Website complexity and usability: is there a role for mental workload?

Giovanni Serra

Department of Psychology,
Sapienza University of Rome,
Rome, Italy
Email: giovanni.serra@uniroma1.it

Federica De Falco

Department of Psychology,
Sapienza University of Rome,
Rome, Italy
and
Health Directorate,
Occupational Medicine Unit,
Bambino Gesù Children's Hospital IRCCS,
Rome, Italy
Email: federica.defalco@opbg.net

Piero Maggi and Rosa De Piano

Department of Psychology,
Sapienza University of Rome,
Rome, Italy
Email: piero.maggi@uniroma1.it
Email: rosa.depiano@uniroma1.it

Francesco Di Nocera*

Department of Psychology,
Sapienza University of Rome,
Rome, Italy
and
Department of Planning, Design, and Technology of Architecture,
Sapienza University of Rome,
Rome, Italy
Email: francesco.dinocera@uniroma1.it
*Corresponding author

Abstract: The concept and the assessment of mental workload – which is of great importance in the human factors community – has been overlooked in usability/user experience research. Mental workload depends on the characteristics of the interface and the nature of the task, and it may affect both user performance and usability. Mental workload assessment may represent a useful addition to human-computer interaction studies, particularly concerning those dealing with the design of user interfaces and the interaction with websites and apps. The objective of the present study was to evaluate the role of mental workload experienced by the individuals when browsing information-abundant websites: a case in which the task load imposed on the user may be particularly relevant. Three Italian Government websites with different levels of information complexity have been selected to test the research hypothesis. Results indicate that mental workload may contribute to the perception of usability and the overall user experience.

Keywords: information architecture; mental workload; eye-tracking; usability; website design.

Reference to this paper should be made as follows: Serra, G., De Falco, F., Maggi, P., De Piano, R. and Di Nocera, F. (2022) 'Website complexity and usability: is there a role for mental workload?', *Int. J. Human Factors and Ergonomics*, Vol. 9, No. 2, pp.182–199.

Biographical notes: Giovanni Serra received his PhD in Psychology and Cognitive Science from Sapienza University of Rome in 2021. His research activity has been mainly devoted to relationship between information architecture, mental workload and usability.

Federica De Falco held an appointment as Research Fellow at the Department of Psychology, Sapienza University of Rome. Currently, she is Research Psychologist at the Occupational Medicine Unit of Bambino Gesù Children's Hospital. Her research activity is mainly devoted to workplace health promotion.

Piero Maggi received his PhD in Psychology and Cognitive Science from Sapienza University of Rome in 2021. His research activity has been mainly devoted to mental workload and adaptive automation, with a focus on eye-tracking.

Rosa De Piano received her PhD in Information Engineering from University of Florence in 2018. She holds an appointment as a Research Fellow at the Department of Psychology since 2020. Her research activity has been mainly devoted to UX research and design for complex systems.

Francesco Di Nocera is an Associate Professor of Work and Organizational Psychology at Sapienza University of Rome. A former faculty member at the Department of Psychology since 2004, in 2021 he joined the Department of Planning, Design, and Technology of Architecture. His research activity has been mainly devoted to the interaction between humans and technology. Currently, he is working at a research program on behaviour-based design for approaching design issues within the behaviour analysis framework.

This paper is a revised and expanded version of a paper entitled 'The role of mental workload in determining the relation between website complexity and usability: an eye-tracking study' presented at Human Factors and Ergonomics Society Europe Chapter 2018 Annual Meeting, Berlin, 8–10 October 2018.

1 Users and operators: two tales on the human interaction with technology

Human interaction with technology is a complex field of studies integrating several traditional disciplines dealing with either humans (e.g., psychology, anthropology) or technology (e.g., engineering, computer science). During the last 50 years, this field of study has been labeled in many different ways as a function of the community hosting this type of research. Engineering psychology, human factors (HF), cognitive ergonomics, human-computer interaction (HCI) and, more recently, user experience are some of those labels indicating academic programs and courses, scientific and professional associations, not to mention professional certifications. Often the same theoretical and methodological background underlie the research carried out under those different labels. Still, sometimes differences in naming involve subtle stances generating a great divide and lack of communication between researchers. An interesting example is the choice of the researchers' object of interest: what 'human' and what 'technology' are taken into account? Roughly two research traditions can be acknowledged: HCI and HF. The HCI community takes into consideration 'users' who interact with consumer products, and it deals with topics including the design of usable user interfaces, the effects of technology on the accidental user, and the development of novel ways for interacting with the technology (Shneiderman et al., 2016). The HF community deals with 'operators' interacting with complex systems in critical settings, and many research studies in this field investigate topics including the assessment of mental workload, the effects of automation on the operator, and the reduction of human error (Wickens et al., 2016). User satisfaction and usability are key issues in the first case, whereas performance and safety are of primary importance for the other. Although many phenomena investigated in one field (but also metrics, constructs, standards) can usefully contribute to the other, knowledge appears somewhat compartmentalised. The reason underlying this divide can be traced back to the very nature of the type of technology existing when those research traditions started. HF was born long before the personal computer, and the only people dealing with certain types of technological artifacts were specialised operators, mainly in the military setting. Usability was not an issue for well-trained operators, but their performance level was crucial. On the contrary, HCI was born with the development of personal computers for common people. The term 'usability' generally refers to issues related to the quality of a system and its capability to be employed by users as a tool to achieve particular goals. In the literature, there are several definitions of usability that differ depending on the theoretical framework. A commonly accepted definition is that provided by the International Organization for Standardization, which defines usability as "The extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use" (International Standards Organization, 2018).

Effectiveness refers to the completeness and accuracy in the achievement of goals by the user. Efficiency refers instead to the optimisation of cognitive and temporal resources of the user. Finally, satisfaction concerns the issues of the comfort of use and the acceptability of the product. The context of use can be considered the fourth dimension of usability, and it concerns users' characteristics, their goals, and the environment in which they operate. Usability is not an intrinsic characteristic of the product. It depends on the characteristics of the user who uses it, the goal to be achieved, and the context in which

the product is used. For this reason, usability should not be traced back to the presence/absence of specific attributes but should always be evaluated by taking into account the skills, perceptions, and objectives of the end-user. However, in the last 50 years, both terms of the interaction we wish to investigate have changed: technology has become pervasive and invisible, and individuals are today more tech-savvy than ever before. Therefore, the distinction between users and operators is now only related to the context where the interaction with technology occurs. Therefore, usability-related and safety-related topics belong rightfully to the study of the interaction with any type of technological artifact.

With that in mind, mental workload assessment – which is of great importance in the HF community – may represent a useful addition to the HCI investigation, particularly for the areas dealing with the design of user interfaces. Mental workload has been traditionally defined as the difference between the demands imposed by the task on the individual and the resources available to perform the task (Hancock and Meshkati, 1988). The 'resource' is limited in nature, and it might be conceptualised as organised in different reservoirs serving tasks that require specific types of resources according to the input modality, the code (i.e., spatial or verbal), the stage of processing, and the response modality (see Wickens, 2008 for a recent account). Several perspectives on mental workload can be found in the literature (see Young et al., 2015 for a recent review), but all of them share the idea that it is a complex phenomenon due to the interaction between the requirements of the task, the circumstances under which it is performed and the skills, behaviors, and perceptions of the individual. While this construct is commonly invoked in studies investigating the interaction with a large set of systems, including some consumer devices (e.g., GPS navigators), it is commonly neglected to assess the interaction with websites and apps. However, there is no doubt that browsing a large website searching for a piece of information can be an activity that taxes attentional resources. Particularly, the mental load imposed by the task due to the characteristics of the interface may contribute to the usability assessment made by the individual interacting with it, therefore affecting both the user performance and their usability perceptions. Although the relation between those two constructs has been theoretically acknowledged by some authors since the 90s (see Bevan and Macleod, 1994), mental workload assessment has never been central in existing usability models (Harrison et al., 2013). Moreover, despite the different fields of application, the research on mental workload and usability share some goals (i.e., improving performance, reducing errors, and mitigating cognitive load), experimental techniques, and methodologies. Both research areas, for example, use performance (e.g., number of errors, execution times, success rate) and physiological metrics such as EEG and eye movement analysis to investigate the interaction between an individual and an interface/system. The reference literature clearly shows how poor performance is associated with high mental workload (Young et al., 2015) and poor usability (Albert and Tullis, 2013). Nevertheless, experimental studies on the intersection between usability and mental workload experienced by users are sparse and not conclusive, sometimes supporting a relation (e.g., Lukanov et al., 2016; Mazur et al., 2019) and sometimes showing weak or no correlation between the two constructs (e.g., Longo, 2018).

2 Visual scanning and mental workload

The analysis of eye movements has achieved popularity during the last decades (Duchowski and Duchowski, 2017), and the availability of affordable eye-tracking systems led to the wide use of ocular metrics in various fields. Many studies have contributed to investigating the effect of mental workload on ocular activity (see McCarley and Kramer, 2008 for an account on eye-tracking in neuroergonomics). One may speculate that any study using eye-tracking while a user interacts with an interface could be considered a study on mental workload, even without explicit reference. That occurs because the ocular activity may be representative of the processing load: how long is the user fixating on an area? How large are the movements needed to reach a specific location? How many transitions are the eyes making to explore the scene? How dispersed or clustered is the pattern of fixations? All those questions address the mental effort needed to perform a task. Since the first studies on this topic, among which there is the seminal work by Paul Fitts in the aviation domain (Fitts et al., 1950), we have gained more insight into the functioning of fixations. We know that the frequency of fixations on a specific area is an indication of its relevance for the individual; the duration of the fixations is directly proportional to the difficulty of information processing; and that short transitions between nearby areas of interest (AOIs), therefore small saccadic amplitude, indicate a correct arrangement of the information in the individual's visual field (see also Van Orden et al., 2001).

Usability studies employing eye movement data consistently showed that non-optimal arrangement of the interface elements is associated with a greater number of fixations and longer duration, as well as with visual search strategies characterised by transitions between non-contiguous AOIs (Ehmke and Wilson, 2007; Goldberg et al., 2002; Kotval and Goldberg, 1998; Wang et al., 2019). Reaching a goal effectively and efficiently through using a system while enjoying the experience (that is what usability is all about) is influenced by the system's characteristics (e.g., interface layout, information architecture, labelling). Of course, there is a connection between ocular activity and usability. However, for a deeper study of the interaction experience, the variable effort the user puts in reaching an objective and, therefore, their experience of different levels of mental workload should be considered.

Among the ocular metrics that have been proposed as a measure of mental workload, those that consider the entire scanpath (that is, the sequence of fixations and saccades) are particularly appealing. The reason behind this interest is that having a single index that provides information about the entire visual scanning during the interaction with a system is a parsimonious strategy. This strategy allows for comparing ocular activity along with segments of the interaction, among different systems, among versions of the same system, or contrasting different categories of users. The first attempt in this direction can date back to the early '80s when a research group at NASA Langley (Ephrath et al., 1980; Harris et al., 1986; Tole et al., 1983) introduced the concept of 'entropy' in the analysis of eye movements. In Thermodynamics, the definition of 'entropy' refers to the quantity of disorder in a system; in those studies, 'entropy' was related to the disorder occurring in visual exploration. According to this concept, as workload increases, the exploration pattern becomes more stereotyped (i.e., less random). In contrast, as mental workload decreases, the randomness of the pattern should increase. One limitation of this approach is that it needs pre-defined AOIs for computing the transitions on which the index is based. Moreover, entropy appears to be sensitive only to variations due to the

visuospatial demand, therefore excluding those changes in workload due to the temporal demand (Kruizinga et al., 2006; Maggi and Di Nocera, 2021). To overcome those limitations, more recent studies by Di Nocera and colleagues (see Di Nocera et al., 2007, 2015) successfully employed a statistical indicator to assess the spatial distribution of eye fixations on the entire visual scene. This indicator, the nearest neighbour index (NNI), is sensitive to variations in mental workload. The NNI compares the mean distance between pairs of (nearest) fixations pairs to that expected based on chance (random distribution). The index expresses a single value that can theoretically vary between 0 (maximum clustering) and 2.1491 (strictly regular hexagonal pattern: same distance between points in the distribution). Values close to 1 indicate that the distribution of fixations is not different from a random distribution; values greater than 1 indicate a dispersion of the fixation pattern, and values less than 1 indicate fixation grouping. The dispersion of fixations appears to be associated with the mental workload when the task load depends on temporal demand. In contrast, fixation clustering seems to be associated with the mental workload when the task load depends on visuospatial demand. In other words, in tasks where the workload is due to changes in the temporal demand, a tendency towards fixation spreading reflects the individual's promptness in detecting stimuli. In contrast, fixation clustering indicates the need to focus on a portion of the scene in tasks where the workload is due to changes in the visuospatial demand. Additionally, one significant feature of this index is that once there are enough fixations (~50), it can be estimated for small epochs (e.g., 1 minute), therefore creating time series that represent the ongoing mental workload. Details on how to compute this index can be found in dedicated publications (Camilli et al., 2008; Di Nocera et al., 2016).

The experimental study reported in the following sections has been designed to assess the mental workload experienced by users interacting with information-abundant websites with varying information architecture complexity. It is expected that greater complexity of the information architecture would be associated with higher levels of mental workload and lower usability evaluations.

3 Experimental study

The objective of this study was twofold:

1 to evaluate the mental workload associated with browsing information-abundant websites with different levels of complexity\

2 to understand whether workload affects perceived usability.

We hypothesised that a greater complexity of the information architecture structure would be related to higher mental workload and poorer usability evaluations.

Three large Italian public administration's (PA) websites (whose identity we are not allowed to disclose) were selected after a heuristic evaluation of their information architecture structures. The heuristic evaluation was based on a breakdown of the information architecture structure. The breakdown allowed the identification of the number of levels (main categories) and the number of categories/labels contained in each level. The sum of the number of categories for each website was used as a criterion to estimate its complexity in terms of information architecture (Table 1). In Italy, the 'Guidelines for the design of public administration websites' were published in January

2017 (http://design.italia.it/) and PA s must adhere to these guidelines to ensure good user experience for citizens. The selected websites are among the first to be designed according to these guidelines, so they were similar in the design and interaction features (i.e., menu structure, colours, fonts, aesthetic) but different in terms of information architecture complexity. Specifically, they have a similar structure in terms of:

- header: the header of all websites contained information regarding contacts, the 'search' bar, the logo, and the website name
- menu: the menu was 'drop-down and horizontal' on the main page (some items were visible, others were hidden but became viewable on mouseover); the menu was 'vertical and fixed' on secondary pages (all menu items were always visible in the left column)
- body: the central part of the page showed the latest news of interest, videos and images
- footer: the bottom part of each website contained information such as contacts and site map.

Hereinafter, the selected websites will be referred to as website 1 (low-complexity), website 2 (medium-complexity), and website 3 (high-complexity) (Table 1).

Table 1 Information architecture of the selected websites

Website	*Information architecture structure (number of levels and number of categories per level)*							
	Level 1	*Level 2*	*Level 3*	*Level 4*	*Level 5*	*Level 6*	*Total*	*Complexity*
1	5	44	134	170	-	-	353	Low
2	5	53	160	190	53	8	469	Medium
3	6	45	109	266	158	56	630	High

3.1 Participants

Twenty volunteers participated in this study (7 females; mean age = 57, st.dev. = 6; mean years of education = 11, st.dev = 2). All of them were native Italian speakers, they were naïve as to the aims, the expected outcomes, and the methodology of the experiment, and they declared to have normal or corrected-to-normal vision. All the subjects declared to use the Internet every day. Participants were employees of an Italian public agency and the age group was compatible with that reported by the Italian National Institute of Statistics (see www.dati.istat.it for the interactive database) as the main group of users accessing PA websites (i.e. 45 to 64 years). This study was performed in accordance with the Helsinki Declaration of the World Medical Association.

3.2 Materials and method

The X2-30 eye-tracking system (Tobii, Sweden) was used to record eye movements during the interaction with the websites. This is a standalone eye tracker that can be used in various set-ups by attaching it to monitors, laptops or to perform eye-tracking on

physical objects with a sampling rate of 30 Hz. In this study we used the Tobii I-VT Fixation Filter with the follow parameters:

Table 2 Fixations filter parameters

Function	*Function parameters*	*Function parameter values*
Gap fill-in (interpolation)	Max gap length	75 ms
Eye Selection	N/A	Average
Noise reduction	N/A	Disabled
Merge adjacent fixations	Max time between fixations	75 ms
	Max angle between fixations	0.5°
Discard short fixations	Minimum fixation duration	100 ms

Fixations are first divided in 1-min length bins and then analysed using the NNI algorithm. NNI values are therefore averaged to be used as a dependent variable. This procedure is necessary because a spatial pattern is the result of a process evolving over time, and analysing the distribution all at once would introduce a bias.

Performance measures were collected during the execution of the tasks:

- *Success rate:* task success is the most widely used performance metric. It measures how effectively users are able to complete a certain task (Nielsen, 2001). Researchers distinguish two different types of task success: 'binary success' and 'levels of success' (Hornbæk, 2006). In this study 'binary success' has been used as a behavioral indicator of usability and mental workload. We consider 'successfully completed' only the task in which the participants reached the correct landing page where they could find the information they were looking for.
- *Task completion time:* it is usually used to measure the efficiency of a system (Albert and Tullis, 2013). In this study we considered task completion time as the amount of time the user needs to complete all the assigned tasks.
- *Backtracking events:* when a user goes back to a previous situation it is possible that he/she performed an action without obtaining the expected result. In this study, we considered backtracking events as the number of times a participant returned back to the previous webpage after visiting one that was useless for his/her goals. Generally, backtracking events reflect user confusion and are associated with poor usability (Akers et al., 2012).

At the end of the interaction with each website, subjective measures of perceived usability and mental workload were collected using the following scales:

- *Net Promoter Score*® (NPS) (Reichheld, 2003; Reichheld and Covey, 2006): it consists of a single item: 'How likely would you recommend this website to a friend or colleague?' to which subjects can answer using an 11-point scale (0 to 10). This tool is based on the classification of users (customer) in three categories: Promoters (provide a score between 9 and 10), Neutrals (provide a score between 7 and 8) and Detractors (provide a score between 0 and 6). The final value of the NPS is obtained by subtracting the percentage of detractors from the percentage of promoters and can be compared to benchmarks. In the present study, the value (0 to 10) will be used as a dependent variable. Bradner and Sauro (2012) reported strong correlations between

good usability design and the likelihood to recommend a software product. Word of mouth is critical for the success of PAs websites and digital services. Therefore, the NPS was included as a measure.

- *Usability Evaluation 2.0* (Us.E. 2.0) (Di Nocera, 2013): a multidimensional questionnaire to evaluate website usability. The questionnaire was developed in Italian and consists of 19 items subdivided into three subscales, representing the framework users would adopt for evaluating the quality of their interaction with the interface: (Mental) Handling, Satisfaction, and Attractiveness. Us.E. 2.0 allows a quick assessment of website perceived usability, identifying critical issues that could be eventually addressed by more extensive testing and re-design. Users are required to answer to all items along with a five-point Likert scale (ranging from 'strongly disagree' to 'strongly agree'). The '(mental) handling' scale includes 11 items and measures the interaction with the structure of the website (e.g., information architecture, layout). Examples of items are: 'In this website I found myself on the point of getting lost' and 'I always feel in control of the operations that are allowed in this website'. Low scores in this scale would suggest the need to make changes to information architecture or page layout. The 'satisfaction' scale includes six items and measures the perceived utility of the website. Examples of items are: 'Exploring this website was a waste of time' and 'I managed to obtain the information/service that I was looking for'. Low scores in this scale may indicate that the website does not meet users' needs, either because the users are not those expected by who created the website or because contents/services are not those expected by the users. Finally, the 'attractiveness' scale measures the interaction with the aesthetic features of the website. This scale is composed only of two items: 'The choice of the colours used in this website is smart' and 'the graphics used in this website are catchy and detailed'. Low scores in this scale would suggest the need for a restyling. Raw scores are standardised as z-scores using normative data (mean and standard deviation) from large pools of previously collected data and divided into four websites categories: portals and communities, universities, authorities and PAs, companies and services. Normative data for each category were reported by Di Nocera (2013). In the current study, the 'PA' pool was used for standardisation.
- *NASA Task Load Index* (NASA-TLX) (Hart and Staveland, 1988): this self-report measure is widely used to assess the mental workload after the interaction with a system. The respondent provides an evaluation of their perceived workload along six scales (responses ranging from 0 to 100): mental demand, physical demand, temporal demand, effort, performance, and frustration. Here we used the NASA-TLX short administration (i.e., raw scores without weighing procedure).

3.3 Procedure

Five equivalent research tasks that included similar activities (e.g., downloading a form, obtaining information about a service, examining a table containing data) have been proposed for each website. The tasks were designed by taking into account the depth of the information architecture. In this way, similar tasks between different websites could be performed successfully with the same minimum number of clicks. The websites, as

explained above, have a similar structure in terms of the organisation of the main menu. Subjects had to search for specific information in different areas of the websites. Each website had the categories 'administration', 'services', and 'open data'. To give an example, task 1 for website #1 asked participants to find information within the 'administration' section and required a certain number of clicks. Similarly, task 1 for website #2 and website #3 asked participants to find the information contained within the same section of the websites (i.e., 'administration') and required the same number of clicks to complete. Prior knowledge of websites was investigated by asking participants if they had ever browsed the selected websites. All participants reported that they had never browsed the websites under investigation.

Participants performed the entire test in three separate sessions. The single sessions were performed at about 15 days apart to limit effects related to fatigue and task duration. Moreover, to avoid effects related to the order of presentation of the stimuli, the websites and the tasks were randomly assigned to the participants. Specifically, each session included:

1 *Familiarisation with the website under investigation:* a free navigation session in which the participants experienced the website structure (duration: ~5 minutes).

2 *Eye-tracker calibration:* participants were positioned at a distance of about 60 cm from a 22" screen, they performed a dynamic nine-point calibration always starting from the centre of the screen (duration: ~3 minutes).

3 *Tasks:* participants in each session performed five research tasks on one of the target websites; the centre of the screen was the starting fixation point for each task. A time limit of six minutes has been assigned for each task; at the end of each task, participants reported their perceptions about the level of complexity of the task on a scale from 1 to 5 (1 = Not difficult at all; 5 = Extremely difficult). (duration: ~30 minutes).

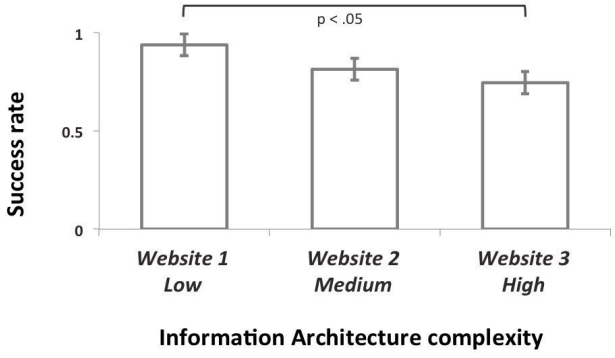
4 *Questionnaires:* after completing all the five tasks, participants filled a questionnaire concerning their personal information (i.e., gender, age, educational qualification, employment, the frequency of internet use), the NPS, the Us.E. 2.0, and the NASA-TLX (duration: ~10 minutes).

3.4 Results

Success rates, completion times, backtracking events, perceived complexity scores, NPS scores, Us.E. 2.0 (Handling, Satisfaction, Attractiveness) scores, NASA-TLX overall scores, and NNI average scores were analysed in repeated-measures ANOVA designs using Complexity (website 1 vs. website 2 vs. website 3) as repeated factor.

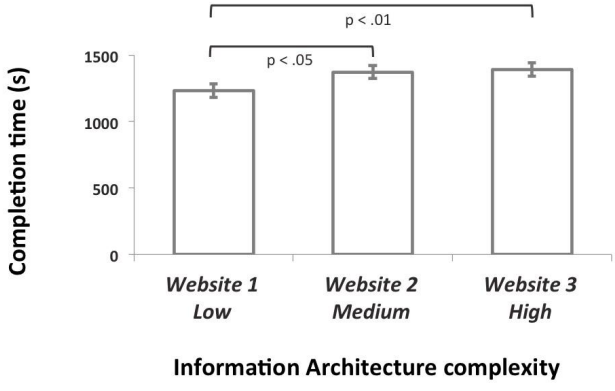
The analysis carried out on the angular transformations of success rates showed a tendency towards statistical significance between websites [$F(2,36) = 3.04$; $p = .06$; $\eta_p^2 = .14$]. Duncan post-hoc testing showed that the success rate for the high-complexity website (website 3) was significantly lower than the low-complexity website (website 1). At the same time, no significant differences emerged between the medium-complexity website (website #2) and the other two (Figure 1).

Figure 1 Success rate per website



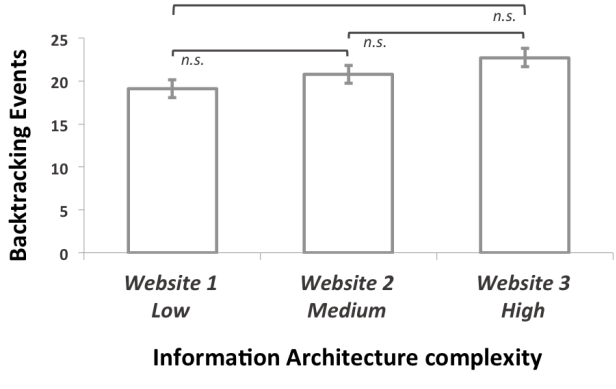
Completion time was significantly different between websites [$F(2,36) = 4.02$; $p < .05$; $\eta_p^2 = .18$]. Duncan post-hoc testing showed that completion time for the low-complexity website (website #1) was significantly faster than the other two (Figure 2).

Figure 2 Completion time (mean) per website



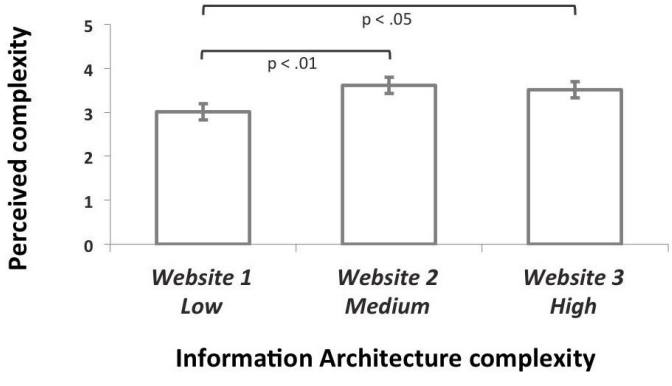
Backtracking events in the website structure were not significantly different between websites [$F(2,36) = .85$; $p = .43$); $\eta_p^2 = .04$] (Figure 3).

Figure 3 Backtracking events per website



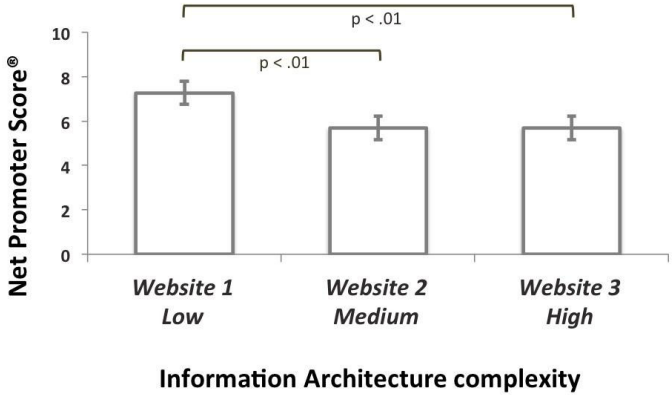
Perceived complexity was significantly different between websites [F(2,36) = 3.92; $p < .05$); $\eta_p^2 = .18$] Duncan post-hoc testing showed that perceived complexity of the low-complexity website (website #1) was significantly lower than the other two (Figure 4).

Figure 4 Perceived complexity per website



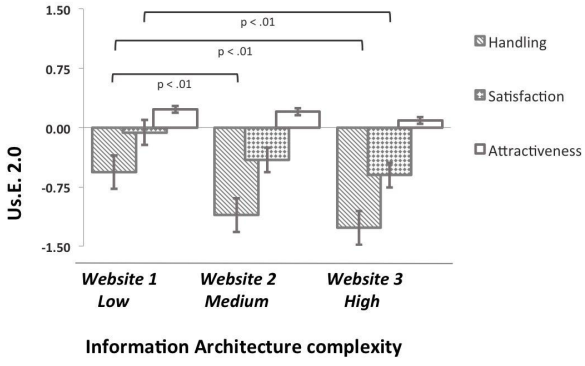
NPS score was significantly different between websites [F(2,36) = 4.52; $p < .05$; $\eta_p^2 = .20$]. Duncan post-hoc testing showed that the proportion of the low-complexity website (website #1) was significantly higher than the other two (Figure 5).

Figure 5 Net promoter score per website



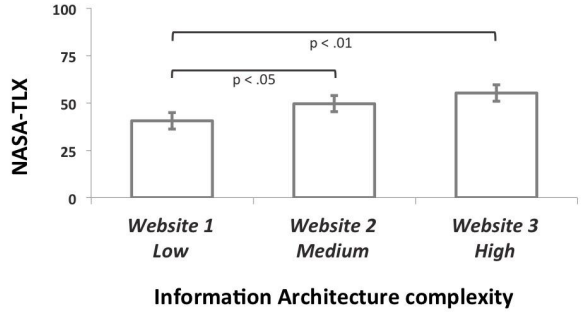
Us.E. 2.0 scores were significantly different between websites. Specifically, Duncan post-hoc testing showed that scores for the low-complexity website (website #1) was significantly higher than the other two for the dimensions (mental) 'handling' [F(2,36) = 6.80, $p < .01$; $\eta_p^2 = .27$] and 'satisfaction' [F(2,36) = 3.45, $p < .05$; $\eta_p^2 = .16$]. No significant differences were found for the dimension 'attractiveness' [F(2,36) = .28, $p > .05$; $\eta_p^2 = .02$] (Figure 6).

Figure 6 Us.E. 2.0 per website (z-scores)



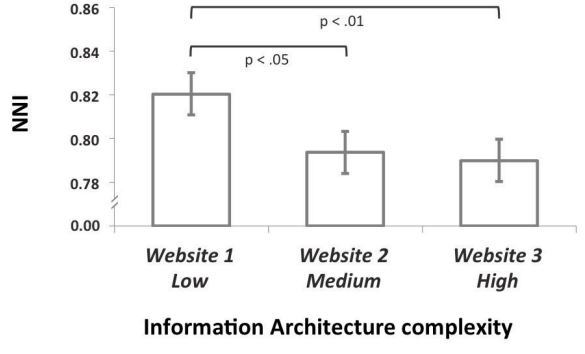
NASA-TLX scores were significantly different between websites [$F(2,36) = 7.38$; $p < .01$; $\eta_p^2 = .29$]. Duncan post-hoc testing showed that the perceived workload for the low-complexity website (website #1) was significantly lower than the other two (Figure 7).

Figure 7 NASA-TLX scores per website



The NNI was significantly different between websites [$F(2,36) = 6.41$; $p < .01$; $\eta_p^2 = .26$]. Duncan post-hoc testing showed that the fixation pattern associated with medium- and high-complexity websites (websites #2 and #3) were significantly more clustered than the low-complexity website (Figure 8).

Figure 8 Nearest neighbour index per website



4 Discussion

The objective of the present study was to evaluate the role of the mental workload experienced by the individual in the formation of the usability perceptions when browsing information-abundant websites: a case in which the task load imposed on the user may be particularly relevant. Three Italian government websites with different levels of information complexity have been selected to test the research hypothesis. Results indicate a correspondence between workload and usability measures in the hypothesised direction. However, the decrement of workload and the increment of usability was not perfectly aligned with the website complexity. A significant difference was expected between all three complexity conditions but the difference was statistically significant only between the low complexity condition and the other two, for both workload and usability indicators. Possibly, our assessment of information complexity was not sufficiently fine-grained, and that is reflected in the results. A more accurate analysis of the information architecture could be obtained by including not only the depth of the website structure and the number of categories, but also a consideration of the labelling system used (i.e., the names used for the website categories). In that way, it would be possible to test the logical effectiveness of the information architecture and to verify whether the labelling system corresponds to the mental model of the end user of the website. Considering that the websites examined all belong to PAs, have the same main structure, and use a similar labelling system, we considered this additional level of analysis unnecessary for the purposes of this research. Nevertheless, this represents a limitation of this study and should be addressed in future investigations by selecting websites that are clearly differentiated in terms of information architecture. Yet, the sites selected for this study had the advantage of belonging to the same category, and to share the same type of layout, as well as the same aesthetics.

As we have reported in the introductory section, studies investigating the relation between workload and usability are sparse and controversial. While some studies found that the mental workload imposed by the characteristics of the interface may contribute to the usability assessment, others show weak or no correlation between the two constructs. For example, Longo (2018) in a recent study involving information-seeking web-based tasks reported no relationship at all between mental workload and usability and suggested that they should be considered separately. However, the study was carried out using a well-known website (i.e., Wikipedia) and manipulating the task load by increasing the difficulty of the task. In that case, changes in workload may have no effect on the perception of usability, because a change in the difficulty of the task does not derive from the website itself, but it's artificially imposed.

In our case, instead, the task load imposed derived from the varying complexity of the information architecture of the websites employed. Participants self-assessed complexity (perceived complexity) and results showed that the low-complexity website was actually rated as significantly less complex than the other two. Consistently, success rate was higher and completion time shorter for the low-complexity website than the high-complexity website (albeit no significant difference was found with respect to the medium-complexity website). That suggests consistency between the users' perception and performance.

Usability evaluations are generally negative when users take too long to complete the task, make mistakes, or fail in its execution (Albert and Tullis, 2013; Nielsen and Levi, 1994; Nielsen, 1999). Similarly, poor efficacy (in terms of number of successes) and poor

efficiency (in terms of number of errors, completion time) are related to higher mental workload (Eggemeier et al., 1991; Young et al., 2015).

Indeed, results showed significantly higher values in both the '(mental) handling' and 'satisfaction' scales of the Us.E. 2.0 questionnaire for the low-complexity website than the other two websites. The '(mental) handling' scale, which measures interaction with website structure (i.e., information architecture), received the lowest scores from participants, indicating that they acknowledged a structural issue with all websites. As expected the 'attractiveness' scale did not show any difference between websites, because they were designed using the same layout and colour scheme. Also, differences in the NPS scores were significantly higher for the low-complexity website than the other two and, as we reported above, high NPS scores (indicating promotion of the website) are associated with more positive usability perceptions (Bradner and Sauro, 2012).

As for the workload, NASA-TLX scores for the low-complexity website were significantly lower for the low-complexity website than the other two, confirming that participants experienced less workload while browsing the less complex website. This result is corroborated by the analysis of the ocular behavior: NNI values associated with the highest-complexity websites were significantly lower than the low- and medium-complexity websites indicating more fixations' grouping and therefore higher mental workload. As reported by Camilli et al. (2008), the visuospatial demand would lead to more grouped pattern of fixations, because the mental operations involved in the spatial task would prevent the use of the specific resources (visual and spatial) needed by visual scanning.

5 Conclusions

Showing the linkage between mental workload and usability does not directly provide any information about which variable causes the other. Nevertheless, as we reported in the introduction of this paper, mental workload depends on the characteristics of the interface and the nature of the task, and may affect both the user performance and the user usability perceptions. Additionally, while the self-assessment of the mental load obtained with the NASA-TLX might be as well a consequence of poor interaction with a system, the changes in the distribution of eye fixations along with the increase in the visuospatial demand, cannot. The consistency of the results obtained with the two measures indicates that the mental workload might not be the consequence of the usability perception, but it might be related to the structural aspects of the websites. For those reasons, we are inclined to consider the experienced mental workload as a key factor affecting usability, particularly for its 'mental handling' component. Of course, this is a first study and we do not consider our results as conclusive. Nevertheless, the finding that mental workload may contribute to the usability assessment, and that may influence the overall user experience should not be ignored. In conclusion, the cognitive effort made by the user should become a commonly assessed variable in this area of study in the same way it is central in the HF literature.

References

Akers, D., Jeffries, R., Simpson, M. and Winograd, T. (2012) 'Backtracking events as indicators of usability problems in creation-oriented applications', *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 19, No. 2, pp.1–40.

Albert, W. and Tullis, T. (2013) *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Elsevier, Waltham, MA.

Bevan, N. and Macleod, M. (1994) 'Usability measurement in context', *Behaviour & Information Technology*, Vol. 13, Nos. 1–2, pp.132–145.

Bradner, E. and Sauro, J. (2012) 'Software user experience and likelihood to recommend: linking UX and NPS', *UPA International Conference*, Henderson, 4–8 June.

Camilli, M., Terenzi, M. and Di Nocera, F. (2008) 'Effects of temporal and spatial demands on the distribution of eye fixations', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52, No. 18, pp.1248–1251.

Di Nocera, F. (2013) *Usability Evaluation 2.0: Una descrizione (s)oggettiva dell'usabilità*, Ergoproject, Roma.

Di Nocera, F., Camilli, M. and Terenzi, M. (2007) 'A random glance at the flight deck: pilot's scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, Vol. 1, No. 3, pp.271–285.

Di Nocera, F., Capobianco, C. and Mastrangelo, S. (2016) 'A simple (r) tool for examining fixations', *Journal of Eye Movement Research*, Vol. 9, No. 4, pp.1–6.

Di Nocera, F., Ranvaud, R. and Pasquali, V. (2015) 'Spatial pattern of eye fixations and evidence of ultradian rhythms in aircraft pilots', *Aerospace Medicine and Human Performance*, Vol. 86, No. 7, pp.647–651.

Duchowski, A.T. and Duchowski, A.T. (2017) *Eye Tracking Methodology: Theory and Practice*, Springer, London.

Eggemeier, F.T., Wilson, G.F., Kramer, A.F. and Damos, D.L. (1991) 'Workload assessment in multi-task environments', in Damos, D.L. (Ed.): *Multiple-Task Performance*, pp.207–216, Taylor & Francis, London.

Ehmke, C. and Wilson, S. (2007) 'Identifying web usability problems from eye-tracking data', *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI ... But Not as We Know It – Volume*, BCS Learning & Development Ltd., Swindon, UK, September, pp.119–128.

Ephrath, A.R., Tole, J.R., Stephens, A.T. and Young, L.R. (1980) 'Instrument scan – is it an indicator of the pilot's workload?', *Proceedings of the Human Factors Society Annual Meeting*, Vol. 24, No. 1, pp.257–258.

Fitts, P.M., Jones, R.E. and Milton, J.L. (1950) 'Eye movements of aircraft pilots during instrument-landing approaches', *Aeronautical Engineering Review*, Vol. 9, No. 2, pp.24–29.

Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N. and Wichansky, A.M. (2002) 'Eye tracking in web search tasks: design implications', *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, Association for Computing Machinery, New York, NY, March, pp.51–58.

Hancock, P.A. and Meshkati, N. (Eds.) (1988) *Human Mental Workload*, Amsterdam, North-Holland.

Harris, R.L., Glover, B.L. and Spady, A.A. (1986) *Analytic Techniques of Pilot Scanning Behavior and their Application*, Technical Paper No. 2525, NASA Langley Research Center, Hampton, VA.

Harrison, R., Flood, D. and Duce, D. (2013) 'Usability of mobile applications: literature review and rationale for a new usability model', *Journal of Interaction Science*, Vol. 1, No. 1, pp.1–16.

Hart, S.G. and Staveland, L.E. (1988) 'Development of NASA-TLX (task load index): results of empirical and theoretical research', in Hancock, P.A. and Meshkati, N. (Eds.): *Human Mental Workload*, pp.139–183, Elsevier Science Publishers, Amsterdam.

Hornbæk, K. (2006) 'Current practice in measuring usability: challenges to usability studies and research', *International Journal of Human-Computer Studies*, Vol. 64, No. 2, pp.79–102.

International Standards Organization (2018) *Ergonomics of Human System Interaction – Part 11: Definitions and Concepts*, ISO 9241-11:2018, International Standards Organization, Geneva, Switzerland.

Kotval, X.P. and Goldberg, J.H. (1998) 'Eye movements and interface component grouping: an evaluation method', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, October, Vol. 42, No. 5, pp.486–490.

Kruizinga, A., Mulder, B. and de Waard, D. (2006) 'Eye scan patterns in a simulated ambulance dispatcher's task', in de Waard, D., Brookhuis, K.A. and Toffetti, A. (Eds.): *Developments in Human Factors in Transportation, Design, and Evaluation*, pp.305–317, Shaker Publishing, Maastricht, The Netherlands.

Longo, L. (2018) 'Experienced mental workload, perception of usability, their interaction and impact on task performance', *PloS ONE*, Vol. 13, No. 8, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199661.

Lukanov, K., Maior, H.A. and Wilson, M.L. (2016) 'Using fNIRS in usability testing: understanding the effect of web form layout on mental workload', *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, May, pp.4011–4016.

Maggi, P. and Di Nocera, F. (2021) 'Sensitivity of the spatial distribution of fixations to variations in the type of task demand and its relation to visual entropy', *Frontiers in Human Neuroscience*, Vol. 15, p.642535.

Mazur, L.M., Mosaly, P.R., Moore, C. and Marks, L. (2019) 'Association of the usability of electronic health records with cognitive workload and performance levels among physicians', *JAMA Network Open*, Vol. 2, No. 4, pp.e191709–e191709.

McCarley, J.S. and Kramer, A.F. (2008) 'Eye movements as a window on perception and cognition', in Parasuraman, R. and Rizzo, M. (Eds.): *Neuroergonomics: The Brain at Work*, Oxford University Press, New York, NY.

Nielsen, J. (1999) *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, Thousand Oaks, CA.

Nielsen, J. (2001) '*Success Rate: The Simplest Usability Metric* [online] https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/ (accessed 20 November 2020).

Nielsen, J. and Levy, J. (1994) 'Measuring usability: preference vs. performance', *Communications of the ACM*, Vol. 37, No. 4, pp.66–76.

Reichheld, F.F. (2003) 'The one number you need to grow', *Harvard Business Review*, Vol. 81, No. 12, pp.46–55.

Reichheld, F.F. and Covey, S.R. (2006) *The Ultimate Question: Driving Good Profits and True Growth*, Harvard Business School Press, Boston, MA.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N. and Diakopoulos, N. (2016) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Pearson, Hoboken, NJ.

Tole, J.R., Stephens, A.T., Vivaudou, M., Ephrath, A.R. and Young, L.R. (1983) '*Visual Scanning Behavior and Pilot Workload*, NASA Contractor Report No. 3717, NASA Langley Research Center, Hampton, VA.

Van Orden, K.F., Limbert, W., Makeig, S. and Jung, T.P. (2001) 'Eye activity correlates of workload during a visuospatial memory task', *Human Factors*, Vol. 43, No. 1, pp.111–121.

Wang, J., Antonenko, P., Celepkolu, M., Jimenez, Y., Fieldman, E. and Fieldman, A. (2019) 'Exploring relationships between eye tracking and traditional usability testing data', *International Journal of Human-Computer Interaction*, Vol. 35, No. 6, pp.483–494.

Wickens, C.D. (2008) 'Multiple resources and mental workload', *Human Factors*, Vol. 50, No. 3, pp.449–455.

Wickens, C.D., Hollands, J.G., Banbury, S. and Parasuraman, R. (2016) *Engineering Psychology and Human Performance*, Routledge, New York, NY.

Young, M.S., Brookhuis, K.A., Wickens, C.D. and Hancock, P.A. (2015) 'State of science: mental workload in ergonomics', *Ergonomics*, Vol. 58, No. 1, pp.1–17.

Appendix

English adaptation of Usability Evaluation 2.0 [Di Nocera, (2013), p.71]. Subscales are mental handling (H), satisfaction (S), and attractiveness (A). Responses are provided along a five-point Likert scale ranging from strongly disagree to strongly agree.

Item		*Sub-scale*
1	While exploring this website I always knew where I was	(H)
2	This website did not meet my expectations	(S)
3	The contents of this website were clear from the beginning	(S)
4	This site is as pretty as it is useless	(S)
5	I felt disoriented while exploring this website	(H)
6	The choice of the colours used in this website is smart	(A)
7	I can easily reach the main menu	(H)
8	This website is useless while pretending to be useful	(S)
9	It is difficult to browse this website	(H)
10	The graphics used in this website are catchy and detailed	(A)
11	Visiting this website was as easy as using the software application I use the most	(H)
12	In this website I can find what I'm looking for without having to explore it all	(H)
13	The contents of this website are updated	(S)
14	In this website I found myself on the point of getting lost	(H)
15	I managed to obtain the information/services that I was looking for	(S)
16	I always feel in control of the operations that are allowed in this website	(H)
17	The information presented in this website is understandable	(S)
18	Exploring this website was a waste of time	(S)
19	This website is made up of long lists that are difficult to examine	(H)