASATLX		WP		SUS	
Task	avg	std	avg	std	avg	std
1	46.04	24.37	39.34	11.54	50.38	21.31
2	41.36	15.72	27.23	9.51	81.98	14.06
3	41.08	14.47	36.49	13.1	73.77	19.71
4	35.36	17.92	34.43	13.61	85.41	8.96
5	45.47	15.75	37.48	13.78	69.22	19.84
6	46.34	14.13	43.09	12.21	86.36	9.26
7	56.21	23.98	37.11	14.92	68.87	16.38
8	49.74	19.98	41.09	13.31	82.16	10.93
9	64.11	12.38	45.99	10.38	80.88	9.91

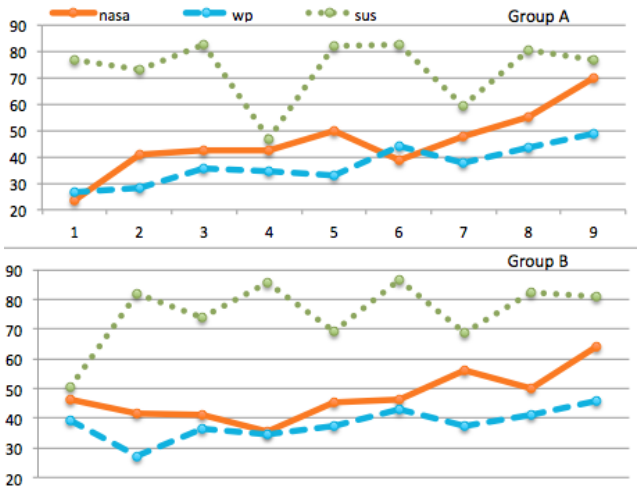

Fig. 1. Means of scores of $NASATLX$, WP, and SUS.

TABLE III. Correlation coefficients considering the full dataset between $NASATLX$, WP, SUS

	Pearson		Spearman	
	WP	SUS	WP	SUS
$NASATLX$	0.586	0.106	0.563	0.085
WP	1	0.026	1	0.027

Despite the fact that perception of usability does not seem to correlate at all with mental workload, a further investigation

[1]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 pairwise, 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 $[0..100] \in \Re$ [14].

of the relation between them has been carried out by performing an analysis of the scores obtained for each task. Tables IV and V list the correlations between the mental workload scores, obtained from the application of the $NASATLX$ and WP instruments against the usability scores obtained from the application of the SUS scale. Generally in the social and behavioural sciences, there may be a greater contribution from complicating factors, as in the case of subjective ratings, thus correlations above 0.5 are regarded as very high [6](page 82). Similarly, values within the range 0.1 and 0.3 are regarded as small correlations and values within the range 0.3 and 0.5 are seen as as medium/moderate correlations (ranges apply symmetrically to negative correlations). For the analysis, only medium and high correlation coefficients are taken into account and these are highlighted in tables IV and V. It is not possible to explain what really happened with the tasks by only examining these correlation coefficients. Figure 2 provides further details of those tasks in which mental workload ($NASATLX$ or WP) was correlated with perception of usability (SUS).

TABLE IV. CORRELATION BETWEEN MENTAL WORKLOAD AND USABILITY SCORES - GROUP A

	Pearson		Spearman	
Task	NASA vs SUS	WP vs SUS	NASA vs SUS	WP vs SUS
1	-0.21	-0.39	-0.24	-0.42
2	-0.22	0.18	-0.10	0.01
3	-0.25	-0.13	-0.23	-0.08
4	-0.05	-0.11	-0.10	-0.09
5	0.13	-0.27	0.10	-0.27
6	-0.17	-0.01	0.03	0.06
7	-0.11	0.03	-0.11	0.03
8	-0.28	0.02	-0.12	-0.13
9	0.48	-0.15	0.57	-0.15

TABLE V. CORRELATION BETWEEN MENTAL WORKLOAD AND USABILITY SCORES - GROUP B

	Pearson		Spearman	
Task	NASA vs SUS	WP vs SUS	NASA vs SUS	WP vs SUS
1	-0.69	-0.06	-0.6	-0.11
2	-0.12	-0.15	-0.15	-0.23
3	-0.07	0.13	-0.05	0.11
4	-0.64	-0.34	-0.60	-0.34
5	-0.34	-0.08	-0.31	-0.08
6	-0.08	-0.14	-0.07	-0.12
7	-0.32	-0.2	-0.36	-0.30
8	-0.08	-0.29	-0.04	-0.24
9	0.36	0.14	0.44	0.14

The following list provides possible interpretations on why mental workload scores were moderately/highly correlated with perception of usability:

- task 1/A and Task 4/B: WP is moderately negatively correlated with SUS. This suggests that when the proportion of attentional resources required by a task is moderated and it decreases, the perception of good usability of interfaces on which tasks are run, increases. In other words, when web-interfaces and the tasks which are carried out over them require a moderate use of different stages and codes of information processing as well as input and output modalities (as in tasks 1 and 4), the usability of those interfaces is increasingly perceived as positive.
- task 9/A and task 9/B: the $NASATLX$ is highly and positively correlated with SUS. This suggests that, even when time pressure is imposed upon tasks (increasing the mental workload experienced), and the answer that is to be provided by users is uncertain (decreasing performance), perception of usability is not affected if the task is pleasant and amusing (like task 9). This advices that even if experienced mental workload increases, and even if the interface is slightly altered (task 9 group B), the perception of good usability is strengthened if tasks are enjoyable.
- tasks 1/B, 4/B, 5/B, 7/B: the $NASATLX$ is highly negatively correlated with SUS. This suggests that when the mental workload experienced by users increases, and tasks are not straightforward, perception of usability can be seriously affected in a negative way with even a slight alteration of the interface.

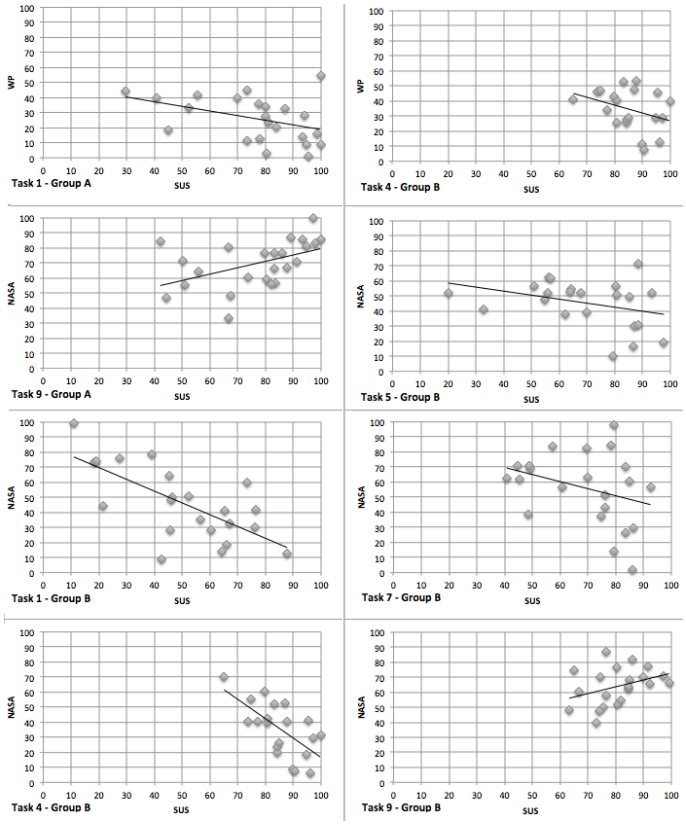

Fig. 2. Scatterplots of mental workload and perception of usability for tasks with moderate and high correlation

A. A/B testing

A further analysis is performed to verify the impact of the structural changes to a web-interface on the perception of mental workload and usability. From a statistical point of view, independent sample t-tests have been performed over the distributions of the mental workload scores and the usability scores of each task for group A against group B, with a confidence interval of 95%. The goal was to study whether there was a statistically significant difference between the scores produced by participants of group A and those of group B. This comparison is well known as A/B testing and it involves the comparison of the scores obtained for each

task, produced by volunteers interacting with the original web-interface and those obtained over their modified counterpart (as detailed in table VII - last 2 columns). The null hypothesis is:

$$H_0 : \mu_A^X = \mu_B^X$$

with X representing the $NASA$, WP and SUS instruments respectively. Informally, H_0: there is no difference between the distributions of the X scores obtained from subjects in group A and those obtained from subjects in group B.

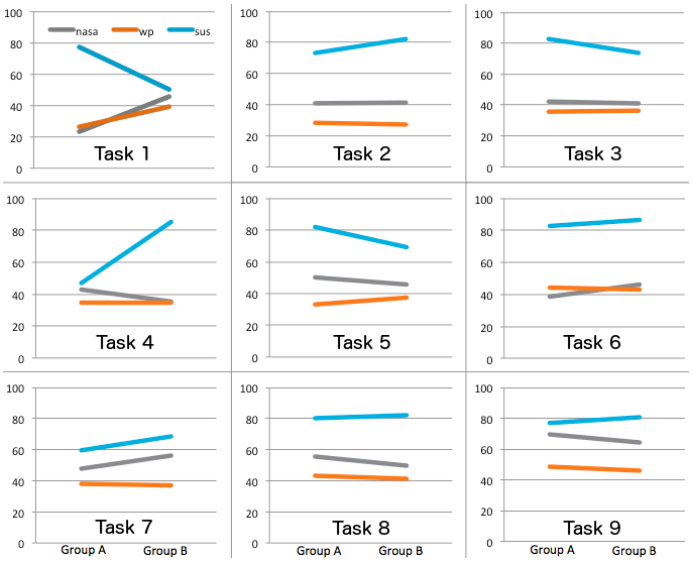

Fig. 3. Comparisons of the means of the mental workload ($NASATLX$, WP) and usability (SUS) scores of the 2 groups

TABLE VI. T-TESTS OF THE $NASATLX$, WP, SUS DISTRIBUTIONS OF SCORES OF GROUP A AGAINST GROUP B

	$NASATLX$			WP			SUS		
	t	p	H_0	t	p	H_0	t	p	H_0
1	-3.79	<0.001	×	-3.27	<0.001	×	4.41	<0.001	×
2	-0.09	0.93		0.28	0.78		-1.87	0.06	
3	0.34	0.74		-0.20	0.84		1.642	0.11	
4	1.50	0.14		0.09	0.93		-9.38	<0.001	×
5	1.06	0.29		-1.07	0.29		2.46	0.02	×
6	-1.79	0.08		0.29	0.78		-1.06	0.29	
7	-1.28	0.21		0.15	0.88		-1.83	0.07	
8	1.08	0.29		0.54	0.59		0.49	0.62	
9	1.38	0.17		0.79	0.43		-0.93	0.35	

The t-tests revealed a statistically significant difference between both the mental workload scores and the usability scores computed for task 1 of group A, and the scores for group B. This was the only task in which the modification of the interface (Wikipedia - task 1 - table VIII) caused, on average, both higher mental workload scores and perception of usability (task 1 of table 3). The removal of the searching box from the interface led users to work harder to find the right answer, imposing higher mental workload and affecting the perception of usability.

> *Intuitively, this suggests that even small structural changes can significantly alter the execution of a typical task, negatively affecting perception of usability and imposing a higher mental workload on end-users.*

Additionally, the T-tests revealed a statistically significant difference of the SUS scores for tasks 4 and 5, but they were not capable of revealing differences in the mental workload scores. The new black background of the google.com interface, the new font color (blue) and the removal of the left menu (only for task 4) affected the perception of usability but in practice did not impose a different mental workload on end-users.

> *This suggests that if the structural change does not modify the execution of a task, the interface that maximises perception of usability should be preferred.*

For the remaining tasks, no statistically significant difference in either mental workload or usability scores was detected. It turns out that interfaces A and B can be used interchangeably. In summary, the findings highlight the difficulty in spotting consistent relationships between the perception of usability of interfaces and the mental workload imposed by typical tasks executed over them. This suggests that usability and mental workload are two distinct, non-overlapping constructs, measuring two different phenomena. It turns out that incorporating mental workload in usability testing might provide designers with a better instrument for the design of interactive systems and interfaces.

VI. CONCLUSION

This study attempted to investigate the correlation between perception of usability and the mental workload imposed by typical tasks executed over three popular web-sites: Youtube, Wikipedia and Google. Prominent definitions of usability and mental workload have been provided, focusing more on the latter rather than the former. On one hand, usability is a central concept in human-computer interaction and a plethora of definitions and applications exists in the literature. On the other hand, the concept of mental workload has a background in Ergonomics and Human Factors with several assessment techniques being proposed. To the best of our knowledge, this research is the first of its kind to link these two notions and empirically investigate their interaction in the popular field of Human-Web Interaction. Specifically, a well known subjective instrument for assessing usability —the System Usability Scale —and two subjective mental workload assessment procedures —the NASA Task Load Index, and the Workload Profile —have been employed in a user study involving 40 subjects.

Empirical evidence suggests that there is no clear relationship between the perception of usability of a set of web-interfaces and the mental workload imposed by a set of designed tasks to be executed on them. It turns out that the two notions seem to model two non-overlapping phenomena and they could be jointly used to better describe the user experience over interacting interfaces, systems and technologies. Further studies will be devoted to making these findings more robust with additional user studies, a set of different interfaces, tasks, systems and with different usability assessment techniques and mental workload assessment procedures. The aims of this study are to offer a new perspective on the application of mental workload to traditional usability inspection methods, to better explain the interaction between humans and digital interfaces and to maximise users' experience.

REFERENCES

[1] J. Addie and T. Niels. Processing resources and attention. In *Handbook of human factors in Web design*, pages 3424–439. Lawrence Erlbaum Associates, 2005.

[2] M. Albers. Tapping as a Measure of Cognitive Load and Website Usability. *Proceedings of the 29th ACM international conference on Design of communication*, pages 25–32, 2011.

[3] D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. A Context-of-Use Taxonomy for Usability Studies. *International Journal of Human-Computer Interaction*, 26(10):941–970, 2010.

[4] J. Brooke. SUS: A quick and dirty usability scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland, editors, *Usability evaluation in industry*. Taylor and Francis, London, 1996.

[5] B. Cain. A review of the mental workload literature. Technical report, Defence Research & Dev.Canada, Human System Integration, 2007.

[6] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates., 1988.

[7] G. E. Cooper and R. P. Harper. The use of pilot ratings in the evaluation of aircraft handling qualities. Technical Report AD689722, 567, Advisory Group for Aerospace Research & Development.

[8] D. De Waard. *The measurement of drivers' mental workload*. The Traffic Research Centre VSC, University of Groningen.

[9] G. Fischer. User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, 11(1-2):65–86, Mar. 2001.

[10] D. Gopher and E. Donchin. Workload - an examination of the concept. In K. R. Boff, L. Kaufman, and J. P. Thomas, editors, *Handbook of perception and human performance*, volume 2, pages 41/1–41/49. John Wiley & Sons, 1986.

[11] J. Gwizdka. Assessing cognitive load on web search tasks. *The ergonomic open journal*, 2(1):114–123, 2009.

[12] J. Gwizdka. Distribution of cognitive load in web search. *Journal of the american society & information science & technology*, 61(11):2167–2187, November 2010.

[13] B. D. Harper and K. L. Norman. Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5. In *1st Annual Mid-Atlantic Human Factors Conference*, pages 224–228, 1993.

[14] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50, 2006.

[15] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): results of empirical and theoretical research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. 1988.

[16] K. Hornbaek. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2):79 – 102, 2006.

[17] B. M. Huey and C. D. Wickens. *Workload transition: implication for individual and team performance*. National Academy Press.

[18] J. R. Lewis. Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, 1995.

[19] L. Longo. *Formalising Human Mental Workload as a Defeasible Computational Concept*. PhD thesis, Trinity College Dublin, 2014.

[20] L. Longo. A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & IT*, 34(8):758–786, 2015.

[21] L. Longo and S. Barrett. A computational analysis of cognitive effort. In *Intelligent Information and Database Systems, Second International Conference, ACIIDS, Hue City, Vietnam*, volume LNCS 5991, pages 65–74. Springer, 2010.

[22] L. Longo, S. Barrett, and P. Dondio. Information foraging theory as a form of collective intelligence for social search. In *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems*.

[23] L. Longo and B. Kane. A novel methodology for evaluating user interfaces in health care. In *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, pages 1–6, June 2011.

[24] L. Longo, F. Rusconi, L. Noce, and S. Barrett. The importance of human mental workload in web-design. In *8th Int. Conference on Web Information Systems and Technologies*, pages 403–409, 2012.

[25] M. Macleod. Usability in context: Improving quality of use. In *Human Factors in Organizational Design and Management*. Elsevier, 1994.

[26] J. Nielsen. Heuristic evaluation. In J. Nielsen and R. L. E. Mack, editors, *Usability Inspection Methods*. Wiley & Sons, New York, 1994.

[27] J. Nielsen. Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems*, CHI '95, pages 377–378, New York, NY, USA, 1995. ACM.

[28] R. D. O' Donnel and T. F. Eggemeier. Workload assessment methodology. In K. Boff, L. Kaufman, and J. Thomas, editors, *Handbook of perception and human performance*, volume 2, pages 42:1–42:49. New York, Wiley-Interscience, 1986.

[29] G. B. Reid and T. E. Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, chapter 8, pages 185–218. 1988.

[30] A. H. Roscoe and G. A. Ellis. A subjective rating scale for assessing pilot workload in flight: a decade of practical use. Technical report 90019, Royal Aerospace Establishment, Farnborough (UK), 1990.

[31] S. Rubio, E. Diaz, J. Martin, and J. M. Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86, 2004.

[32] B. Shackel. Usability - context, framework, definition, design and evaluation. *Interact. Comput.*, 21(5-6):339–346, Dec. 2009.

[33] L. A. Slaughter, B. D. Harper, and K. L. Norman. Assessing the equivalence of paper and on-line versions of the quis 5.5. In *nd Annual Mid-Atlantic Human Factors Conference*, pages 87–91, 1994.

[34] J. P. Tracy and M. J. Albers. Measuring Cognitive Load to Test the Usability of Web Sites. *Usability and Information Design*, pages 256–260, 2006.

[35] P. S. Tsang. Mental workload. In W. Karwowski, editor, *International Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 166. Taylor & Francis, 2006.

[36] P. S. Tsang and V. L. Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.

[37] P. S. Tsang and M. A. Vidulich. Mental workload and situation awareness. In G. Salvendy, editor, *Handbook of Human Factors and Ergonomics*, pages 243–268. John Wiley & Sons, Inc., 2006.

[38] T. S. Tullis and J. N. Stetson. A Comparison of Questionnaires for Assessing Website Usability. In *Annual Meeting of the Usability Professionals Association*, 2004.

[39] M. A. Vidulich and S. J. Ward Frederic G. Using the subjective workload dominance (sword) technique for projective workload assessment. *Human Factors Society*, 33(6):677–691, December 1991.

[40] C. D. Wickens. Multiple resources and mental workload. *Human Factors*, 50(2):449–454, 2008.

[41] C. D. Wickens and J. G. Hollands. *Engineering Psychology and Human Performance*. Prentice Hall, 3rd edition, Sept. 1999.

[42] G. F. Wilson and T. F. Eggemeier. Mental workload measurement. In W. Karwowski, editor, *Int. Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, volume 1, chapter 167. Taylor & Francis, 2006.

[43] B. Xie and G. Salvendy. Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. *Work and Stress*, 14(1):74–99, 2000.

[44] M. S. Young and N. A. Stanton. Mental workload. In N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, and H. W. Hendrick, editors, *Handbook of Human Factors and Ergonomics Methods*, chapter 39, pages 1–9. CRC Press, 2004.

[45] M. S. Young and N. A. Stanton. Mental workload: theory, measurement, and application. In W. Karwowski, editor, *International encyclopedia of ergonomics and human factors*, volume 1, pages 818–821. Taylor & Francis, 2nd edition, 2006.

[46] H. Zhu and M. Hou. Restrain mental workload with roles in hci. In *Proceedings of Science & Technology for Humanity*, pages 387–392, 2009.

[47] F. R. H. Zijlstra. Efficiency in work behaviour. Doctoral thesis, Delft University, The Netherlands, 1993.

TABLE VII. LIST OF EXPERIMENTAL TASKS

Task	Description	Typology	Task condition	Web-site	Group A	Group B
T_1	Find out how many people live in Sidney	Fact finding	Simple search	Wikipedia		*altered*
T_2	Read the content of simple.wikipedia.org/wiki/Grammar	Browsing	Not goal-oriented and no time pressure	Wikipedia	*altered*	
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	Fact finding	dual-task and mental arithmetical calculations	Google		*altered*
T_4	Find out the difference (in years) between the foundation of the Microsoft Corp. & the year of the 23^{rd} Olympic games	Fact finding	dual-task and mental arithmetical calculations	Google	*altered*	
T_5	Find out the year of birth of the 1^{st} wife of the founder of playboy	Fact finding	Single task by timbre pressure (2-min limit). Each 30 secs user is warned of time left	Google		*altered*
T_6	Find out the name of the man (interpreted by Johnny Deep) in the video www.youtube.com/watch?v=FfTPS-TFQ_c	Fact finding	Constant demand on visual and auditory modalities. Participant can replay the video if required	Youtube		*altered*
T_7	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	Fact finding	Demand on visual modality and inference on auditory modality. The song is extremely irritating	Wikipedia	*altered*	
T_8	Find out how many times Stewie jumps in the video www.youtube.com/watch?v=TSe9gbdkQ8s	Fact finding	Demand on visual resource and external inference: participant is distracted twice & can replay video	Youtube	*altered*	
T_9	Find out the age of the blue fish in the video www.youtube.com/watch?v=H4BNbHBcnDI	Fact finding	Demand on visual and auditory modality, plus time-pressure:150-sec limit. User can replay the video. There is no answer.	Youtube		*altered*

TABLE VIII. RUN-TIME MANIPULATION OF WEB-INTERFACES

Task	Manipulation
1	Left menu of wikipedia.com and the internal searching box have been removed. The background colour has been set to light yellow.
2	Left menu of wikipedia.com and the internal searching box have been removed. The background colour has been set to light yellow. (task 1)
3	Each result returned by Google has been wrapped with a box with thin borders and the font has been altered.
4	The left menu of google.com has been removed, the background colour set to black and the font colour to blue.
5	The background colour of google.com has been set to black and the font colour to blue.
6	The background colour of youtube.com has been set to dark grey.
7	The background colour of wikipedia.com has been set to light blue and headings to white.
8	The background colour of youtube.com has been set to black and each video was always displayed in 16:9, removing the right list of related videos.
9	The background colour of youtube.com has been set to dark grey. (task 6)

TABLE IX. SYSTEM USABILITY SCALE (SUS) QUESTIONNAIRE

Label	Question
SUS_1	I think that I would like to use this interface frequently
SUS_2	I found the interface unnecessarily complex
SUS_3	I thought the interface was easy to use
SUS_4	I think that I would need the support of a technical person to use this interface
SUS_5	I found the various functions in this interface were well integrated
SUS_6	I thought there was too much inconsistency in this interface
SUS_7	I would imagine that most people would learn to use this interface quickly
SUS_8	I found the interface very unmanageable (irritating or tiresome) to use
SUS_9	I felt very confident using the interface
SUS_{10}	I needed to learn a lot of things before I could get going with this interface

TABLE X. NASA TASK LOAD (NASATLX) QUESTIONNAIRE

Label	Question
NT_1	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
NT_2	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
NT_3	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
NT_4	How hard did you have to work (mentally and physically) to accomplish your level of performance?
NT_5	How successful do you think you were in accomplishing the goals, of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
NT_6	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

TABLE XI. WORKLOAD PROFILE (WP) QUESTIONNAIRE

Label	Question
WP_1	How much attention was required for activities like remembering, problem-solving, decision-making, perceiving (detecting, recognising, identifying objects)?
WP_2	How much attention was required for selecting the proper response channel (manual - keyboard/mouse, or speech - voice) and its execution?
WP_3	How much attention was required for spatial processing (spatially pay attention around you)?
WP_4	How much attention was required for verbal material (eg. reading, processing linguistic material, listening to verbal conversations)?
WP_5	How much attention was required for executing the task based on the information visually received (eyes)?
WP_6	How much attention was required for executing the task based on the information auditorily received (ears)?
WP_7	How much attention was required for manually respond to the task (eg. keyboard/mouse)?
WP_8	How much attention was required for producing the speech response (eg. engaging in a conversation, talking, answering questions)?