Subjective Usability, Mental Workload Assessments and Their Impact on Objective Human Performance

Luca Longo

Subjective usability, mental workload assessments and their impact on objective human performance

Luca Longo*

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

Abstract. Self-reporting procedures and inspection methods have been largely employed in the fields of interaction and web-design for assessing the usability of interfaces. However, there seems to be a propensity to ignore features related to end-users or the context of application during the usability assessment procedure. This research proposes the adoption of the construct of mental workload as an additional aid to inform interaction and web-design. A user-study has been performed in the context of human-web interaction. The main objective was to explore the relationship between the perception of usability of the interfaces of three popular web-sites and the mental workload imposed on end-users by a set of typical tasks executed over them. Usability scores computed employing the System Usability Scale were compared and related to the mental workload scores obtained employing the NASA Task Load Index and the Workload Profile self-reporting assessment procedures. Findings advise that perception of usability and subjective assessment of mental workload are two independent, not fully overlapping constructs. They measure two different aspects of the human-system interaction. This distinction enabled the demonstration of how these two constructs cab be jointly employed to better explain objective performance of end-users, a dimension of user experience, and informing interaction and web-design.

1 Introduction

In recent decades the demands of evaluating usability of interactive web-based systems have produced several assessment procedures. Very often, during usability inspection, there is a tendency to overlook features of the users, aspects of the context and characteristics of the tasks. This tendency is also justified by the lack of a model that unifies all of these aspects. Considering features of users is fundamental for the User Modeling community [1,16]. Similarly, taking into consideration the context of use is of extreme importance for inferring reliable assessments of usability [3,36]. Additionally, during the usability assessment process, accounting for the demands of the task executed is core for describing user experience [20]. Building a cohesive model is not trivial, however we believe the construct of human *mental workload* (MWL) – often referred to as cognitive load – can significantly contribute to such a goal and inform interaction and

web-design. MWL, with roots in Psychology, has been mainly applied within the fields of Ergonomics and Human Factors. Its assessment is key to measuring performance, which in turn is fundamental for describing user experience and engagement. A few studies have tried to employ the construct of MWL to explain usability [2,24,41,46,50]. Despite this interest, not much has yet been done to investigate their relationship empirically. The aim of this research is to empirically test the relationship between subjective perception of usability and mental workload as well as their impact on objective user performance, which means tangible quantifiable facts.

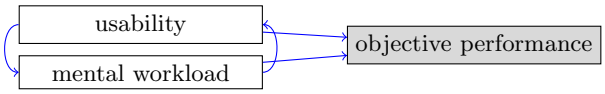

Fig. 1: Schematic overview of the empirical study

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 Human-Computer Interaction (HCI). Related work is also presented, highlighting how the two constructs have been employed so far, distinctly and jointly. 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 (youtube, wikipedia and google) and the mental workload experienced by users after interacting with them. Results are presented and critically discussed, showing how these constructs interact and how they impact objective user performance. A summary concludes this paper pointing to future work and highlighting the contribution to knowledge.

2 Core notions and definitions

Widely employed in the broader field of HCI, usability and mental workload are two constructs from Ergonomics, with no crystal and generally applicable definitions. There is an acute debate on their assessment and measurement [4,5,6]. Although ill-defined, they remain extremely important for describing the user experience and improving interaction, interface and system design.

2.1 Definitions of usability

The amount of literature covering definitions [21,48], frameworks and methodologies for assessing usability is vast. 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 [38], is a method for improving ease-of-use in the design of interactive systems and technologies. It embraces other concepts such as efficiency, learnability and satisfaction. It is often associated with the functionalities of a product rather than being merely a feature of the user interface [39].

2.2 Measures of usability

Often when selecting an appropriate procedure in the context of interaction and web-design, it is desirable to consider the effort and expense that will be incurred in collecting and analysing data. For this reason, designers have tended to adopt subjective usability assessment techniques for collecting feedback from users [21]. On one hand, 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. On the other hand, these techniques appear to be the most sensitive and diagnostical [21]. Nielsen's principles, thanks to their simplicity in terms of effort and time, are frequently employed to evaluate the usability of interfaces [38]. The evaluation is done iteratively by systematically finding usability problems in an interface and judging them according to the principles [39]. The main problem associated to these principles is that they mainly focus on the user interface forgetting contextual factors, the cognitive state of the users and the underlying tasks. The System Usability Scale [9] is a questionnaire that consists of ten questions (table 9). It is a highly cited usability assessment method and it has been massively applied [7]. It is a very easy scale to administer, demonstrating reliability to distinguishing usable and unusable systems and even with small sample sizes [54]. Alternatives include the Computer System Usability Questionnaire (CSUQ), developed at IBM and the Questionnaire for User Interface Satisfaction (QUIS), developed at the HCI lab at the University of Maryland. The former is a survey that consists of 19 questions on a seven-point Likert scale of 'strongly disagree' to 'strongly agree' [25]. The latter was designed to assess users' satisfaction with aspects of a computer interface [49]. It includes: a demographic questionnaire, a measure of system satisfaction along six scales, and a hierarchy of measures of nine specific interface factors. Each of these factors 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, QUIS has shown high reliability across several interfaces [19]. Many other usability inspection methods and techniques have been proposed in the literature [21,54].

2.3 Definitions of mental workload

Human Mental Workload (MWL) is an important design concept and it is fundamental for exploring the interaction of people with technological devices [29,31,32]. It has a long history in Psychology with applications in Ergonomics, especially in the transportation industry [14,20]. The principal reason for MWL assessment is to quantify the cognitive cost associated to performing a task for predicting operator or system performance [10]. However, it has been largely reported that mental underload and overload can negatively influence performance [60]. On one hand, during information processing, when MWL is at a low level, individuals may frequently feel frustrated or annoyed. On the other hand, when MWL is at a high level, this can lead individual to confusion and

decrease their performance in processing information and increases the chances of mistakes. Hence, designers who are interested in human or system performance require answers about operator workload at all stages of system design and operation so design alternatives can be explored and evaluated [20]. MWL is not a linear concept [30,43] but it can be intuitively defined as the volume of cognitive work necessary for an individual to accomplish a task over time. 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' [20]. However, this is only a practical definition, as many other factors influence mental workload [33].

2.4 Measures of mental workload

The measurement of MWL is an extensive area where several assessment techniques have been proposed [10,51,59,61,62,37]: a) *self-assessment measures*; b) *task measures*; c) *physiological measures*; The category of *self-assessment measures* is often referred to as self-report measures. It relies on the subject perceived experience of the interaction with an underlying interactive system through the direct estimation of individual differences such as the emotional state, attitude and stress of the operator, the effort devoted to the task and its demands [14,20]. It is strongly believed that only the individual concerned with the task can provide an accurate judgement with respect to the MWL experienced, hence self-assessment measures have always attracted many practitioners. This has also been adopted in this study. The class of *performance measures* is based upon the assumption that the mental workload of an operator, interacting with a system, gain relevance only if it influences system performance. In turn, this class appears as the most valuable options for designers [45,53]. The category of *physiological measures* considers bodily responses derived from the operator's physiology. These responses are believed to be correlated to MWL and are aimed at interpreting psychological processes by analysing their effect on the state of the body. Their advantage is that they can be collected continuously over time, without requiring an overt response by the operator [40] but they require specific equipment and trained operators mitigating their employability in real-world tasks.

3 Related work

In a recent review, it was acknowledge that usability and performance are two core elements for assessing user experience [46]. Lehmann et al. also emphasise the importance of adopting multiple metrics for tackling the problem of user engagement measurement, being usability and cognitive engagement part of these metrics [24]. OBrien and collaborators identified mental workload and usability as elements of user engagement, suggesting that a little correlation exists between the two constructs [41]. Nonetheless, this is under-investigated in their

environment and, to the best of our knowledge, this study is the first real attempt aimed at exploring the relationship between subjective mental workload and perception of usability. Additionally, because the former area is less explored in interaction and web design, while the latter area has an extensive research endeavour [21,54], this section mainly covers related work on mental workload.

3.1 Applications of MWL for design

At an early design phase, a system/interface can be optimised taking mental workload into consideration, guiding designers in making appropriate structural changes [60]. Specifically, in the context of web-applications, modern interfaces have become increasingly complex [35], often requiring more mentally demanding tasks with a consequent increments in the degree of mental workload imposed on operators [17,18]. As the difficulty of these task increases, due to interface complexity, mental workload also increases and performance usually decreases [10]. In turn, operator's response time increases, error are more recurrent and fewer tasks are completed per time unit [22]. In contrast, when task difficulty is minor, interfaces and systems can impose a low mental workload on operators. This situation should be avoided as it leads to difficulties in maintaining attention and increasing reaction time [10]. [63] noted how roles can be useful in interface design and proposed a role-based method to measure MWL applicable in HCI for dynamically adjusting mental workload and enhance performance in interaction.

3.2 Application of MWL self-assessment measures

Self-assessment measures of MWL include multidimensional approaches such as the NASA's Task Load Index ($NASATLX$) [20], the Subjective Workload Assessment Technique [42], the Workload Profile [52] as well as uni-dimensional measures such as the Copper-Harper scale [13], the Rating Scale Mental Effort [64], the Subjective Workload Dominance Technique [55] and the Bedford scale [44]. These procedures have low implementation requirements, low intrusiveness and high subject acceptability. The $NASATLX$ has been used for evaluating user interfaces in health-care [26] or in e-commerce, along with a dual-task objective methodology for investigating the effects on user satisfaction [47]. The Workload Profile [52], the $NASATLX$ and the Subjective Workload Assessment Technique [42] have been compared in a user study to evaluate different web-based interfaces [35]. Tracy and Albers adopted three different techniques for measuring MWL in web-site design: $NASATLX$, the Sternberg Memory Test and a tapping test [2,50]. They proposed a technique to identify sub-areas of a web-site in which end-users manifested a higher mental worklaod during interaction, allowing designers to modify those critical regions. Similarly, [15] investigated how the design of query interfaces influence stress, workload and performance during information search. Here stress was measured by physiological signals and a subjective assessment technique – Short Stress State Questionnaire. Mental workload was assessed using the $NASATLX$ and log data was used as objective indicator of performance to characterise search behaviour.

4 Design of experiments

A study involving human participants executing typical tasks over 3 popular web-sites (youtube, google, wikipedia) was set to investigate the relationship between perception of usability, mental workload and objective performance. One self-assessment procedure for measuring usability and two for mental workload:

- the System Usability Scale (SUS) [9]
- the Nasa Task Load Index ($NASATLX$), developed at NASA [20]
- the Workload Profile (WP) [52], based on Multiple Resource Theory [57,56].

Five classes of the objective performance of participants on tasks were set:

1. the task was not completed as the user gave up
2. the execution of the task was terminated because the available time was over
3. the task was completed and no answer was required by the user
4. the task was completed, the user provided an answer, but it was wrong
5. the task was completed and the user provided the correct answer

These are sometimes conditionally dependent (figure 2). The experimental hypotheses are defined in table 1 and illustrated in figure 3.

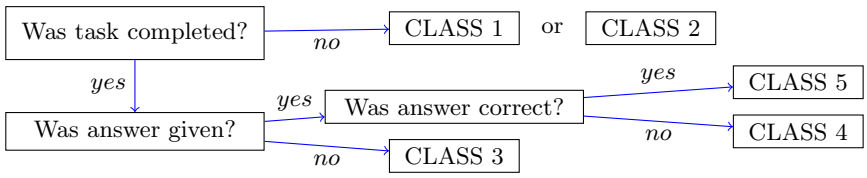

Fig. 2: Partial dependencies of classes of objective performance

Table 1: Research hypotheses

H_1	Usability and Mental workload are two uncorrelated constructs capturing difference variance (as measured with self-reporting techniques - SUS, NASATLX, WP).
H_2	A unified model incorporating a usability and a MWL measure can better predict objective performance than MWL alone.

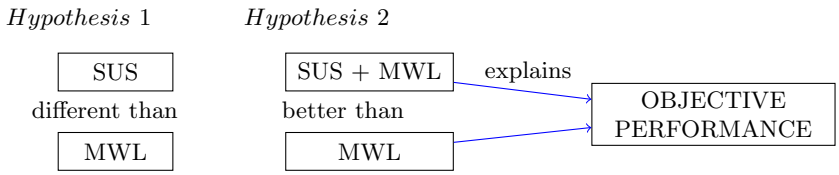

Fig. 3: Illustration of research hypotheses

4.1 Details of experimental subjective self-reporting techniques

The System Usability Scale is a subjective usability assessment instrument that uses a Likert scale, bounded in the range 1 to 5 [9]. Questions can be found in table 9. 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 $[1..100] \in \Re$ with $i_1 = \{1, 3, 5, 7, 9\}$ $i_2 = \{2, 4, 6, 8, 10\}$

$$SUS = 2.5 \cdot \left[\sum_{i_1} (SUS_i - 1) + \sum_{i_2} (5 - SUS_i) \right]$$

The NASA Task Load Index instrument [20] belongs to the category of self-assessment measures. It has been validated in the aviation industry and other contexts in Ergonomics [20,45] with several applications in many socio-technical domains. It is a combination of six factors believed to influence MWL (questions of table 10). Each factors is quantified with a subjective judgement coupled with a weight computed via a paired comparison procedure. Subjects are required to decide, for each possible pair (binomial coefficient, $\binom{6}{2} = 15$) of the 6 factors, *'which of the two contributed the most to mental workload during the task'*, such as 'Mental or Temporal Demand?', and so forth. The weights w are the number of times each dimension was selected. In this case, the range is from 0 (not relevant) to 5 (more important than any other attribute). The final MWL score is computed as a weighed average, considering the subjective rating of each attribute d_i and the correspondent weights w_i:

$$NASATLX : [0..100] \in \Re \qquad NASATLX = \left(\sum_{i=1}^{6} d_i \times w_i \right) \frac{1}{15}$$

The Workload Profile (WP) assessment procedure [52] is built upon the Multiple Resource Theory proposed in [56,57]. In this theory, individuals are seen as having different capacities or 'resources' related to: • *stage of information processing* – perceptual/central processing and response selection/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 11) 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 range $0..1 \in \Re$. A rating of 0 means that the task placed no demand while 1 indicates that it required maximum attention. The aggregation strategy is a simple sum of the 8 rates d (averaged here, and scaled in $[1..100] \in \Re$ for comparison purposes):

$$WP : [0..100] \in \Re \qquad WP = \frac{1}{8} \sum_{i=1}^{8} d_i \times 100$$

4.2 Participants and procedure

A sample of 46 people fluent in english volunteered to participate in the study after signing a consent form. Subjects were divided into 2 groups of 23 each: those in group A were different to those in group B. Participants could not interact with instructors during the tasks and they did not have to be trained. Ages ranges from 20 to 35 years; 24 females and 22 males evenly distributed across the 2 groups (Total - Avg.: 28.6, Std. 3.98; g.A - Avg. 28.35, Std.: 4.22; g.B - Avg: 28.85, Std.: 3.70) all with a daily Internet usage of at least 2 hours. Participants were required to execute a set of 9 information-seeking web-based tasks (table 13) 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 psychological modalities. Two groups were created because the tasks were executed on web-based interfaces, sometimes altered at run-time (through a CSS/HTML manipulation) (as in table 12). This manipulation was implemented, as part of a larger study [27,28,34], to enable A/B testing of web-interfaces (not included here). Interface alteration was not extreme, like making things very hard to read. Rather the goal was to alter the original interface to manipulate task difficulty and usability independently. The order of the tasks administered was the same for all the participants. Computerised versions of the SUS (table 9), the $NASATLX$ (table 10) and the WP (table 11) instruments were administered immediately after task completion. Note that the question of the $NASA-TLX$ related to 'physical load' was set to 0 as well as its weight. Consequently, the pairwise comparison procedure was shorter. Some volunteer did not execute all the tasks and the final dataset contains 405 cases.

5 Results

Table 2 contains the means and standard deviations of the usability and the mental workload scores for each task, depicted also in figure 4.

Table 2: Mental workload & usability - Groups A, B (G.A/G.B)

G. A	NASATLX		WP		SUS		G. B	NASATLX		WP		SUS	
Task	avg	std	avg	std	avg	std	Task	avg	std	avg	std	avg	std
1	46.03	24.30	39.34	11.54	50.38	21.31	1	23.66	13.93	26.57	14.85	77.00	19.49
2	41.38	15.71	27.23	9.51	81.98	14.06	2	40.97	16.62	28.27	14.73	73.24	16.92
3	41.08	14.47	36.50	13.10	73.77	19.71	3	42.63	14.21	35.60	15.81	82.33	14.58
4	35.36	17.92	34.43	13.61	85.41	8.96	4	42.70	14.09	34.87	15.25	46.61	17.90
5	45.47	15.74	37.49	13.78	69.22	19.84	5	51.15	13.78	33.54	13.88	84.64	12.77
6	46.35	14.13	43.09	12.20	86.36	09.26	6	39.31	14.57	44.61	13.50	82.68	14.12
7	56.20	23.97	37.11	14.92	68.87	16.38	7	47.86	19.97	37.84	18.02	59.62	17.97
8	49.76	19.96	41.09	13.31	82.16	10.93	8	55.34	14.75	42.97	16.98	81.41	13.73
9	64.61	12.92	46.65	10.46	81.85	09.81	9	70.75	16.29	50.51	14.06	75.39	18.02

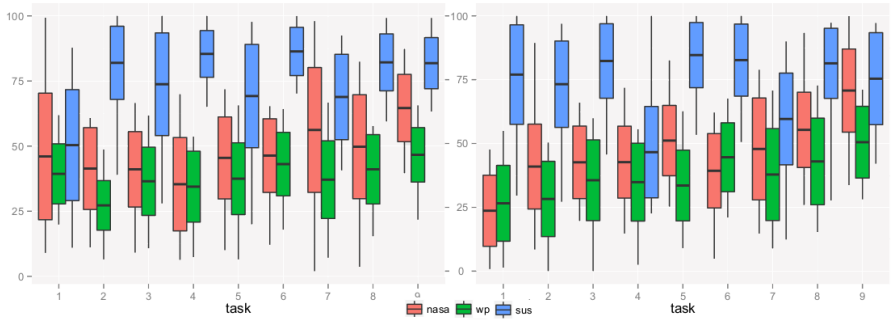

Fig. 4: Summary statistics by task

5.1 Testing hypothesis 1 - Difference usability and mental workload

From an initial analysis of figure 5, it seems clear that there is no correlation between the usability scores (SUS) and the mental workload scores ($NASATLX$, WP). This is statistically confirmed in table 3 by the Pearson and Spearman correlation coefficients computed over the full dataset (Groups A, B). Person was chosen for exploring linear correlation while Spearman for monotonic relationship, not necessarily linear.

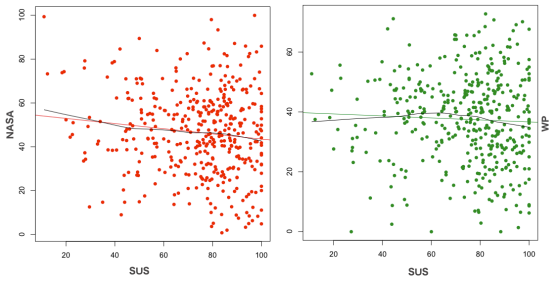

Table 3: Correlation coefficients

pearson	WP	SUS
$NASA$	0.55	-0.13
WP		-0.05

spearman	WP	SUS
$NASA$	0.53	-0.1
WP		-0.08

Fig. 5: Scatterplots of $NASATLX$, WP vs SUS.

Despite perception of usability does not seem to correlate at all with mental workload, a further investigation of their relationship was performed on the scores obtained for each task. Tables 4 lists the correlations between the MWL scores ($NASATLX$, WP) against the usability scores (SUS), and figure 6 their densities. Generally, in behavioural/social sciences, there may be a greater contribution from complicating factors, as in the case of subjective ratings. Hence, correlations above 0.5 are regarded as very high, within $[0.1 - 0.3]$ small and within $[0.3 - 0.5]$ as medium/moderate (symmetrically to negative values) [12](page 82). For this analysis, only medium/high coefficients are considered. Yet, a clearer picture does not emerge and just a few tasks show some form of correlation between mental workload and usability. Figure 7 provides further details aiming at extracting further information and possible interpretations on why workload scores were moderately/highly correlated with usability.

Table 4: Correlations MWL vs usability. Groups A and B

G. B	Pearson		Spearman		G. A	Pearson		Spearman	
Task	Nasa/SUS	WP/SUS	Nasa/SUS	WP/SUS	Task	Nasa/SUS	WP/SUS	Nasa/SUS	WP/SUS
1	-0.21	-0.39	-0.24	-0.42	1	-0.69	-0.06	-0.6	-0.11
2	-0.22	0.18	-0.1	0.01	2	-0.12	-0.15	-0.15	-0.23
3	-0.25	-0.13	-0.23	-0.08	3	-0.07	0.13	-0.05	0.11
4	-0.05	-0.11	-0.10	-0.09	4	-0.64	-0.34	-0.60	-0.34
5	0.14	-0.26	0.10	-0.27	5	-0.34	-0.08	-0.31	-0.08
6	-0.17	-0.01	0.04	0.06	6	-0.08	-0.14	-0.07	-0.12
7	-0.11	0.03	-0.10	0.03	7	-0.32	-0.2	-0.37	-0.30
8	-0.28	0.02	-0.13	-0.13	8	-0.08	-0.29	-0.04	-0.24
9	0.48	-0.15	0.57	-0.15	9	0.36	0.14	0.44	0.14

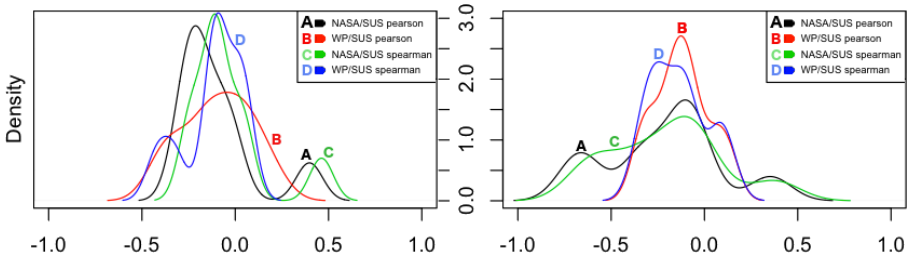

Fig. 6: Density plots of the correlations by task - Group A, B

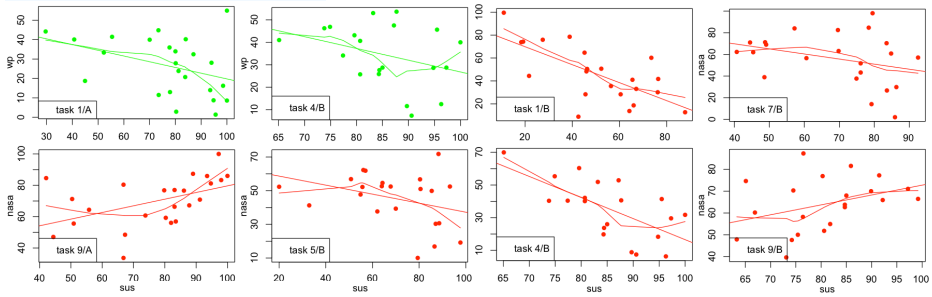

Fig. 7: Details of tasks with moderate/high correlation

- task 1/A and task 4/B: WP is moderately negatively correlated with SUS. This suggests that *when the proportion of attentional resources being taxed by a task is moderated and decreases, the perception of good usability increases.* In other words, when web-interfaces and the tasks executed over them require a moderate use of different stages, codes of information processing and input, output modalities (section 4.1), 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 causing an increment in the workload experienced, and the perception of performance decreases because task answer is not found, than perception of usability is not affected if the task is pleasant and amusing (like task 9). In other words, *even if experienced workload increases but is not excessive,*

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 MWL experienced by users increases, perhaps because tasks are not straightforward, perception of usability can be negatively affected even with a slight alteration of the interface.*

The above interpretations do not aim to be exhaustive; they are just our own interpretations, they cannot be generalised and are only confined to this study. To further strengthening the data analysis, an investigation of the correlation between the MWL and the usability scores has been performed by considering users on an individual-basis (table 5 and figure 8).

Table 5: Correlation MWL-usability by user

	Pearson		Spearman			Pearson		Spearman	
User	Nasa/SUS	WP/SUS	Nasa/SUS	WP/SUS	User	Nasa/SUS	WP/SUS	Nasa/SUS	WP/SUS
1	-0.5	-0.43	-0.45	-0.32	24	0.19	0.32	-0.25	0.19
2	0.41	-0.11	0.57	-0.23	25	-0.62	-0.07	-0.38	-0.4
3	-0.4	0.18	-0.27	0.45	26	-0.69	0.29	-0.62	0.38
4	0.38	0.37	0.15	0.17	27	-0.38	-0.36	-0.55	-0.58
5	-0.66	-0.57	-0.7	-0.63	28	-0.13	-0.43	-0.2	-0.48
6	-0.15	-0.34	-0.06	-0.14	29	-0.11	0.28	-0.03	0.15
7	-0.17	-0.2	-0.17	-0.4	30	0.17	-0.22	0.22	-0.38
8	0.02	0.21	-0.36	0.01	31	-0.6	-0.42	-0.78	-0.48
9	-0.16	-0.4	-0.25	-0.08	32	-0.7	-0.4	-0.2	-0.22
10	0	0.26	-0.05	0.33	33	0.06	-0.67	0	-0.32
11	-0.47	-0.74	-0.52	-0.78	34	-0.41	-0.45	-0.32	-0.27
12	0.58	-0.33	0.46	-0.4	35	0.58	0.12	0.8	0.4
13	-0.17	0.18	-0.23	0.18	36	-0.39	-0.31	-0.54	-0.37
14	0.24	0.39	-0.22	0.16	37	-0.47	-0.08	-0.17	0.38
15	0.06	0.17	0.21	0.47	38	0.21	0.43	0.32	0.51
16	0.46	0.34	0.57	0.55	39	-0.17	-0.07	0.2	0.12
17	0.27	0.02	0.15	0.23	40	-0.34	0.93	0.1	0.87
18	-0.14	0.16	-0.15	-0.2	41	0.25	-0.23	0.37	-0.35
19	-0.57	0.13	-0.41	0.1	42	-0.67	-0.6	-0.65	-0.38
20	0.05	-0.21	0.27	0.18	43	0.36	0.34	0.28	0.25
21	0.36	-0.05	-0.07	0.07	44	-0.69	-0.69	-0.67	-0.51
22	-0.99	0.05	-1	0.4	45	-0.51	-0.22	-0.42	-0.27
23	0.29	-0.05	0.45	-0.17	46	0.39	0.59	0.2	0.36

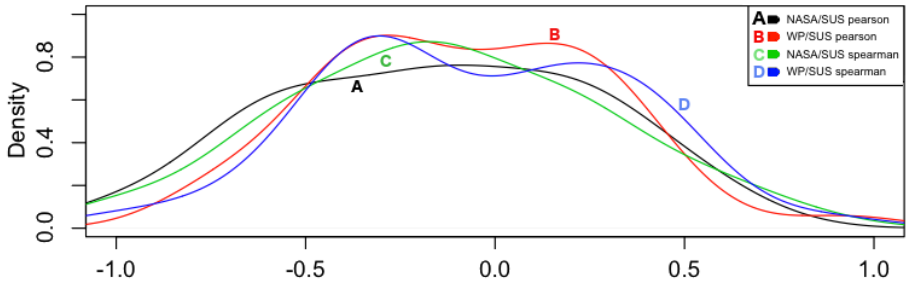

Fig. 8: Density plots of the correlations by user

As in the previous analysis (by task), just medium and high correlation coefficients (> 0.3) are considered for deeper investigation. Additionally, because the results of table 3 and tables 4 were not able to systematically show common trends, the analysis on the individual-basis was reinforced by considering only those users for which a medium/high linear relationship (Pearson) and a monotonic relationship (Spearman) was detected between both the two MWL scores ($NASA$, WP) and the usability scores (SUS). Table 5 highlights these users (1, 5, 11, 12, 21, 22, 27, 39, 40, 46). The objective was to look for the presence of any particular pattern of user's behaviour or a complex deterministic structure. Figure 9 depicts the linear scatterplots associated to these users with a linear straight regression line and a local smoothing regression line (Lowess algorithm [11]). The former type of regression is parametric and stands on the normal distribution, while the latter is non-parametric and it is aimed at supporting exploration and identification of patterns, enhancing the ability to see a line of best fit over data not necessarily normally distributed. Outliers from scatterplot are not removed: the rationale behind this decision is justified by the limited amount of points – maximum 9 points that coincides with the number of tasks.

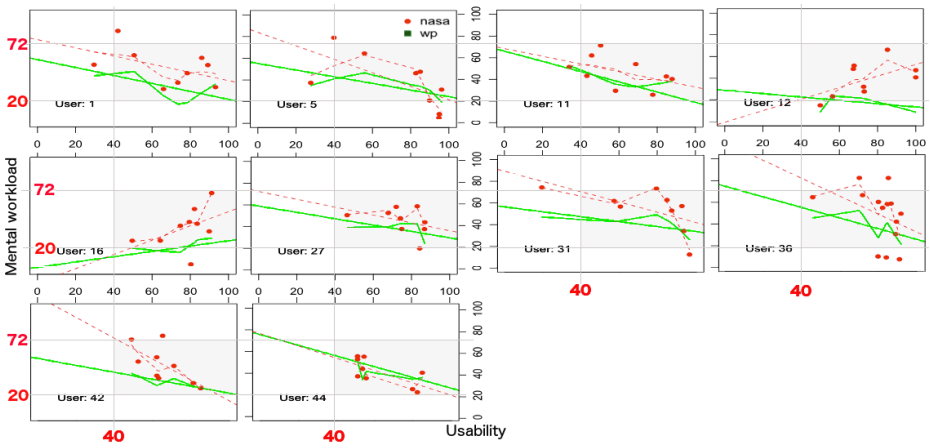

Fig. 9: Correlations MWL-usability for users with moderate/high Pearson and Spearman coefficients

No clear and consistent patterns emerge from figure 9. However, by analysing the mental workload scores ($NASATLX$ and WP), it is possible to note that the 10 selected users have all achieved, except a few outliers, a score of optimal mental workload (on average between 20-72). In other words, these users did not perceive underload or overload while executing the nine tasks. From an analysis of the usability assessments, all the users achieved scores higher than 40, indicating that no interface was perceived not usable at all. This might indicate that *when the mental workload experienced by users is within an optimal range, and usability is not bad, then the combination of mental workload and usability*

in a joint model might not be fully powerful in explaining objective performance more than mental workload alone. In the other cases, where correlation of mental workload and usability is almost inexistent, then a joint model might better explain objective performance. The following section is devoted to test this.

5.2 Testing hypothesis 2 - usability and mental workload impact performance more than just workload

From the previous analysis it appears that the perception of usability and the mental workload experienced by users are not related, except few cases in which mental workload was optimal and usability was not bad. Nonetheless, as previously reviewed, literature suggests that these constructs are important for describing and exploring the user's experience with an interactive system. For this reason a further investigation of the impact of the perception of usability and mental workload on objective performance has been conducted to test hypothesis 2 (section 4). In this context, objective performance refers to objective indicators of the performance of the volunteers who participated in the user study, categorised in 5 classes (section 4). During the experiment, the measurement of the objective performance of users was in some case faulty. These were discarded and a new dataset with 390 valid cases was formed. The exploration of the impact of the perception of usability and mental workload on the 5 classes of objective performance was treated as a classification problem, employing supervised machine learning. In detail, 4 different classification methods were chosen to predict the objective performance classes, according to different types of learning:

- information-based learning: decision trees (with Gini coefficient);
- similarity-based learning: k-nearest neighbors;
- probability-based learning: Naive Bayes;
- error-based learning: support vector machine (with a radial kernel) [8,23].

The distribution of the 5 classes is depicted in figure 10 and table 6:

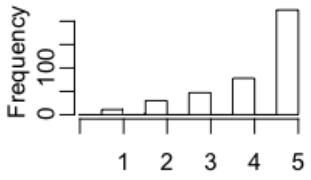

Fig. 10: Distribution of performance classes - original dataset

Table 6: Frequencies of classes

Class	Original	Oversampled
1	11	224
2	30	224
3	47	224
4	78	224
5	224	224
total	390	1120

Clearly, the above frequencies are unbalanced. For this reason a new dataset has been formed through oversampling, a technique to adjust class distributions and to correct for a bias in the original dataset, aimed at reducing the negative impact of class unbalance on model fitting. Random sampling (with replacement)

the minority classes to be the same size as the majority class is used (table 6). The two mental workload indexes ($NASA$ and WP) and the usability index (SUS) were treated as independent variables (features) and they were used both individually and in combination to form models aimed at predicting the 5 classes of objective performance (figure 11).

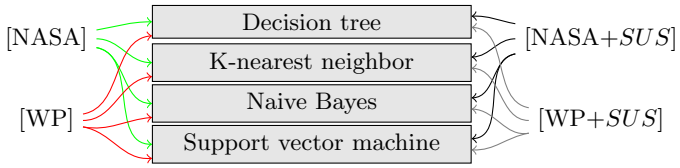

Fig. 11: Independent features and classification techniques

The independent features were normalised in the range $[0..1] \in \Re$ to facilitate the training of models and 10-fold stratified cross validation has been adopted in the training phase. In other words, the oversampled dataset was divided in 10 folds and in each fold, the original ratio of the distribution of the objective performance classes (figure 10, table 6) was preserved. 9 folds were used for training and the remaining fold for testing against accuracy and this was repeated 10 times changing the testing fold. This generated 10 models and produced 10 classification accuracies for each learning technique and for each combination of independent features (figure 12, table 7). It is important to note that training sets (a combination of 9 folds) and test sets (the remaining holdout set) were always the same across the classification techniques and the different combination of independent features (paired 10-fold CV). This is critical to perform a fair comparison of the different trained models using the same training/test sets.

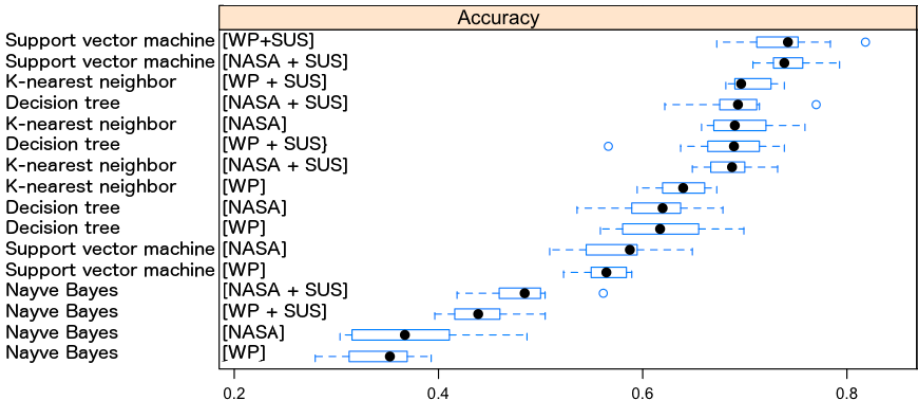

Fig. 12: Independent features, classification technique, distribution of accuracies with 10-fold stratified cross validation

To test hypothesis 2, the 10-fold cross-validated paired Wilcoxon statistical test has been chosen for comparing two matched accuracy distributions and to assess whether their population mean ranks differ (it is a paired difference test) [58]. This test is a non-parametric alternative to the paired Student's t-test selected because the population of accuracies (obtained testing each holdout set) was assumed to be not normally distributed. Table 8 lists these tests for the individual models (containing only the mental workload feature) against the combined models (containing the mental workload and the usability features). Except in one case (k-nearest neighbor, using the NASA-TLX as feature), the addition of the usability measure (SUS) to the mental workload feature (NASA or WP) always statistically significantly increased the classification accuracy of the induced models, trained with the 4 selected classifiers. This suggests how mental workload and usability can be jointly employed to explain objective performance measure, an extremely important dimension of user experience.

Classifier	Independent features	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Support vector machine	WP, SUS	0.6726	0.7140	0.7422	0.7368	0.7506	0.8182
Support vector machine	NASATLX, SUS	0.7080	0.7285	0.7387	0.7430	0.7534	0.7928
K-nearest neighbors	WP, SUS	0.6754	0.6971	0.7027	0.7091	0.7185	0.7748
Decision tree	NASATLX, SUS	0.6216	0.6769	0.6933	0.6937	0.7111	0.7699
K-nearest neighbors	NASATLX	0.6339	0.6497	0.6815	0.6822	0.7101	0.7297
Decision tree	WP, SUS	0.5664	0.6645	0.6894	0.6816	0.7136	0.7387
K-nearest neighbors	NASATLX, SUS	0.6549	0.6704	0.6861	0.6848	0.6971	0.7143
K-nearest neighbors	WP	0.5676	0.6182	0.6355	0.6331	0.6510	0.6818
Decision tree	NASATLX	0.6216	0.6470	0.6578	0.6615	0.6696	0.7027
Decision tree	WP	0.5586	0.5813	0.6170	0.6179	0.6511	0.6991
Support vector machine	NASATLX	0.5664	0.6097	0.6233	0.6189	0.6323	0.6757
Support vector machine	WP	0.5225	0.5503	0.5644	0.5625	0.5812	0.5893
Naive Bayes	NASATLX, SUS	0.4182	0.4596	0.4844	0.4827	0.4989	0.5614
Naive Bayes	WP, SUS	0.3964	0.4194	0.4389	0.4411	0.4602	0.5045
Naive Bayes	NASATLX	0.2973	0.3400	0.3527	0.3597	0.3943	0.4091
Naive Bayes	WP	0.2793	0.3139	0.3524	0.3428	0.3671	0.3929

Table 7: Ordered distributions of accuracies of trained models

			Accuracy (mean)			
Classifier	Model 1	Model 2	Model 1	Model 2	p-value	difference
Decision tree	NASA	NASA, SUS	0.6615	0.6937	0.032	yes
Decision tree	WP	WP, SUS	0.6179	0.6816	0.019	yes
K-nearest neighbor	NASA	NASA, SUS	0.6822	0.6848	1	no
K-nearest neighbor	WP	WP, SUS	0.6331	0.7091	0.005	yes
Nayve Bayes	NASA	NASA, SUS	0.3597	0.4827	0.001	yes
Nayve Bayes	WP	WP, SUS	0.3428	0.4411	0.001	yes
Support vector machine	NASA	NASA, SUS	0.6189	0.743	0.001	yes
Support vector machine	WP	WP, SUS	0.5625	0.7368	0.001	yes

Table 8: Wilcoxon test of distributions of accuracies with different independent features and learning classifiers

5.3 Summary of findings

In summary, from empirical evidence, the two hypotheses can be accepted.

- H_1: *Usability and Mental workload are two uncorrelated constructs (as measured with the selected self-reporting techniques (SUS, NASA-TLX, WP).*

They capture different variance in experimental tasks. This has been tested by a correlation analysis (both parametric and nonparametric) which confirmed that the two constructs are not correlated. The obtained Pearson coefficients suggest that there is no linear correlation between usability (SUS scale) and mental workload (NASA-TLX and WP scales). The Spearman coefficients confirmed that there is no tendency for usability to either increase or decrease when mental workload increases. The large variation in correlations within different tasks and for different individuals is interesting and worth of future investigation.

- H_2: *A unified model incorporating a usability and a MWL measure can better explain objective performance than MWL alone.*

This has been tested by inducing combined and individual models, using four supervised machine learning classification techniques, to predict objective performance of users (five classes of performance). The combined models were most of the times able to predict objective user performance significantly better than the individual models, according to the Wilcoxon non-parametric test.

6 Conclusion

This study attempted to investigate the correlation between the 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 were presented, with a particular focus on the latter. This because usability is a central notion in human-computer interaction, with a plethora of definitions and applications existing in the literature. Whereas, the construct of mental workload has a background in Ergonomics and Human Factors, but less mentioned in HCI. 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 46 subjects. Empirical evidence suggests that there is no relationship between the perception of usability of a set of web-interfaces and the mental workload imposed on users by a set of tasks executed on them. In turn, this suggests that the two constructs seem to describe two not overlapping phenomena. The implication of this is that they could be jointly used to better describe objective indicator of user performance, a dimension of user experience. Future work will be devoted to replicate this study employing a set of different interfaces, tasks and with different usability and mental workload assessment instruments. The contributions of this research are to offer a new perspective on the application of mental workload to traditional usability inspection methods, and a richer approach to explain the human-system interaction and support its design.

Table 9: System Usability Scale (SUS)

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 10: The NASA Task Load Index (NASA-TLX)

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 11: Workload Profile (WP)

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)?
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?
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)?

Table 12: 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.
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.

Table 13: Experimental tasks (M=manipulated; g=Group)

Task	**Description**	**Type**	**Task condition**	**Web-site**	**g. A**	**g. B**
T_1	Find out how many people live in Sidney	Fact finding	Simple search	Wikipedia		*M*
T_2	Read the content of `simple.wikipedia.org/wiki/Grammar`	Browsing	Not goal-oriented and no time pressure	Wikipedia	*M*	
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		*M*
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	*M*	
T_5	Find out the year of birth of the 1^{st} wife of the founder of playboy	Fact finding	Single task by time pressure (2-min limit). Each 30 secs user is warned of time left	Google		*M*
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		*M*
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	*M*	
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	*M*	
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		*M*

References

1. Addie, J., Niels, T.: Processing resources and attention. In: Handbook of human factors in Web design, pp. 3424–3439. Lawrence Erlbaum Associates (2005)
2. Albers, M.: Tapping as a Measure of Cognitive Load and Website Usability. Proceedings of the 29th ACM international conference on Design of communication pp. 25–32 (2011)
3. Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., Moret-Bonillo, V.: A Context-of-Use Taxonomy for Usability Studies. International Journal of Human-Computer Interaction 26(10), 941–970 (2010)
4. Annett, J.: Subjective rating scales in ergonomics: a reply. Ergonomics 45(14), 1042–1046 (Nov 2002)
5. Annett, J.: Subjective rating scales: science or art? Ergonomics 45(14), 966–987 (2002)
6. Baber, C.: Subjective evaluation of usability. Ergonomics 45(14), 1021–1025 (2002)
7. Bangor, A., T. Kortum, P., T. Miller, J.: An empirical evaluation of the system usability scale. International Journal of Human-Computer Interaction 24(6), 574–594 (2008)
8. Bennett, K.P., Campbell, C.: Support vector machines: Hype or hallelujah? SIGKDD Explor. Newsl. 2(2), 1–13 (December 2000)
9. Brooke, J.: Sus: A quick and dirty usability scale. In: Jordan, P.W., Weerdmeester, B., Thomas, A., Mclelland, I.L. (eds.) Usability evaluation in industry. Taylor and Francis, London (1996)
10. Cain, B.: A review of the mental workload literature. Tech. rep., Defence Research & Dev.Canada, Human System Integration (2007)
11. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. American Statistical Association 74, 829–836 (1979)
12. Cohen, J.: Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates. (1988)
13. Cooper, G.E., Harper, R.P.: The use of pilot ratings in the evaluation of aircraft handling qualities. Technical Report AD689722, 567, Advisory Group for Aerospace Research & Development (April 1969)
14. De Waard, D.: The measurement of drivers' mental workload. The Traffic Research Centre VSC, University of Groningen (1996)
15. Edwards, A., Kelly, D., Azzopardi, L.: The Impact of Query Interface Design on Stress, Workload and Performance, pp. 691–702. Springer International Publishing, Cham (2015), `http://dx.doi.org/10.1007/978-3-319-16354-3_76`
16. Fischer, G.: User modeling in human-computer interaction. User Modeling and User-Adapted Interaction 11(1-2), 65–86 (March 2001)
17. Gwizdka, J.: Assessing cognitive load on web search tasks. The ergonomic open journal 2(1), 114–123 (2009)
18. Gwizdka, J.: Distribution of cognitive load in web search. Journal of the american society & information science & technology 61(11), 2167–2187 (November 2010)
19. Harper, B.D., Norman, K.L.: Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5. In: 1st Annual Mid-Atlantic Human Factors Conference. pp. 224–228 (1993)
20. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Human Factors and Ergonomics Society Annual Meeting. vol. 50. Sage Journals (2006)
21. Hornbaek, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64(2), 79–102 (2006)

22. Huey, B.M., Wickens, C.D.: Workload transition: implication for individual and team performance. National Academy Press, Washington, DC. (1993)
23. Karatzoglou, A., Meyer, D.: Support vector machines in r. Journal of Statistical Software 15(9), 1–32 (2006)
24. Lehmann, J., Lalmas, M., Yom-Tov, E., Dupret, G.: Models of user engagement. In: Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization. pp. 164–175. UMAP'12, Springer-Verlag, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31454-4_14
25. Lewis, J.R.: Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 7, 57–78 (1995)
26. Longo, L., Kane, B.: A novel methodology for evaluating user interfaces in health care. In: Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on. pp. 1–6 (June 2011)
27. Longo, L.: Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In: INTERACT (4). pp. 402–405 (2011)
28. Longo, L.: Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In: UMAP, pp. 369–373 (2012)
29. Longo, L.: Formalising Human Mental Workload as a Defeasible Computational Concept. Ph.D. thesis, Trinity College Dublin (2014)
30. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behaviour and Information Technology 34(8), 758–786 (2015)
31. Longo, L.: Designing medical interactive systems via assessment of human mental workload. In: Int. Symposium on Computer-Based Medical Systems. pp. 364–365 (2015)
32. Longo, L.: Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). pp. 106–111 (June 2016)
33. Longo, L., Barrett, S.: A computational analysis of cognitive effort. In: Intelligent Information and Database Systems, Part II. pp. 65–74 (2010)
34. 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)
35. 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. SciTePress (April 2012)
36. Macleod, M.: Usability in context: Improving quality of use. In: Human Factors in Organizational Design and Management, Proceedings of the International Ergonomics Association 4th International Symposium. Elsevier (1994)
37. Moustafa, K., Saturnino, L., Longo, L.: Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In: 2017 1st International Symposium on Human Mental Workload: models and applications. vol. CCIS 726, pp. 30–50. Springer International Publishing (June 2017)
38. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L.E. (eds.) Usability Inspection Methods. Wiley & Sons, New York (1994)
39. Nielsen, J.: Usability inspection methods. In: Conference Companion on Human Factors in Computing Systems. pp. 377–378. CHI '95, ACM, New York, NY, USA (1995)

40. O' Donnel, R.D., Eggemeier, T.F.: Workload assessment methodology. In: Boff, K., Kaufman, L., Thomas, J. (eds.) Handbook of perception and human performance, vol. 2, pp. 42/1–42/49. New York, Wiley-Interscience (1986)
41. O'Brien, H.L., Toms, E.G.: What is user engagement? a conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology 59(6), 938–955 (2008), http://dx.doi.org/10.1002/asi.20801
42. Reid, G.B., Nygren, T.E.: The subjective workload assessment technique: A scaling procedure for measuring mental workload. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology, vol. 52, chap. 8, pp. 185–218. North-Holland (1988)
43. Rizzo, L., Dondio, P., Delany, S.J., Longo, L.: Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile, pp. 215–229. Springer International Publishing, Cham (2016), http://dx.doi.org/10.1007/978-3-319-44944-9_19
44. Roscoe, A.H., Ellis, G.A.: A subjective rating scale for assessing pilot workload in flight: a decade of practical use. Technical report 90019, Royal Aerospace Establishment, Farnborough (UK) (March 1990)
45. Rubio, S., Diaz, E., Martin, 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)
46. Saket, B., Endert, A., Stasko, J.: Beyond usability and performance: A review of user experience-focused evaluations in visualization. In: Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization. pp. 133–142. BELIV '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2993901.2993903
47. Schmutz, P., Heinz, S., Métrailler, Y., Opwis, K.: Cognitive load in ecommerce applications: Measurement and effects on user satisfaction. Advances in Human-Computer Interaction 2009, 3/1–3/9 (2009)
48. Shackel, B.: Usability - context, framework, definition, design and evaluation. Interact with Computers 21(5–6), 339–346 (December 2009)
49. Slaughter, L.A., Harper, B.D., Norman, K.L.: Assessing the equivalence of paper and on-line versions of the quis 5.5. In: 2nd Annual Mid-Atlantic Human Factors Conference. pp. 87–91 (1994)
50. Tracy, J.P., Albers, M.J.: Measuring Cognitive Load to Test the Usability of Web Sites. Usability and Information Design pp. 256–260 (2006)
51. Tsang, P.S.: Mental workload. In: Karwowski, W. (ed.) International Encyclopedia of Ergonomics and Human Factors (2nd ed.), vol. 1, chap. 166. Taylor & Francis (2006)
52. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics 39(3), 358–381 (1996)
53. Tsang, P.S., Vidulich, M.A.: Mental workload and situation awareness. In: Salvendy, G. (ed.) Handbook of Human Factors and Ergonomics, pp. 243–268. John Wiley & Sons, Inc. (2006)
54. Tullis, T.S., Stetson, J.N.: A Comparison of Questionnaires for Assessing Website Usability. In: Annual Meeting of the Usability Professionals Association (2004)
55. Vidulich, M.A., Ward Frederic G., S.J.: Using the subjective workload dominance (sword) technique for projective workload assessment. Human Factors Society 33(6), 677–691 (December 1991)
56. Wickens, C.D.: Multiple resources and mental workload. Human Factors 50(2), 449–454 (2008)

57. Wickens, C.D., Hollands, J.G.: Engineering Psychology and Human Performance. Prentice Hall, 3rd edn. (September 1999)
58. Wilcoxon, F.: Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6), 80–83 (Dec 1945), http://dx.doi.org/10.2307/3001968
59. Wilson, G.F., Eggemeier, T.F.: Mental workload measurement. In: Karwowski, W. (ed.) International Encyclopedia of Ergonomics and Human Factors (2nd ed.), vol. 1, chap. 167. Taylor & Francis (2006)
60. Xie, B., Salvendy, G.: Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. Work and Stress 14(1), 74–99 (2000)
61. Young, M.S., Stanton, N.A.: Mental workload. In: Stanton, N.A., Hedge, A., Brookhuis, K., Salas, E., Hendrick, H.W. (eds.) Handbook of Human Factors and Ergonomics Methods, chap. 39, pp. 1–9. CRC Press (2004)
62. Young, M.S., Stanton, N.A.: Mental workload: theory, measurement, and application. In: Karwowski, W. (ed.) International encyclopedia of ergonomics and human factors, vol. 1, pp. 818–821. Taylor & Francis, 2nd edn. (2006)
63. Zhu, H., Hou, M.: Restrain mental workload with roles in hci. In: Proceedings of Science & Technology for Humanity. pp. 387–392 (2009)
64. Zijlstra, F.R.H.: Efficiency in work behaviour. Doctoral thesis, Delft University, The Netherlands (1993)