Thesis dissertation in fulfilment
of requirements for the XXXIII PhD Program
in Psychology and Cognitive Science

SAPIENZA
UNIVERSITÀ DI ROMA

WEBSITE COMPLEXITY AND USABILITY: THE ROLE OF MENTAL WORKLOAD

PhD Student:
Giovanni Serra
Sapienza University of Rome,
Italy

Advisor:
Prof. Francesco Di Nocera

Co-Advisor:
Prof. Stefano Sdoia

Abstract

The evolution of digital interfaces has changed the way people perform their main daily activities (Parameswaran & Whinston, 2007; Wattal et al., 2010). The spread of the first personal computers, between the second half of the 1970s and the first half of the 1980s, marks the beginning of the digitisation process. The first machines were complex computing systems that only experts were able to use. With the "revolution" of graphic interfaces and the introduction of the paradigm "What you see is what you get" (Goldberg, 1988; Myers, 1984; Smith et al., 1982), there was a rapid spread of computers and digital interfaces for non-expert users. All individuals today interact with web-based interfaces and systems for most of the day, both at work and in their private life. This technological development, which involved the transition from physical to graphical interfaces, has been accompanied and guided by the development of a multidisciplinary scientific discipline known as "Human-Computer Interaction" (Grudin, 1990; Preece et al., 2015).

The term "Human-Computer Interaction" (HCI) generally refers to the study of the design, evaluation and development of technological interfaces (Hewett et al., 1996). HCI benefits from the contribution of disciplines such as engineering, computer science, cognitive psychology, sociology and anthropology (Carroll, 1997; Rogers, 2004; Lazar, Feng & Hochheiser, 2017). The main research areas of HCI include the study of cognitive processes involved during human-machine interaction, the development of principles, guidelines and heuristics to be applied in the design and evaluation of interfaces (Preece et al., 2015). Cognitive psychology has strongly contributed to the development of human-centred interfaces. HCI has integrated the main assumptions of cognitive models in the design phases of digital instruments and devices (Rogers, 2004). Some noteworthy examples refer to the studies on memory conducted by Baddeley (1992) and Miller (1956) or the tripartite theory of Atkinson and Shiffrin (1968). These studies have primarily influenced how human-machine interaction models are conceived and implemented. (Card et al., 1983). The main objective of the researchers has been to reduce the cognitive load resulting from individual-interface interaction (Mandel, 1997; Preece et al., 2015). Several studies have shown that when cognitive demands are either too high, or too low, these

negatively affect the performance and user experience (Xie and Salvendy, 2000). Mental workload is crucial in the design and evaluation of digital interfaces. Digital system designers must be able to predict the mental workload imposed by the system. Paying attention to users' mental workload is essential to enable them to use digital systems with effectiveness, efficiency and satisfaction, taking care to improve "perceived usability" and decrease user "mental workload".

Research in this area has produced numerous studies and identified different metrics capable of estimating the mental workload experienced by a user during the execution of a task (O'Donnel & Eggemeier, 1986; Wierwille & Eggemeier, 1993). However, we are still far from being effectively integrated into usability assessments. This deficiency in the reference literature is mainly due to the fact that the constructs of mental workload and usability have been investigated in different fields of application and with reference to different types of users.

HCI Researchers (see Carrillo et al., 2017 for review) distinguished "inexperienced" or "occasional users" by "operators" and "expert users". While the first ones have been the focus of User eXperience (UX) studies, the second ones have been more involved in Human Factor (HF) research. Research involving occasional users focused on the user experience that emerges from the interaction with an interface. The topics of greatest interest in UX research are "User satisfaction", "Ease of use", "Consistency", "Affordance". Researchers focus their efforts on designing intuitive and easy-to-use interfaces that support the user in achieving their goals. In this field, the concept of "usability" is fundamental and generally refers to aspects related to the quality of a system. Human Factors studies on operators and expert users, on the other hand, focused on the human-machine interaction in "high complexity" work environments (e.g., aviation, aerospace, healthcare, etc.) where specific skills and knowledge are required to use certain interfaces.

HF deals with psychological and physiological aspects of human capability, able to influence the interaction between the operator, the systems and the procedures of his working environment, and to directly affect the outcome of events. The high safety standard required in domains, such as aviation and healthcare industry, led to a wide exploration and application of HF theories and practices to the design of services, system interfaces and procedures. Some of the topics of greatest interest in HF research are "human error", "situation awareness", "automation" and the "mental workload" experienced by an operator who interacts with a system.

This compartmentalised division has the consequence that for a long time, there has never been any mention of "mental workload" in usability studies.

However, the concept of mental workload could provide important guidance for both the design phase of a system and the evaluation and improvement of its usability. Identifying reliable metrics that can provide objective information on these phenomena as well as their relation remains an open challenge for the scientific community. However, in this regard, eye movements analysis is a promising measure. Eye tracking has gained increasing popularity in the scientific community as a technique to evaluate usability and mental workload (May et al., 1990; Jacob & Karn, 2003; Pan et al., 2004; Poole & Ball, 2006; Majaranta & Bulling, 2014). The analysis of eye behaviour (e.g. fixations, saccades and scanpath analysis) allows researchers to obtain real-time information on user-interface interaction. Moreover, this technique has high ecological validity and returns objective measurements.

Considering this brief premise, the purpose of this work is to contribute to the investigation of the relationship between mental workload and usability in relation to the use of digital interfaces. The constructs of usability and mental workload, and the main subjective and objective metrics used for their evaluation will be illustrated. A particular focus will be made on the eye movements recording techniques used for the usability and mental workload evaluation. In the last part of this work, some experimental studies will be presented. The structure of the manuscript consists of four parts. In the first part, the concept of usability will be described. The most widely used definition in the literature is reported, and the basic principles of usability are explained. Considerable attention is devoted to the different techniques used for its measurement. The differences between formative assessment methods and summative or comparative assessment methods (Scholtz, 2004), techniques involving users and heuristics-based techniques will be explained. In addition, the main subjective usability questionnaires will be described in detail.

The second part will focus on the mental workload construct. After illustrating the main definitions and theories behind this construct, the various techniques useful for its measurement will be explained. In particular, the manuscript will focus on the most used physiological and subjective measures in the reference literature.

The third part provides an examination of the use of eye movements in psychology, the main characteristics of eye behaviour, the metrics and recording methodologies used. Also, this part of the work focuses on the most widely used ocular metrics for a measure of usability and mental workload, highlighting their differences and convergences.

In the fourth part, different experimental studies will be presented. This section will deal with some research questions accompanied by the results of the relative experimental analyses. The first study had the objective to validate a new tool for a quick and reliable estimation of the usability perceived during the use of digital interfaces. The comparisons between the new tool and other usability questionnaires will be analysed. The second and third studies investigate the relationship between usability perceptions and mental workload associated with browsing complex websites. A promising metric based on the analysis of eye movements, the NNI, was used in these studies. This metric provides real-time information about the mental workload experienced by a user when using an interface. Although the results obtained are encouraging, defining these comprehensive studies would be ambitious. In conclusion, the limitations that emerged during the studies are discussed, and ideas for future research are proposed.

Table of content

List of tables

List of figures

1. Usability

The term "*usability*" generally refers to issues related to the quality of a system and its capability to be used by users as a tool to achieve particular objectives. In the literature, there are several definitions of usability that differ depending on the theoretical framework. The most shared by the international community refers to the ISO 9241-11 introduced by the International Organization for Standardisation, 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".

Effectiveness refers to the completeness and accuracy in the achievement of objectives by users. Efficiency refers instead to the optimisation of the use of cognitive and temporal resources of the user. Finally, satisfaction concerns the issues of the comfort of use and acceptability of the product. The context of use can be considered the fourth dimension of usability and refers to users' individual characteristics, their objectives 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 objective 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 taking into account the skills, perceptions and objectives of the end-user (Di Nocera, 2013). Usability is one of the most important concepts in HCI (Chalmers, 2003; Sharp, Rogers & Preece, 2007).

According to Nielsen (1994a), a usable system should respect these fundamental principles:

- "Learnability": the system must guarantee easy and fast learning of its functionalities, input and output modes.
- "Efficiency": once the user has learned how to use the system, he must be able to perform his tasks accurately and efficiently.
- "Memorability": the interaction modes and functionalities of the system must be easy to remember for a user even after non-use periods.
- "Error frequency": the system should support users in order to minimise errors during its use. When an error occurs, the system should help users to

fix it and continue to perform the task easily.

- “User Satisfaction”: this principle concerns the user's subjective impressions about the use of the system.

The exponential development of web services has led to the concept of “web usability” (Nielsen, 1999). Individuals use the Internet to find and understand information (consulting news, downloading documents), share information (work, social networks), discover and access services (online shopping, administrative practices). Web usability, therefore, refers to the ability of websites and web tools to enable the user to perform these tasks effectively, efficiently and satisfactorily (Nielsen & Loranger, 2006).

Nielsen’s principles can also be adapted for the web context:

- “Learnability”: in this case, it refers to the ease with which users understand how to use the content and services offered by the website, such as searching for specific information using hypertext links. Each page of the front-end hypertext should be structured to ensure easily understandable content and easy to identify navigation mechanisms.
- “Efficiency”: users should be able to orient themselves within the website and achieve their goals quickly. The website should offer at any time the possibility to go back to the starting point and make the user understand in which part of the website he is.
- “Memorability”: within a web page, after a non-use period, users must be able to remember the navigation mechanisms and know how to browse around the website.
- “Error frequency”: in case users have made a wrong action (e.g., selecting an item for purchase, downloading a document) the website should offer the possibility to cancel it and return to the starting point.
- “Satisfaction”: also in this case it refers to the positive impressions related to the interaction with the website. Users should feel that they have control over the website, and that can orient their choices and navigation according to their goals.

In short, usability is very important to ensure a comfortable and satisfying user experience. Several researchers have shown how usable websites lead to better performance for users and how, on the contrary, the lack of attention to usability principles greatly increases the probability that a user will fail in the task and leave the website (Took, 1990; Buschke 1997; Chain Store Age 1997; Nielsen, 1999).

1.1. Measuring usability

The only application of usability principles in the design phase of a product/service is not enough to guarantee its effective usability. It is essential to evaluate, step by step, the usability of the product through assessments targeted to detect possible problems that occur when the user uses the product (Di Nocera, 2013).

A valid and reliable measurement of usability should follow certain principles. First of all, it is good practice to perform usability checks during the very early stages of an interface design, so that changes can be implemented early, avoiding excessive changes to the prototypes. In addition, the evaluation should be iterative, i.e. it should not be limited to a single observation, but consist of several measurements, integrating new tests each time a change is implemented. Therefore, while the first phases will consist of orientation for future design, in the following phases, quantitative evaluations will be carried out, such as those related to the achievement of objectives. From then on, therefore, it will be possible to assess the degree of adequacy of the system with respect to the context of use (ISO 9241-210, 2008).

Therefore, the assessment is composed of two activities: (a) *verification*, i.e. checking that the product is consistent with what is expressed in the requirements documents by comparing the characteristics of the product and what is indicated in the requirements; (b) *validation*, i.e. checking that the product actually meets the needs for which it was designed through tests involving users and stakeholders (Polillo, 2010). It is possible to define usability evaluation as a systematic process of collecting data, in order to have a better understanding of users and how user groups use the product to perform a specific task under specified conditions.

The scientific community distinguishes two main methods for usability assessment: “formative assessment methods” and “summative or comparative assessment methods” (Scholtz, 2004). Formative evaluations take place during product design and their purpose is to verify that the design of the product respects the usability principles and the needs of the final user. Formative evaluations are carried out on relatively small samples (3-5 subjects), so they do not allow statistical control on the data, although they contribute to "shape" the product. These methods are called "quick & dirty", because they allow us to evaluate in advance whether the project is developing in the right direction and, consequently, decide whether to continue to invest or not in solutions that will then have to be discarded or redesigned. In fact, the early identification of potential problems will allow companies to save money during the subsequent troubleshooting phases (Boscarol, 2010). As suggested by Nielsen (1994a), these techniques fall within the phenomenon of "discount usability", as they are both smooth and cheap in the identification of usability problems. In fact, according to the "Nielsen's rule", to identify 75% of the problems of an interface it would be appropriate to perform repeated usability tests with five users, rather than a single test with twenty users (Nielsen & Molich, 1990). In fact, the first five users will highlight most of the relevant usability problems, while subsequent users would only confirm the same result.

Summative methods, on the other hand, consist of an overall evaluation of the product, which is carried out later or at the end of the design and development process. The summative or comparative evaluations are useful to detect the problems that emerge during the interaction with the user, their purpose is to improve this interaction and to model the characteristics of the final product based on the real use that users make of it.

Within these macro-categories, we can distinguish two types of evaluations:

- Evaluations performed through the involvement of users;
- Evaluations performed without the involvement of users.

1.1.2. Usability evaluations performed through user testing

Commonly known as "User testing", evaluations involving users are intended to analyse user behaviour during the interaction with a website. The researcher asks users to perform several tasks within the website, detecting a number of indicators useful to estimate its usability such as main problems, execution time, type and number of errors, user satisfaction or frustration (Shneiderman et al., 2016). The test results allow researchers to identify usability problems and to implement solutions for product/service improvement. Usability researchers use specific laboratory settings that include a room where the user performs the test (i.e., user room) and a room where usability experts, product managers and designers follow the interaction between the user and the product (i.e., control room). After the completion of the task, the researchers analyze the user's behaviour by viewing the audio/video materials recorded during the test. The recordings allow researchers to understand the causes of any critical issues that would not otherwise be detected.

Currently, there are also several web platforms for conducting tests remotely, which allow to reach a larger number of subjects and to optimise time.

A method widely used during user testing is the “Think-aloud protocol”. This technique is an empirical method that consists of a protocol of “verbal research”. Thinking aloud was first developed in the psycho-social sciences (Ericsson & Simon, 1993) and later introduced in the field of usability by Lewis (1982). Today, the thinking aloud is widely used to locate misjudgements and other usability issues that occur at the exact moment of interaction (Nielsen, 1994b). This technique consists in asking the user to express and report aloud all thoughts related to the interaction with the system, product or service, verbalizing any difficulty, real or perceived. The main objective of the empirical evaluation of usability is to obtain practical indications about the difficulties encountered by the user, with the ultimate aim of improving the system, product or service. For this reason, the “think-aloud” is a particularly functional method for this purpose, which allows observers to gather precise and direct information about the strategies and difficulties encountered by users during the execution of tasks.

Two types of “think-aloud” are commonly used: the “concurrent” version and the “retrospective” version (Van den Haak et al., 2003). The concurrent “think-aloud” ensures that the researcher interacts with the participant during the performance of the tasks. On the contrary, in the retrospective “think aloud” the participant is invited to verbalize after the test the problems emerged during the navigation, providing, if necessary, a video replay of the performance just performed. It is necessary that the choice of using one type rather than the other is always contextualized. For example, the retrospective “think-aloud” is recommended when the participant and researcher speak different languages or when the participants show difficulties in verbalization or reading (Borsci et al., 2013). To ensure good reliability of the results, it is essential that the test is carefully planned. Usability testing should generally be carried out in accordance with the following steps:

- *Identify the test objectives:* before proceeding with the design of the test, it is necessary to define clear objectives on which to set the assessment. The objectives may be general (testing the usability of the website) or specific (testing the usability of specific sections or features of the website).
- *Select a sample of participants*: it is generally preferable to select a representative sample of the real users of the website. However, the choice of the sample should not be rigid but should vary according to the objectives of the study. Regarding the number of users to involve, most studies in the literature (Virzi, 1992; Nielsen, 1994b; Turner et al., 2006) state that the involvement of five users is sufficient to reveal more than 80% of the usability problems of a website.
- *Design the usability tasks:* the tasks to be assigned to the participants must be representative of the activities carried out on the website, they must be structured in a clear and understandable way, moreover all the objectives must be really achievable by the users.
- *Define the usability indicators*: before the start of each test it is necessary to establish what will be measured, i.e. what will be the indicators of the website usability. Generally researchers use both subjective measures, such as user satisfaction or difficulty of use that can be investigated through interviews and questionnaires, and objective and quantitative measures, such as the number of tasks successfully completed, the task completion time, the number and type of errors.
- *Prepare the experimental setting and materials:* it is preferable to carry out the tests in a quiet environment where the participant feels comfortable and can carry out the assigned tasks without interruption. The room should be organized in such a way that the researcher can observe the participant during the execution of the

test, take notes and communicate with him/her, but without interfering with the execution of the task. Among the materials it is good to include both paper observation grids and video recording systems.

- *Launch a pilot test:* before involving users it is necessary to conduct one or more pilot tests. The pilot test is useful to validate the tasks or modify them if they are unclear or impossible to perform due to website problems.
- *Launch tests with users*: in the first moment the researcher explains to participants the objectives and the methods of conducting the test; then he takes note of the behaviour of the participant during the various tasks. In this phase, the researcher collects data related to the usability indicators (e.g., time of execution, the success rate of the different tasks, questionnaires, etc.).
- *Analyse the data and prepare the final report*: at the end of the user tests the researcher proceeds with the data analysis in order to obtain global measures on the website usability. The final report must indicate crucial information such as the number of participants and tasks assigned, the success rate for each task, the results of the questionnaires administered, the list of the main problems encountered.

1.1.3. Usability evaluations without users

The evaluations based on user involvement have the advantage of ensuring accuracy and ecological validity. These techniques, in fact, consider the needs and ways of interacting with the website of its end users the centre of the assessment. However, they require the investment of many economic and temporal resources. For this reason, faster and cheaper usability assessments that do not involve users are often used. These types of assessments are generally referred to as “inspection methods” and are based on the judgment of a few experts (Nielsen, 1994c).

The most used inspection methods are two:

1. *methods based on heuristics,* in which a team of experts judges whether the website respects the fundamental principles of usability;
2. *methods based on "cognitive walkthrough"*, a technique that, through a detailed task analysis, tries to reconstruct the cognitive path that the user follows while performing certain operations within the website. This technique analysesin particular, the way in which the system interacts with the user to identify the correct actions to perform.

The evaluation through heuristics is commonly based on the application of the ten heuristics proposed by Nielsen (1994c):

- *"Visibility of system status"*: users should always be informed of system operations with easy to understand and highly visible status displayed on the screen within a reasonable amount of time;
- "*Match between system and the real world"*: designers should endeavour, to mirror the language and concepts users would find in the real world based on who their target users are. Presenting information in logical order and piggybacking on user's expectations derived from their real-world experiences will reduce cognitive strain and make systems easier to use;
- *"User control and freedom"*: offer users a digital space where backward steps are possible, including undoing and redoing previous actions.
- *"Consistency and standards":* interface designers should ensure that both the graphic elements and terminology are maintained across similar platforms. For example, an icon that represents one category or concept should not represent a different concept when used on a different screen;
- *"Error prevention"*: whenever possible, design systems so that potential errors are kept to a minimum. Users do not like being called upon to detect and remedy problems, which may on occasion, be beyond their level of expertise. Eliminating or flagging actions that may result in errors are two possible means of achieving error prevention.
- *"Recognition rather than recall"*: minimize the cognitive load by maintaining task-relevant information within the display while users explore the interface. Human attention is limited and we are only capable

of maintaining around five items in our short-term memory at one time. Due to the limitations of short-term memory, designers should ensure users can simply employ recognition instead of recalling information across parts of the dialogue. Recognizing something is always easier than recall because recognition involves perceiving cues that help us reach into our vast memory and allowing relevant information to surface. For example, we often find the format of multiple-choice questions easier than short answer questions on a test because it only requires us to recognize the answer rather than recall it from our memory.

- *"Flexibility and efficiency of use"*: with increased use comes the demand for fewer interactions that allow faster navigation. This can be achieved by using abbreviations, function keys, hidden commands and macro facilities. Users should be able to customize or tailor the interface to suit their needs so that frequent actions can be achieved through more convenient means.
- *"Aesthetic and minimalist design"*: keep clutter to a minimum. All unnecessary information competes for the user's limited attentional resources, which could inhibit a user's memory retrieval of relevant information. Therefore, the display must be reduced to only the necessary components for the current tasks, whilst providing clearly visible and unambiguous means of navigating to other content.
- *"Help users recognize, diagnose and recover from errors"*: designers should assume users are unable to understand technical terminology, therefore, error messages should almost always be expressed in plain language to ensure nothing gets lost in translation.
- *"Help and documentation"*: ideally, users should navigate the system without having to resort to documentation. However, depending on the type of solution, documentation may be necessary. When users require help, ensure it is easily located, specific to the task at hand and worded in a way that will guide them through the necessary steps towards a solution to the issue they are facing.

The cognitive walkthrough method is based on a reconstruction of the problem-solving strategies of the website users (Wharton, 1994). In particular, a team of expert evaluators analyzes the relationships between users' objectives, how they achieve them and the visible states of the interface in order to highlight any problems of interaction with the website. This method is particularly suitable for identifying problems related to the ability of users to learn the functionality of the website.

These evaluation methods are particularly useful when the researcher has few resources available. However, the effectiveness of these methods is highly dependent on the experience and skills of the evaluators. Moreover, they are based on a prototypical user, for this reason, they do not take into account the real experience of the end-users of the website (Nielsen, 1994c; Di Nocera, 2013).

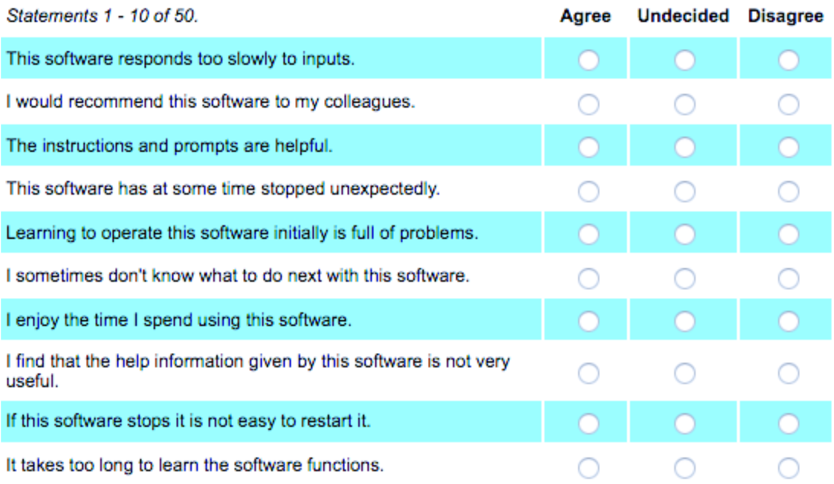
1.1.4. Subjective usability measures

Among subjective usability measures, questionnaires are the most used. Questionnaires are standardized tools that allow us to obtain information on users' perceptions about the usability of a given system/product or website. Questionnaires can be administered quickly and easily, and they are also cost-effective, as they allow us to reach a large sample in a short time. Typically, a questionnaire, as a subjective measure of usability, requires the user to express an evaluation of the browsing experience just completed. In order to be used as an evaluation tool, the questionnaire must respect some fundamental principles of statistics, including reliability and validity: the first refers to the accuracy of the instrument, while the second, the ability of a measurement to actually capture the characteristic under consideration (Ercolani, Areni & Leone, 2001).

The following pages will illustrate the main usability questionnaires used in the reference literature. In particular, the Software Usability Measurement Inventory - SUMI (Kirakowsky & Corbett, 1993), the Website Analysis and Measurement Inventory - WAMMI (Kirakowski et al., 1998) and the Questionnaire for User Interaction Satisfaction - QUIS (Chin, Diehl & Norman, 1988), the System Usability Scale - SUS (Brooke, 1996), Usability Metric for User Experience - UMUX (Finstad, 2010; 2013) and UMUX-LITE (Lewis, 2013), the Net Promoter Score - NPS (Reichheld, 2003), the Standardized User Experience Percentile Rank Questionnaire - SUPR-Q (Sauro, 2015), WebQual 4. 0 (Barnes & Vidgen, 2001) and Usability System Evaluation - Us.E. 2.0 (Di Nocera, 2013) will be described.

Software Usability Measurement Inventory - SUMI

The Software Usability Measurement Inventory - SUMI (Kirakowsky & Corbett, 1993), developed by the University College of Cork (Bevan & Macleod, 1994), is a standardized pencil-paper tool that measures user satisfaction and, consequently, the perceived quality of specific software. Specifically, the questionnaire consists of 50 items that investigate the attitudes of the user engaged in a particular task, i.e. involved in a particular context of use with a system (ISO 9241-11, 1991). The questionnaire is applicable to any software that requires interaction with an interface (e.g. display, keyboard, etc.). For each item the user can respond by choosing between three options "Agree", "Undecided" and "Disagree".



Statements 1 - 10 of 50.	Agree	Undecided	Disagree
This software responds too slowly to inputs.	○	○	○
I would recommend this software to my colleagues.	○	○	○
The instructions and prompts are helpful.	○	○	○
This software has at some time stopped unexpectedly.	○	○	○
Learning to operate this software initially is full of problems.	○	○	○
I sometimes don't know what to do next with this software.	○	○	○
I enjoy the time I spend using this software.	○	○	○
I find that the help information given by this software is not very useful.	○	○	○
If this software stops it is not easy to restart it.	○	○	○
It takes too long to learn the software functions.	○	○	○

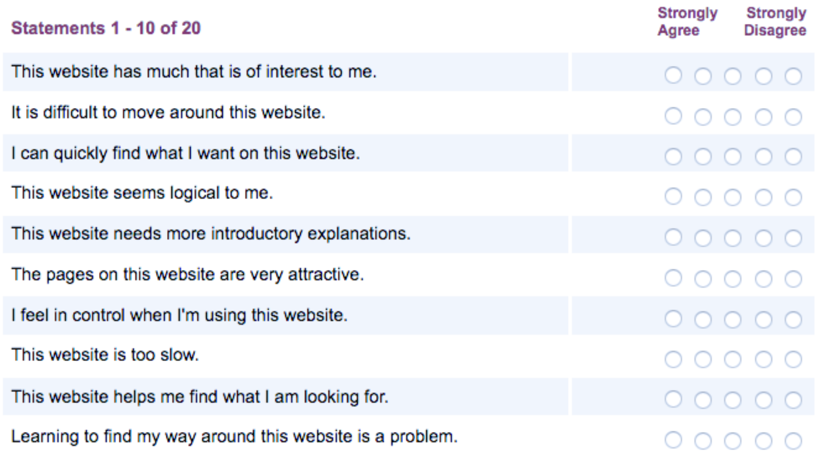
Figure 1.1. Some items from the SUMI Questionnaire.

A minimum of 12 subjects is required to obtain reliable measures (Kirakowsky & Corbett, 1993).

SUMI consists of various dimensions divided into three hierarchical levels of output. The first level provides global information related to usability; the second level, instead, investigates users' cognitive load considering five sub-dimensions (i.e., Interest, Efficiency, Learning, Availability and Control); the third level coincides with the item "Consensual Analysis", a descriptive index that allows to compare the single answers to the questionnaire, with the corresponding standardized scores related to the SUMI database. The standardization of SUMI was carried out starting from the analysis of a database consisting of more than 200 different types of software (word processors, spread sheets, communication programs, etc.). Moreover, the questionnaire shows a good reliability and a good ability to distinguish between different types of software. SUMI has been used effectively to assess new products during product evaluation, make comparisons between products or versions of products, and set targets for future application developments. Moreover, SUMI has been used specifically within development environments to set verifiable goals for user experience, track achievement of targets during product development, highlight good and bad aspects of an interface (Kirakowsky & Corbett, 1993).

Website Analysis and Measurement Inventory - WAMMI

The research center of the University College of Cork within the MUSiC project (Bevan & Macleod, 1994) proposed the use of a specific questionnaire, the Website Analysis and Measurement Inventory - WAMMI (Kirakowski, Claridge & Whitehand, 1998) for the evaluation of web usability. The WAMMI questionnaire consists of 20 items to which the user responds providing a degree of agreement on a Likert scale from 1 “Strongly agree” to 5 “Strongly disagree”. WAMMI proposes five factors to assess the usability of websites: attractiveness, controllability, efficiency, helpfulness and learnability.



Statements 1 - 10 of 20	Strongly Agree				Strongly Disagree
This website has much that is of interest to me.	○	○	○	○	○
It is difficult to move around this website.	○	○	○	○	○
I can quickly find what I want on this website.	○	○	○	○	○
This website seems logical to me.	○	○	○	○	○
This website needs more introductory explanations.	○	○	○	○	○
The pages on this website are very attractive.	○	○	○	○	○
I feel in control when I'm using this website.	○	○	○	○	○
This website is too slow.	○	○	○	○	○
This website helps me find what I am looking for.	○	○	○	○	○
Learning to find my way around this website is a problem.	○	○	○	○	○

Figure 1.2. Some items from the WAMMI Questionnaire.

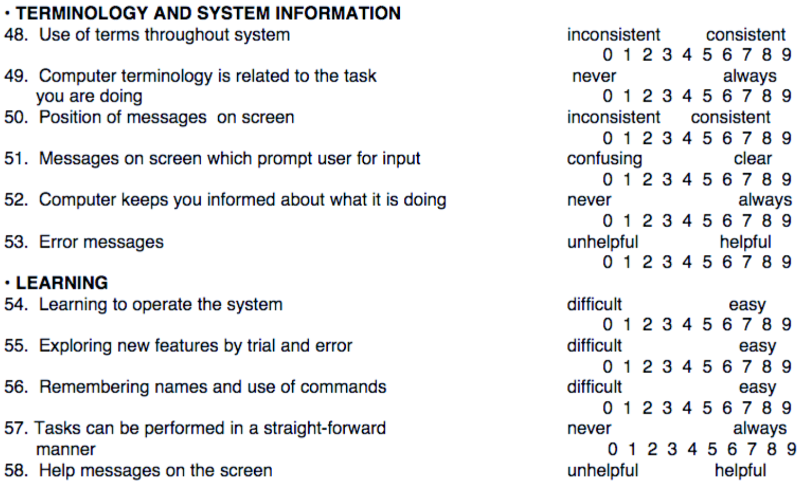
However, the data on the validity of SUMI and WAMMI give conflicting judgments. While they are considered the best validated tools available (Baber, 2002), there is a lack of comparative validation that demonstrates their real capacity for analysis (Annett, 2002). In fact, although the analyses conducted on

SUMI and WAMMI have demonstrated their effectiveness and efficiency (multidimensional aspects of usability), it is not entirely clear how the multidimensional metrics that constitute the sub-scale and global scale have been derived (Federici, Borsci & Meloni, 2009).

Questionnaire for User Interaction Satisfaction - QUIS

The Questionnaire for User Interaction Satisfaction - QUIS (Chin, Diehl & Norman, 1988), is a tool developed by the Human Computer-Interaction Laboratory - HCIL, at the College Park of the University of Maryland. The tool is based on the assumption that user satisfaction is a relevant indicator of system usability.

The QUIS is described as one of the most reliable and valid instruments to measure how users evaluate their interaction with a system (Chin, Diehl & Norman, 1988). This questionnaire mitigates the typical usability evaluation problems, which refer to validation, reliability and standardization (Ives, Olson, & Baroudi, 1983), providing a highly reliable measurement for many types of interfaces (Harper & Norman, 1993). The QUIS contains a section dedicated to the demographic data of the users, a section dedicated to the measurement of the overall satisfaction of the system and a hierarchically organized section of the measures related to specific elements of the interface. These measures vary depending on the version of the questionnaire. For example, QUIS 5.5 (Harper & Norman, 1993) has four dimensions or factors: screen factors, terminology and feedback feedback, learning factors, system capabilities (Chin, Diehl & Norman, 1988). The QUIS is currently at Version 7.0 with demographic questionnaire, a measure of overall system satisfaction along 6 scales, and measures of 9 specific interface factors. These 9 factors are: screen factors, terminology and system feedback, learning factors, system capabilities, technical manuals, on-line tutorials, multimedia, teleconferencing, and software installation. Users through the QUIS give a usability rating on a 9-point scale based on the semantic differential paradigm, indicating general satisfaction with each individual feature of a specific interface.



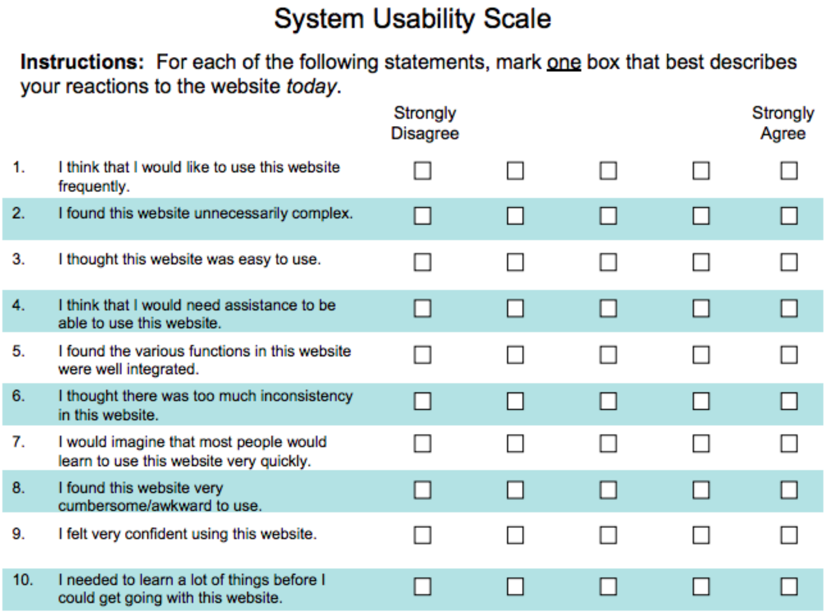
• TERMINOLOGY AND SYSTEM INFORMATION			
48. Use of terms throughout system	inconsistent	consistent	0 1 2 3 4 5 6 7 8 9
49. Computer terminology is related to the task you are doing	never	always	0 1 2 3 4 5 6 7 8 9
50. Position of messages on screen	inconsistent	consistent	0 1 2 3 4 5 6 7 8 9
51. Messages on screen which prompt user for input	confusing	clear	0 1 2 3 4 5 6 7 8 9
52. Computer keeps you informed about what it is doing	never	always	0 1 2 3 4 5 6 7 8 9
53. Error messages	unhelpful	helpful	0 1 2 3 4 5 6 7 8 9
• LEARNING			
54. Learning to operate the system	difficult	easy	0 1 2 3 4 5 6 7 8 9
55. Exploring new features by trial and error	difficult	easy	0 1 2 3 4 5 6 7 8 9
56. Remembering names and use of commands	difficult	easy	0 1 2 3 4 5 6 7 8 9
57. Tasks can be performed in a straight-forward manner	never	always	0 1 2 3 4 5 6 7 8 9
58. Help messages on the screen	unhelpful	helpful	

Figure 1.3. Some items from the QUIS Questionnaire.

System Usability Scale - SUS

The System Usability Scale - SUS (Brooke, 1996) is a 10 item questionnaire that provides a global and subjective assessment of usability.

This tool is described as “quick and dirty” because it was created without any evaluation of its validity and reliability. Nevertheless, the questionnaire has been widely used in the UX field and adapted to different contexts. The SUS is mentioned in over 600 technical-scientific publications and is one of the most robust and proven psychometric usability tools to date (Sauro, 2011a). Users are asked to answer each statement (e.g. “I found the website very easy to use”), providing a degree of agreement on a Likert scale from 1, “not at all agree”, to 5, “completely agree”. The scoring of the questionnaire returns a final score that can vary from 0 to 100, allowing researchers to obtain a one-dimensional measure of perceived usability (Brooke, 1996; Borsci et al., 2015).



System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	□	□	□	□	□
2.	I found this website unnecessarily complex.	□	□	□	□	□
3.	I thought this website was easy to use.	□	□	□	□	□
4.	I think that I would need assistance to be able to use this website.	□	□	□	□	□
5.	I found the various functions in this website were well integrated.	□	□	□	□	□
6.	I thought there was too much inconsistency in this website.	□	□	□	□	□
7.	I would imagine that most people would learn to use this website very quickly.	□	□	□	□	□
8.	I found this website very cumbersome/awkward to use.	□	□	□	□	□
9.	I felt very confident using this website.	□	□	□	□	□
10.	I needed to learn a lot of things before I could get going with this website.	□	□	□	□	□

Figure 1.4. The System Usability Scale.

The average value of a SUS questionnaire on more than 500 applications is 68, so scores lower or higher than 68 will indicate negative or positive variation from the reference average value. Bangor, Kortum and Miller (2008, 2009) compared the results of more than 3,500 studies using SUS. In particular, the authors examined the relationship between SUS scores and the evaluations of systems and products that users provided in terms of adjectives such as "good", "poor" or "excellent" and found a close correlation. According to Bangor, Kortum and Miller, it is possible to assign a grade score based on the SUS raw score.

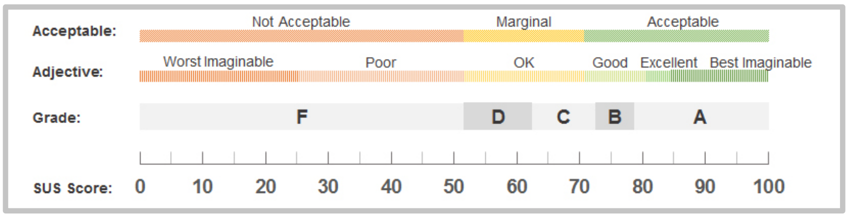


Figure 1.5. Grade rankings of SUS scores (from Bangor, Kortum & Miller, 2009).

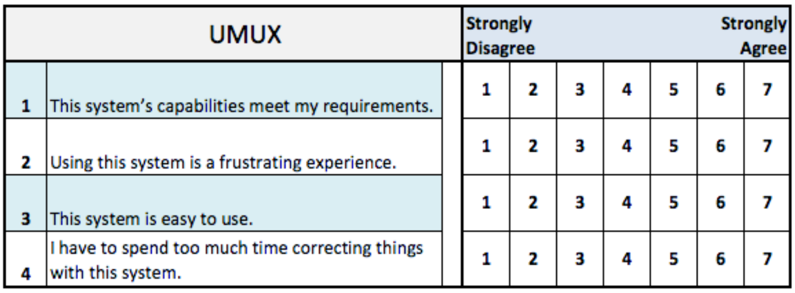
As reported by Borsci and colleagues (2015), the popularity of this tool within the HCI field is mainly due to three factors, including low usage costs, good psychometric properties - being a highly reliable and valid instrument - and fast administration time.

Although SUS was designed as a one-dimensional measure, several researchers have shown that its items could be divided into two dimensions: "usability" and "learnability" (Bangor, Kortum & Miller, 2008; Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009; Lewis, Utesch & Maher, 2013; Sauro and Lewis, 2012). However, this aspect has not yet been confirmed, requiring further studies to clearly identify the dimensional structure of the SUS and the variables that would contribute to its determination (Lewis, 2014).

Usability Metric for User Experience - UMUX / UMUX-LITE

Administration time is a crucial issue in matters of subjective questionnaires. Although the SUS is considered a "quick" questionnaire sometimes it is necessary to use even shorter scales. This further reduces the time and cost of research and is important when the subjective measure of usability is part of a broader investigation context (Lewis, 2014). The scales Usability Metric for User Experience - UMUX (Finstad, 2010; 2013) and UMUX-LITE (Lewis, Utesch, & Maher, 2013) were developed with this purpose.

Specifically, the UMUX scale consists of 4 items, while the UMUX-LITE scale consists of only 2 items. Users are asked to answer through a 7-point Likert scale, where 1 indicates lack of agreement and 7 total agreement with the proposed statement. The total scores obtained at both scales can vary from a minimum of 0 to a maximum of 100 points.

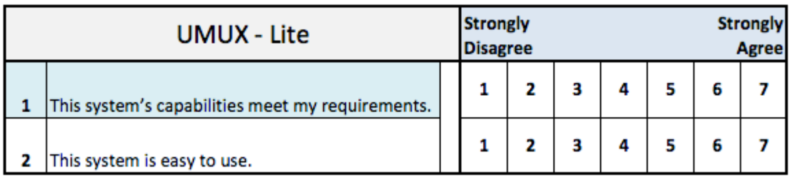


UMUX		Strongly Disagree						Strongly Agree
1	This system's capabilities meet my requirements.	1	2	3	4	5	6	7
2	Using this system is a frustrating experience.	1	2	3	4	5	6	7
3	This system is easy to use.	1	2	3	4	5	6	7
4	I have to spend too much time correcting things with this system.	1	2	3	4	5	6	7

Figure 1.6. The UMUX questionnaire.

For the calculation of scores in the UMUX to odd items one point is subtracted, while for even items the score provided by the user is subtracted from 7. The sum of the scores of the items thus obtained is then divided by 24 and multiplied by 100 (Finstad, 2010). With regard to the factorial structure of the UMUX, a study by Lewis, Utesch, & Maher (2015) identifies that the formulations of items in a positive or negative sense determine a two-factor structure, rather than a one-dimensional structure according to which the scale was designed. This phenomenon follows what was previously reported for the SUS questionnaire, although the UMUX also tends to be interpreted as a one-dimensional measure (Borsci et al., 2015).

For the scoring of the UMUX-LITE, instead, there is a different procedure: first one point is subtracted from the value expressed by the subject through the Likert scale; then the result obtained from the sum of the two items is divided by 12 and multiplied by 100 (Lewis, Utesch, & Maher, 2013). The validation of the UMUX-LITE (Lewis, Utesch, & Maher, 2015) has shown a high internal reliability, similar to that shown by the SUS (i.e., UMUX-LITE alpha = 0.86; SUS alpha = 0.91). In addition, the UMUX-LITE showed high correlation rates with both the SUS (r = 0.83) and the Net Promoter Score (r = 0.72). Finally, the regression equation showed how the UMUX-LITE scores can predict the SUS scores with an accuracy of 99% (Lewis et al., 2015; Borsci et al., 2015; Berkman & Karahoca, 2016).



	UMUX - Lite	Strongly Disagree						Strongly Agree
1	This system's capabilities meet my requirements.	1	2	3	4	5	6	7
2	This system is easy to use.	1	2	3	4	5	6	7

Figure 1.7. The UMUX-Lite questionnaire.

As reported by Borsci and colleagues (2015), the comparative analyses carried out for both questionnaires were conducted for the sole purpose of scale validation. Any effects due to the level of experience of the users were therefore not taken into account. The use of these tools would be particularly indicated in the preliminary design phases of a system/interface, in order to quickly test how users interact with the prototype. In conclusion, the UMUX-LITE is a promising tool both in terms of speed of administration and high effectiveness in the evaluation of the User Experience.

Net Promoter Score®

Developed by Reichheld (2003), the Net Promoter Score®[1] - NPS® is a widely

[1] Net Promoter® and NPS® are registered trademarks of Bain & Company, Inc., Satmetrix Systems, Inc., and Fred Reichheld.

[2] This study refers to the following publication: Serra, G., De Falco, F., Maggi, P., Forsi, R., Cocco, A., Gaudino, G., ... & Di Nocera, F. The role of mental workload in determining the

used indicator to assess customer loyalty and satisfaction. This tool consists of a single item: “How likely is it that you would recommend our company/product/service to a friend or colleague? “. Users can answer the questionnaire on a scale from 0 (“Not at all likely”) to 10 (“Extremely likely”). Depending on the score that is given to the Net Promoter question, three categories of people can be distinguished:

1. Promoters = respondents giving a 9 or 10 score;
2. Passives = respondents giving a 7 or 8 score;
3. Detractors = respondents giving a 0 to 6 score.

Promoters are satisfied users who, as a result of positive interactions, promote the brand or website with colleagues and friends; Passives can be considered satisfied but not enthusiastic users (score between 7 and 8); detractors are unhappy customers who can spread negative reviews of a company.

The final value of the NPS is obtained by subtracting the percentage of detractors from the percentage of promoters.

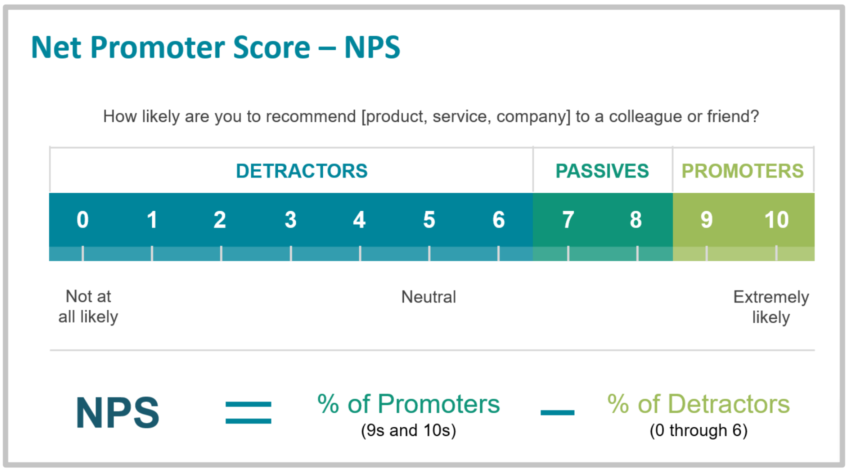


Figure 1.8. The Net Promoter Score questionnaire.

Several studies have found a strong correlation between perceived usability and NPS scores. For example, Sauro and Lewis (2010), in a study involving 146 subjects, identified a significant correlation (r = 0.61) between the NPS and the System Usability Scale (SUS). From this research, it emerges that promoters would obtain a higher average score in the SUS than detractors (Sauro & Lewis, 2010). Groth and Haslwanter (2015), in a study evaluating changes in usability between desktop and mobile versions of two different websites, found a strong correlation between the SUS and the NPS. They highlighted that participants who gave low usability scores through the SUS were more likely to be categorized as detractors.

The success of the NPS and its broad use could be attributed to its simplicity and its openly available methodology. These two attributes are extremely useful for companies, allowing them to reduce administration costs and to easily monitor perceived quality over time. On the other hand, the NPS is also advantageous for users. They gladly accept to answer only one question rather than long questionnaires. The use of the NPS is particularly suitable for e-commerce. In fact, it is easy to integrate a single statement into an online purchasing process, increasing response rates and user acceptance (Artz, 2017).

The NPS also has several limitations, mainly related to the measurement method. The first limitation refers to the use of an 11-range scale that excessively reduces the distance between judgment levels, especially if each level of the scale is not accurately described. This scale variability creates a greater risk of subjectivity in scoring. As a consequence, the instrument proposes a categorization of users in three clusters starting from the 11 step scale, of which only 2 points represent the promoters, 2 points represent the passives and 7 points represent the detractors. If, on the one hand, Reichheld (2003) states that, by limiting the "most enthusiastic" promoters to only 2 points, it was also possible to limit the "inflation" effect, so that even the passives would be evaluated as satisfied, on the other hand, Artz (2017) argues that this division would lead, in reality, to misleading and deceptive results, especially if only the detractors and promoters would be considered in the final evaluation. In support of this, Artz (2017) reports, in fact, that a company with 5% promoters, 90% passives and 5% detractors would have a final result of 0, while another company with 50% promoters, 0 passives and 50% detractors, would always have a final result of 0. In this way, therefore, although the conditions described are completely different and would require completely different business management, if we dwell on the NPS scores, we would observe the same value relative to the level of growth.

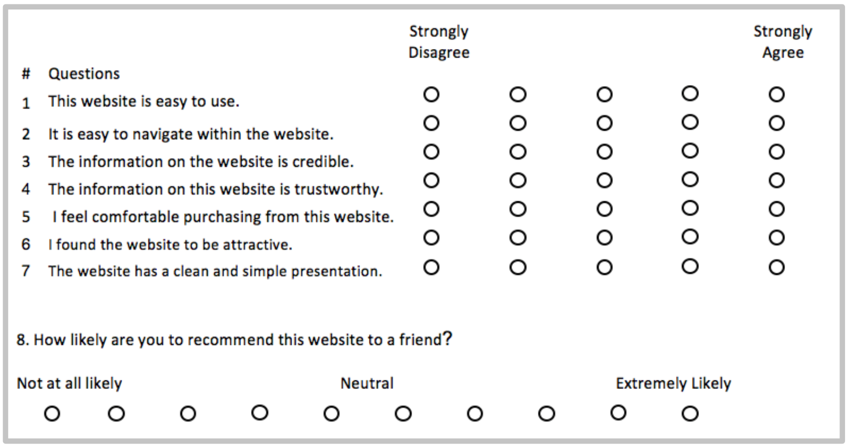
A further limit concerns the average score of the scale. In fact, although 5 should potentially indicate the average value of the scale that goes from 0 to 11, however, the score 6 is already considered as belonging to the negative judgments that contribute to determine the category of detractors. In this case, therefore, the user could be misled, because, believing to provide a judgment that tends towards a positive evaluation, he would find himself in a non-explicit way to provide, instead, a negative judgment.

In the final analysis, asking “How likely is it that you would recommend [...]” could lead to an excess of subjectivity by users. In a scale of this type, in fact, a criterion formulated in a generic way could lead to excessive interpretations by users. One solution might be to ask users “why” they have given a particular score. Again, in UX research it would be appropriate to anchor users' judgment to specific dimensions of the evaluated system/website.

Standardized User Experience Percentile Rank Questionnaire – SUPR-Q

The SUPR-Q questionnaire (Sauro, 2015) consists of 8 multiple-choice items, which assess the quality of a user's user experience with respect to a given website. The SUPR-Q provides both a measure of the overall quality of the interaction (similar to the satisfaction dimension) and a measure of specific aspects such as usability, credibility/trust, loyalty and aesthetics of the website. The SUPR-Q is a widely used questionnaire: it is estimated that more than 100 organizations in various sectors, from e-commerce to travel agencies and public services, use it to evaluate their websites.

The SUPR-Q can therefore be used both as part of a broader usability test and as a preliminary evaluation tool in the construction of a website.



#	Questions	Strongly Disagree				Strongly Agree
1	This website is easy to use.	○	○	○	○	○
2	It is easy to navigate within the website.	○	○	○	○	○
3	The information on the website is credible.	○	○	○	○	○
4	The information on this website is trustworthy.	○	○	○	○	○
5	I feel comfortable purchasing from this website.	○	○	○	○	○
6	I found the website to be attractive.	○	○	○	○	○
7	The website has a clean and simple presentation.	○	○	○	○	○

8. How likely are you to recommend this website to a friend?

Not at all likely				Neutral				Extremely Likely	
○	○	○	○	○	○	○	○	○	○

Figure 1.9. The SUPR-Q questionnaire.

The SUPR-Q scores indicate the level of usability of the evaluated website: low scores suggest substantial changes to the website and the possible need to conduct a broader and more structured usability test. In contrast, high scores indicate a good level of usability. For the first 7 items an answer is given on a 5-point Likert scale (where 1 corresponds to "not at all agree" and 5 to "totally agree"). The last item, instead, follows the single item of the NPS ("How likely would you recommend this website to a friend or colleague?"), with 11 answer options ranging from 0 to 10.

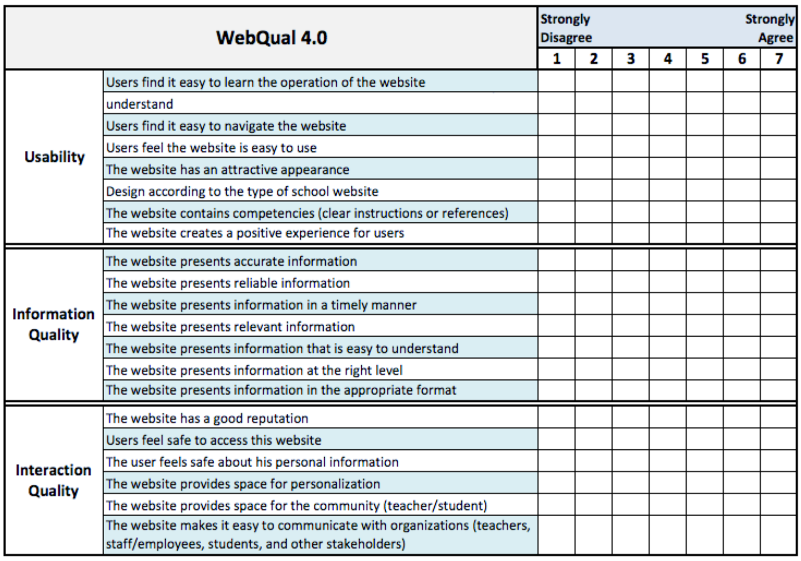
The SUPR-Q returns scores in percentile ranks, it follows that a score higher than 75% would indicate that the total score is higher than 75% of websites. It is also important to compare each scale of the questionnaire with the percentile ranks obtained, as the same website may show higher scores in one dimension and lower scores in the other.

The SUPR-Q shows high internal reliability (α= 0.86) and good validity. The overall score to the questionnaire has, in fact, shown significant correlations with both the SUS (r= 0.87) and the WAMMI (r= 0.88). Some SUPR-Q limitations regarding the pool of the data (the dataset of organisations has a North American bias) and the absence of qualitative insights (the proverbial "why").

WebQual 4.0

The WebQual questionnaire was developed by Barnes and Vidgen (2001; 2002; 2005), in order to assess the usability and quality of websites and, specifically, e-commerce websites. The WebQual 4.0 is based on quality function deployment (QFD), i.e. "a structured and disciplined process that provides a means to identify and transport the customer's voice at each stage of product and/or service development and implementation" (Slabey, 1990). For this reason, the authors describe WebQual as a tool able to "capture the voice of users" (Barnes & Vidgen, 2002).

The WebQual 4.0 version is composed of 22 items, plus item 23 related to the overall evaluation of the website, to which users provide a score on a 7-point Likert scale (where 1 indicates "Strongly disagree" and 7 "Strongly agree"). The 22 items refer to the following 3 dimensions: (a) usability, which indicates how easily the user can learn and interact with the content of the website; (b) quality of information, which indicates the degree of relevance or detail of the information on the website; (c) quality of interaction, measured by the level of user satisfaction with the website.



WebQual 4.0		Strongly Disagree						Strongly Agree
		1	2	3	4	5	6	7
Usability	Users find it easy to learn the operation of the website							
	understand							
	Users find it easy to navigate the website							
	Users feel the website is easy to use							
	The website has an attractive appearance							
	Design according to the type of school website							
	The website contains competencies (clear instructions or references)							
	The website creates a positive experience for users							
Information Quality	The website presents accurate information							
	The website presents reliable information							
	The website presents information in a timely manner							
	The website presents relevant information							
	The website presents information that is easy to understand							
	The website presents information at the right level							
	The website presents information in the appropriate format							
Interaction Quality	The website has a good reputation							
	Users feel safe to access this website							
	The user feels safe about his personal information							
	The website provides space for personalization							
	The website provides space for the community (teacher/student)							
	The website makes it easy to communicate with organizations (teachers, staff/employees, students, and other stakeholders)							

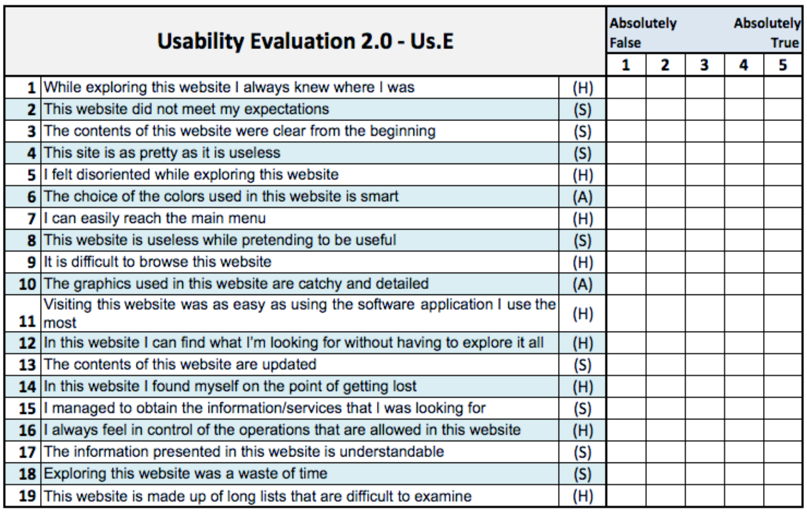
Figure 1.10. The SUPR-Q questionnaire.

WebQual 4.0 is currently very popular (Ahmad & Khan, 2017). However, some authors (Chen & Chang, 2010) have pointed out that the questionnaire gives general information about the user experience, rather than specific information about the specific dimensions of website usability.

Usability System Evaluation 2.0 - Us.E. 2.0

Usability Evaluation 2.0 (Us.E. - Di Nocera, 2013) is a multidimensional questionnaire for the assessment of website usability, intended as the quality of the interaction experienced by a user who visits a website to achieve a goal (e.g., searching information, purchasing a product, forum compilations).

The questionnaire was officially presented during the 1° Italian Day on Human-Computer Interaction (Di Nocera, Ferlazzo & Renzi, 1999). The first version of the questionnaire (Us.E. 1.0) contained numerous sentences gathered through users interviews on their experience with various websites. This version contained 70 items. The research group of Di Nocera and colleagues (Di Nocera, Ferlazzo & Renzi, 2003), through factorial analysis found the existence of a four-factors structure: i) “Handling”; ii) “Satisfaction”; iii) “Attractiveness” and iv) “Predictability”. Further investigations suggested a more economical solution with three factors, where the “Predictability” items converged in the “Handling” dimension. Initially, this version was disseminated with a first standardization norm set, and it was used for professional and research purposes. The high number of items was a critical aspect of Us.E. 1.0 version, making it difficult for online administration (Boscarol, 2003). The current version of the questionnaire, Us.E. 2.0, obtained by reducing the number of items through factorial analysis, is composed of 19 items. The Us.E. 2.0 requires users to express a judgment with respect to the three dimensions of Handling, Satisfaction and Attractiveness, through a 5-point Likert scale (from 1= “absolutely false” to 5= “absolutely true”).



Usability Evaluation 2.0 - Us.E			Absolutely False				Absolutely True
			1	2	3	4	5
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 colors 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)					

Figure 1.11. The Us.E. 2.0 questionnaire.

- *H Scale – Handling:* This dimension refers to the interactions with the website structure -i.e., the architecture information, the pages hierarchy and the menu options. This scale expresses the ease with which the user can reach a specific goal. Eleven items define this dimension: four refer to the absence of handling (they classify the H dimension negatively), and, for this reason, the scores must be inverted. Low scores suggest the need to modify some aspects of the website structure, like the menu, the contents hierarchy and the elements position in the website pages.
- *S Scale – Satisfaction:* This dimension deals with the “perceived utility” and it must not be confused with “customer satisfaction”, a notion used in the marketing field. They are two different concepts: customer satisfaction indicates the satisfied users' proportion with a product or a service. Instead, the term “satisfaction”, adopted in the usability domain, refers to the users' perceived utility of a website. Satisfaction is intrinsically linked to the user's goals achievement or to the problems solution search that led the

visitor to a specific website. Six items define this dimension: three refer to the absence of Satisfaction (they define the dimension negatively), and, for this reason, the scores must be inverted. Low scores could suggest the website was built without considering the users characteristics. The reasons behind that could be at least two: (a) the website visitors don't correspond to the designer's expectations, and (b) the contents don't match the users' needs.

- *A Scale – Attractiveness:* This dimension refers to the user's appreciation of some aesthetic characteristics (colour and composition) considered to obtain indirect information on the website design accuracy. This information can convey usability even before interacting with the website (the so-called “apparent usability”: Kurosu and Kashimura, 1995). The items that contribute to this scale are only two, and both define the dimension positively. Low scores may indicate the need for a redesign and/or for a more considerable attention to the use of colours, images and care for details.

The “raw scores” obtained by the administration are standardized in “z scores” according to calibration standards (average and standard deviation of previous assessments) divided into four types of websites (Portals and Communities, Universities, Authorities and Public Administrations, Companies and Services). The administration of the test requires a sample of at least 30 subjects. The scores obtained vary within a range from +1 to -1, where 0 indicates the average value, i.e. a score exactly in the norm. On the contrary, values below and above zero indicate a website evaluation, respectively, more negative or more positive than websites in the same category. Over time, Us.E. was administered to a large number of users to evaluate several types of websites. To calculate the standardized score three kinds of information are needed: i) the score obtained from the evaluation carried out; ii) the average score collected in the previous assessments; iii) the standard deviation of the scores obtained in the earlier estimates. These last two pieces of information constitute the calibration standards. The calibration standards are divided into four types of websites: i) Portals and Communities; ii) Universities; iii) Public Authority and Public Administrations; iv) Companies and Services.

Furthermore, for each category, normative values have been identified separately for males and females. Usability assessment through Us.E. is affected by gender differences. Therefore, for each single user evaluation, it is necessary to compare the scores to the specific rules for the gender.

1.1.5. Behavioural measures

Usability could also be evaluated through behavioural measures, i.e. measures collected while the user is performing a certain activity. The HCI community agrees that poor usability affects the user's performance. The most commonly used behavioural metrics in usability assessments refer to the "task success rate", "error rate" and "task completion time" (Albert & Tullis, 2013; Hornbæk, 2006; Esmeria & Seva, 2017).

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).

Binary success is a special type of discrete data (i.e. success – complete task, failure – did not complete). This metric is appropriate when the achievement of a user's goal depends on the accomplishment of a task or series of tasks. More specifically, in the case of binary success, the users can obtain a "success" or "failure" score. There is no middle ground. Typically, these scores are in the form of 1's for success and 0's for failure. Binary success rates are often analyzed and presented by task. For example, if 5 out of 10 users complete a task, the completion rate is .50 or 50% when expressed as a percentage. This means that it is possible to draw conclusions about the usability of a system depending on the percentage of participants who completed each task successfully (Nielsen, 2001). Sauro (2011b) analysed success rate data from 115 usability tests (1189 tasks taken from both lab-based and unmoderated usability tests). By observing the cumulative distribution of this data set, he found that a success rate of 78% corresponds to the 50th percentile, so 78% is an average success rate.

On the other hand, considering "levels of success" is useful when the user receives some value from completing a task partially. In order to collect data on success levels it is necessary to define what “complete success” and “complete failure” mean. This allows researchers to break out different success levels of completion. For example:

- Complete success:
  - With assistance;
  - Without assistance.
- Partial success:
  - With assistance;
  - Without assistance.
- Failure:
  - User thought it was complete, but it wasn’t;
  - User abandoned.

The use of this particular metric is useful to highlight design interventions useful to improve the usability of the system even if the user has completed the assigned task (Nielsen, 2001).

Error Rates

Errors are a useful metric to evaluate user performance. First of all, it is crucial to distinguish errors by usability issues. While a usability issue is the cause of a problem, an error is a possible outcome of an issue. For example, if users encounter a problem downloading a document to a Public Administration website, the problem could be incorrect menu labelling. The error, or the result of the issue, could be the act of choosing the wrong pages when searching for the document. In summary, errors are incorrect actions that may lead to task failure. Each error is considered as a defect and error counts are considered as discrete data. The total opportunities for defects are calculated by multiplying the total number of participants with the number of tasks. The analysis of errors is useful to understand how many errors were made, where they were made, and, in general, how much a system is usable (Albert & Tullis, 2013). Measuring errors is useful in different situations, for example:

1. When an error leads to a loss of efficiency - for example, when an error leads to a waste of time or requires the user to do the same activities

several times;

2. When an error results in significant costs to the organisation or end-user - for example, if an error results in an increase in complaints to customer care operators;
3. When an error results in the failure of a task - for example if an error causes a user to download the wrong document, or to make a wrong payment.

When using this metric it is often necessary to investigate and understand why certain errors occur. For this purpose, it is useful to examine in detail and code each type of error. An error encoding could include, for example, “typing errors”, “interpretation errors”, “selection errors”, and so on. Error coding allows usability experts to count the frequency of errors and to understand their impact on the user’s goals.

Task completion time

Task completion time (or task on time) is usually used to measure the efficiency of a system (Nielsen, 2001). This metric considers the amount of time the user needs to complete the task. It can be expressed in minutes or seconds. The most common way to report this type of data is to present the average time spent on each task, or to perform a series of tasks. Generally, the smaller the task completion time, the better the usability. However, there are some exceptions to the assumption that faster is better. For example, in the field of e-learning, users can learn more if they spend more time completing tasks rather than performing tasks too quickly. Task completion time is particularly important for the evaluation of systems that require the user to repeat the same tasks several times. For example, for websites that require the user to purchase a train ticket or book a hotel room, the time needed to complete these tasks is an important measure of efficiency.

UX researchers (Albert & Tullis, 2013; Hornbæk, 2006; Nielsen, 2001) distinguish between task completion time of only successful tasks and task completion time of all tasks. Including only successful tasks allows for a cleaner measure of efficiency. In fact, the inclusion of tasks in which the participant was unsuccessful increases the variability of the data and complicates its interpretation.

On the other hand, including time data for all tasks, successful or unsuccessful, reflects more accurately the overall user experience. Moreover, it is a measure independent of the success rate (Albert & Tullis, 2013).

1.1.6. Physiological measures

Recently, the use of physiological metrics has also been extended to UX (Cowley et al., 2016; Ganglbauer et al., 2009). Variations in physiological indices, such as galvanic skin response (GSR), breathing, heart rate (HR) and blood volume pulse (BVP) are associated with task difficulty, attention levels, frustration experiences and emotionally toned stimuli (Andreassi, 2013; Cowley et al., 2016). Thanks to this type of metrics, UX researchers can obtain real-time feedback without interfering with the performance of the user. Moreover, the use of physiological measures can help to reduce social desirability and avoid distortions related to subjective metrics.

Skin Conductance

Skin conductance (or Electrodermal activity - EDA; or galvanic skin response - GSR) has been used extensively as an indicator of UX experience (Boucsein, 2012; Wilson & Sasse, 2000; Wilson, 2001; Ward et al., 2002; Ward & Marsden, 2003).

Variations in EDA are associated with the user's emotional state as stress, excitement or frustration (Lang et al., 1993). EDA data are usually collected from the fingers, wrists, or hand palms.

Several studies investigated the relation between skin conductance and usability. Wilson and Sasse (2000) used GSR measures to evaluate subject responses to audio and video degradations. Significant increases in GSR were found for poor quality videos, even though most subjects didn't report noticing differences in media quality.

Ward and Marsden (2003) compared EDA of users while browsing a well-designed and a poorly designed website. The authors found a decrease in skin conductance for the well-designed website. For the poorly designed site, skin conductance data showed an increase over the first minutes of the session indicating a high level of stress.

Trimmel, Meixner-Pendleton, and Haring (2003) used skin conductance to assess the level of stress induced by the load time of different websites. They found significant increases in skin conductance as page load time increased.

Despite these encouraging results skin conductance measures show the intensity of arousal, but not its valence. Therefore it is important to integrate these measures with other explicit measures to better understand the user experience.

Cardiovascular measures

Measures of cardiovascular activity include HR, heart rate variability (HRV), blood pressure (BP), blood volume pulse (BVP) and electrocardiogram (EKG). Several authors (Winton et al., 1984; Papillo & Shapiro, 1990) use HR to differentiate between positive and negative emotions.

Wilson and Sasse (2000) found an increase in HR and a decrease in BVP related to a degradation of the quality of a video conferencing software.

Similar results were found by Ward et al. (2003) in a study that compared the cardiac measure of users while browsing well and poorly designed websites. Participants showed a high level of stress while browsing the poorly designed website, indicated by increases in HR.

Other studies (Drachen et al., 2010; Yannakakis et al., 2008) in the video-games field found that HR is associated with self-report measures of player experience, both positive and negative. Specifically, Drachen and colleagues found that a higher average HR is correlated with players frustration, while a low HR average indicates positive affect.

Electroencephalography

Like other physiological measures, electroencephalography (EEG) allows UX researchers to draw conclusions about the emotional states of users in real time (do Amaral et al., 2013, Tatum, 2014; Van Camp et al., 2018). Generally, to record EEG data, some electrodes are placed on the user's head to detect and capture brain signals from the underlying cortical regions. The frequencies are then analysed with sophisticated algorithms that allow researchers to identify the user's cognitive and emotional activity. Today, EEG is increasingly used by UX researchers to investigate whether a user is frustrated or happy while using a specific product or interface. The latest commercial EEG analysis software analyzes the user's brain activity according to different electrical frequency bands. Researchers (do Amaral et al., 2013, Tatum, 2014; Kimura et al., 2009; Van Camp et al., 2018) distinguish four primary bands: *i) delta waves* (1-4 Hz) generally occur during deep sleep; *ii) theta waves* (4-8Hz) are associated with creativity, daydreaming and emotions; *iii) alpha waves* (8-14Hz) indicate low brain activity and relaxation; *iiii) beta waves* (10-30Hz) indicate cognitive processes such as problem solving and information processing.

Van Camp and collaborators (Van Camp et al., 2018) analysed EEG of 8 participants while watching three different types of videos: one inducing a high positive emotional response (i.e. enjoyment), one inducing a high negative emotional response (disgust) and one inducing a neutral or low emotional response. The authors found differences in the participants' EEG between the positive and negative condition, demonstrating the usefulness of EEG analysis within user experience research.

Another study conducted by Nacke (2010) in the entertainment sector compares the affective gameplay interaction modes between two different video games consoles while gamers played a horror game. The results indicated a significant positive correlation between alpha waves and negative affect ratings for both console types.

Similarly, the work of do Amaral and colleagues (do Amaral et al., 2013) highlighted differences in the EEG of participants while performing "easy" or "difficult" tasks on a social network. Results like these suggest that EEG could also be a valid metric in usability studies allowing researchers to measure users' emotions and cognitive processes in real-time.

Facial Emotion Recognition

In recent years there has been a growing interest in the analysis of non-verbal communication (Knapp, Hall, & Horgan, 2013; Mandal, 2014; Witkowski, 2012). In particular, in the UX field, the recognition and coding of the user's emotions starting from an analysis of his facial expressions has acquired more and more interest (Dubey and Singh, 2016; Winton et al., 1984). The pioneering studies of Ekman and Friesen (1976) highlighted how basic emotions (happiness, sadness, surprise, fear, disgust and anger) are recognizable by all human beings through the observation of facial expression, regardless of their cultural background.

Several companies have developed software that analyses a user's emotions by examining his facial reactions captured through a webcam. These systems analyse basic facial features as eyebrow, mouth, nose and eye frame by frame and, through algorithms, provide as output the type of emotion the user is experiencing at that particular moment.

Although this technique would seem promising, for the moment the results obtained are still weak and show a lot of individual variability (Staiano et al., 2012; Terzis, Moridis & Economides, 2010; Zaman & Shrimpton-Smith, 2006). Moreover, emotions can occur with or without facial expressions and vice versa (Russell, 1995). For these reasons, facial recognition software should not be used in isolation.

2. Mental workload

The concept of mental workload (MWL) historically focused on the human-machine interaction in “high complexity” work environments (e.g: aviation, aerospace, healthcare, etc.) where specific skills and knowledge are required to use certain interfaces (De Waard, 1996; Hart & Staveland, 1988; Hart, 2006).

MWL deals with psychological and physiological aspects of human capability, able to influence the interaction between the operator, the systems and the procedures of his working environment, and to directly affect the outcome of events. MWL has often been used in an attempt to explain the differences that occur between the performance of individuals with the same skills and abilities. The term is usually used to indicate the expenditure of “cognitive resources” and is therefore aligned with attentional theories that assume the existence of one (Kahneman, 1973) or more (Wickens, 1984, 2002, 2008) limited resources that must be used to perform certain tasks by the individual. An explanation for this phenomenon can be found in Wickens' “theory of multiple resources” (2008). This theory is based on two main assumptions: the first is that each individual during the execution of a task has limited resources available, the second, instead, that different tasks require different amounts and types of resources. MWL is a cross-disciplinary phenomenon, for this reason there is no univocal definition of MWL in the literature. However, despite the divergence of views on the nature and definition of its construct, MWL is a crucial and measurable phenomenon in HCI. One of the most important reasons behind the study of MWL is to quantify the mental resources needed to perform an activity in order to predict, and possibly integrate through automation systems, the operator’ performance (Cain, 2007).

Generally, MWL is defined as “the difference between the demands imposed by the task on the subject and its resources available to perform the task” (O' Donnel & Eggemeier, 1986). MWL is determined by different factors as task demands (when the difficulty, number, frequency, or complexity of demands increases, MWL increases), the level of performance of the operator (when the number of errors increases, or when the accuracy of the control exercised decreases, MWL increases) the operator's effort to perform the task (in this case MWL reflects the operator's response to the task, rather than the demands imposed) and the operator’s perceptions of his amount of effort (when an operator feels under severe stress, his MWL may increase even if the requirements of the task have not changed). Therefore MWL is considered a multidimensional construct (Kramer, 1991; Leplat, 1978; Moray, 1979). It represents the individual level of attentive involvement and mental effort (Wickens, 1984) integrating both the objective difficulties of the task and the effort (physical and mental) that the operator experiences (Gopher & Donchin, 1986).

Several studies (Lysaght et al., 1989; Young and Stanton, 2002) observed that too high or too low levels of mental workload can adversely affect user performance and efficiency (Xie and Salvendy, 2000). In the literature (Xie, Salvendy, 2000) there is also a distinction between an “effective” mental workload, i.e. mainly due to the structure of the task and the demands it imposes on the operator (it cannot be avoided), and an “ineffective” mental workload, due to the characteristics of the individual and his skills (it can be reduced). An important implication of this classification is that the mental workload can be reduced by controlling those factors that contribute to the ineffective workload. Controlling these factors, starting at the design stage of a system or interface, is fundamental to optimize the levels of safety, productivity, satisfaction and user involvement.

2.1. Measuring mental workload

There are different techniques to evaluate the mental workload. The heterogeneity of these techniques is due both to the variety of fields of application of the mental workload and to its multidimensionality. A common element between these different techniques is that they evaluate MWL indirectly, i.e. through the analysis of variables related to it.

Typically, mental workload measures are divided into subjective, behavioural and physiological measures (O' Donnel & Eggemeier, 1986; Wierwille & Eggemeier, 1993). Behavioural measures derive from indices recorded during the performance of a task (number of errors, reaction time, etc.); subjective measures refer to the impressions reported by users during or at the end of the task and investigated through specific questionnaires (e.g. NASA-TLX, SWAT, etc.); finally, physiological measures are based on the analysis of changes in physiological indices such as heart rate, breath rate, event-related potentials or eye movements.

The choice of measurements must take into account different elements such as the research objectives, the type of task assigned to the user and the context in which the measurements take place. Some authors (Eggemeier et al., 1991) stress the importance of aspects related to the validity and reliability of the measurements used. In fact, these must allow us to discriminate between the different types of demands (physical, mental, etc.) that the task imposes on the individual.

The most important properties that a MWL measure must have are mainly three, namely: “sensitivity”, i.e. the ability to discriminate between different levels of mental workload; “diagnosticity”, in order to differentiate between different types of MWL, in terms of the type of demand (physical, cognitive, etc..) imposed on the individual by the task; and, finally, “intrusiveness”, i.e. the measure must not be obtained in an invasive manner and must not interfere with the performance of the primary task. In addition, it is important that the measures chosen are able to provide real-time estimates, so as to constantly monitor the levels of mental workload experienced by the individual.

2.1.1 Subjective measures of mental workload

These measures are based on the perceptions and experiences reported by the operator during or at the end of a task. Specifically, through the administration of questionnaires or interviews, the individual is asked to answer questions about the degree of fatigue experienced in relation to different dimensions related to the mental workload. The advantages related to the use of these tools mainly concern the ease and cost-effectiveness of administration, as they allow to obtain indications on the mental workload quickly and without the need for sophisticated analysis or equipment. Moreover, thanks to subjective measures it is possible to obtain an evaluation of the subjectively perceived workload. The subjective perceptions of MWL are very important and can not be investigated with other tools except through interviews and self-report questionnaires.

Hart and Wickens (1990) propose a classification of MWL subjective measurements based on three categories: one-dimensional measurements, hierarchical measurements and multidimensional measurements. These self-assessment techniques differ in the complexity of administration and scoring of results. Multidimensional scales explicitly represent the dimensions of MWL and allow to obtain an evaluation for each dimension, such as NASA-TLX (Hart & Staveland, 1988), or the Subjective Workload Assessment questionnaire (SWAT; Reid & Nygren, 1988). These scales provide, in addition to an overall score, a score related to the individual dimensions of MWL.

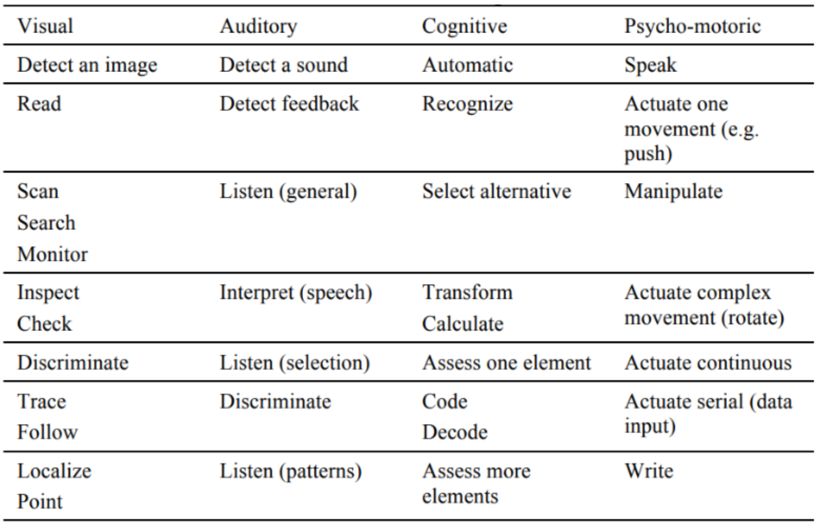
According to Nygren (1991) it is preferable to use multidimensional scales as they have a greater diagnostic capacity and therefore greater sensitivity. One limitation of these instruments is that they take a long time to be correctly administered. Hendy, Hamilton and Landry (1993) claim that one-dimensional scales are better than multidimensional scales in providing an overall assessment of MWL. One-dimensional scales are also faster and easier to administer. The short administration time allows a real-time assessment of MWL and, in addition, does not require the operator to remember his past MWL. For these reasons, one-dimensional scales are often preferred in operational contexts where a real-time MWL evaluation is required. Examples include Bedford Scale (Roscoe, 1984), Modified Cooper Harper (MCH; Wierwille & Cascali, 1983), Instantaneous Self Assessment (ISA; Brennan, 1992), and the US Air Force Flight Test Centre (AFFTC; Ames & George, 1993). Their ease of use favors a high level of acceptance by operators who are willing to respond to the request for self-assessment several times during the execution of the task.

2.1.1.2 Multidimensional measures

Visual, Auditory, Cognitive, Psychomotor method - VACP

This method has been developed since 1984 by Aldrich and McCracken (Aldrich, Szabo & Bierbaum, 1989; McCracken & Aldrich 1984) based on Wickens' resource theory (1984). Subjects evaluate the demands of the task by referring to four dimensions: visual, auditory, cognitive and psychomotor (Visual, Auditory, Cognitive, Psychomotor method - VACP). The assessment involves a response on a 7-point scale that uses verbal descriptors to simplify the self-assessment and increase consistency between the responses of various operators or between the various activities (Figure 2.1). The descriptor formulation is designed to be adaptable in various contexts. However, the responses obtained must be analyzed taking into account the nature of the task, the individual performing the task and the relationship between the individual and the task. For example, the cognitive demand for a driving task for an inexperienced pilot may be very high and the subject may choose the maximum descriptor for the scale. Conversely, a person who has been driving for a long time and who is following a familiar route (e.g., daily driving to work) may experience lower cognitive demands and choose the minimum descriptor level. An effect similar to the "context" effect can therefore influence the response. The investigator must therefore be very careful in interpreting the result. Also in relation to the use of websites or digital interfaces, for example, the type of assessment may vary depending on familiarity with the use of a specific website or device, requiring careful reflection in order to provide a reliable result.

This method has proven useful in providing guidance during system design, indicating possible critical issues for users.



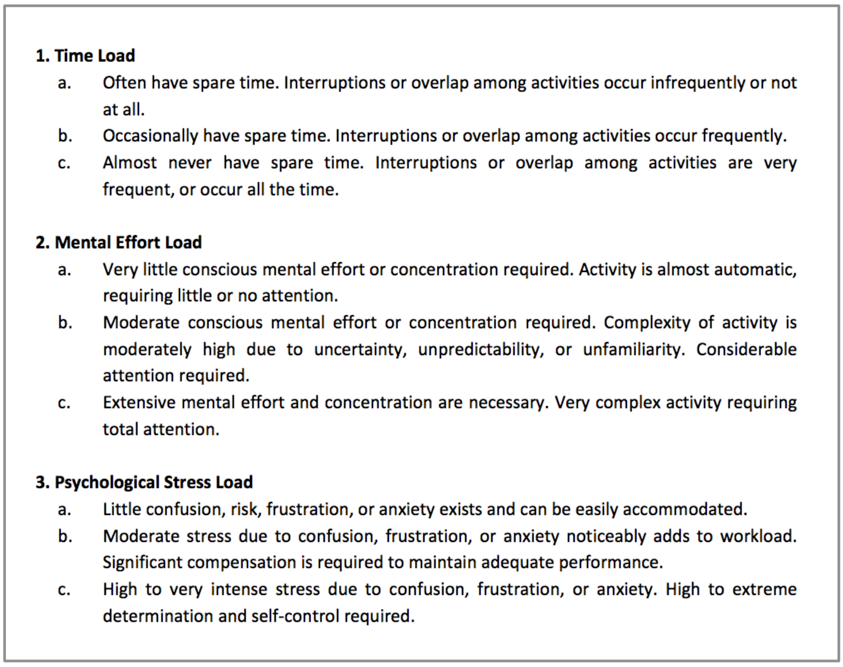
Visual	Auditory	Cognitive	Psycho-motoric
Detect an image	Detect a sound	Automatic	Speak
Read	Detect feedback	Recognize	Actuate one movement (e.g. push)
Scan Search Monitor	Listen (general)	Select alternative	Manipulate
Inspect Check	Interpret (speech)	Transform Calculate	Actuate complex movement (rotate)
Discriminate	Listen (selection)	Assess one element	Actuate continuous
Trace Follow	Discriminate	Code Decode	Actuate serial (data input)
Localize Point	Listen (patterns)	Assess more elements	Write

Figure 2.1. Descriptors of the four dimensions used in the VACP method.

Subjective Workload Assessment Technique - SWAT

The Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988) provides a three dimension estimate of MWL:

1. *The time load:* it refers to the time available to the operator and the percentage of time the operator is busy performing the task;
2. *The mental effort load*: it refers to the amount of resources spent on the task;
3. *The psychological stress load:* it concerns the levels of confusion, frustration and/or anxiety that affect the operator's workload.



1. Time Load

a. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.

b. Occasionally have spare time. Interruptions or overlap among activities occur frequently.

c. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

2. Mental Effort Load

a. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.

b. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.

c. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

3. Psychological Stress Load

a. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.

b. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.

c. High to very intense stress due to confusion, frustration, or anxiety. High to extreme determination and self-control required.

Figure 2.2. Swat sub-scales and descriptors.

Each dimension is described on the basis of three statements that correspond to a “low”, “medium” or “high” level of MWL on that dimension.

The application of the SWAT includes a card sorting task before the experimental task assignment. At this stage, the nine statements referring to the three dimensions are presented to participants in the form of 27 cards. Each participant should order these 27 cards to reflect his or her perception of an increase in MWL. The objective of this phase is to obtain a hierarchy between the three scales of Time Load (T), Mental Effort (E) and Stress (S). For example, TES is the order when there is a higher emphasis on the T scale, lower emphasis on E and lower emphasis on S. Then, during the experimental task, the subject is asked to provide a discrete evaluation of his MWL level with reference to the three statements representative of the three dimensions of MWL. This assessment can be made at the end of the activity or one of its segments. Finally, each three-dimensional assessment is converted into numerical scores between 0 and 100 using the interval scale developed in the first phase.

Although the SWAT has been tested on several occasions and is widely used, studies such as that of Hart and Staveland (1988) and Hill and collaborators (Hill et al., 1992) have shown that the NASA-TLX is superior to the SWAT in terms of sensitivity, especially when the MWL level is very low (Nygren 1991).

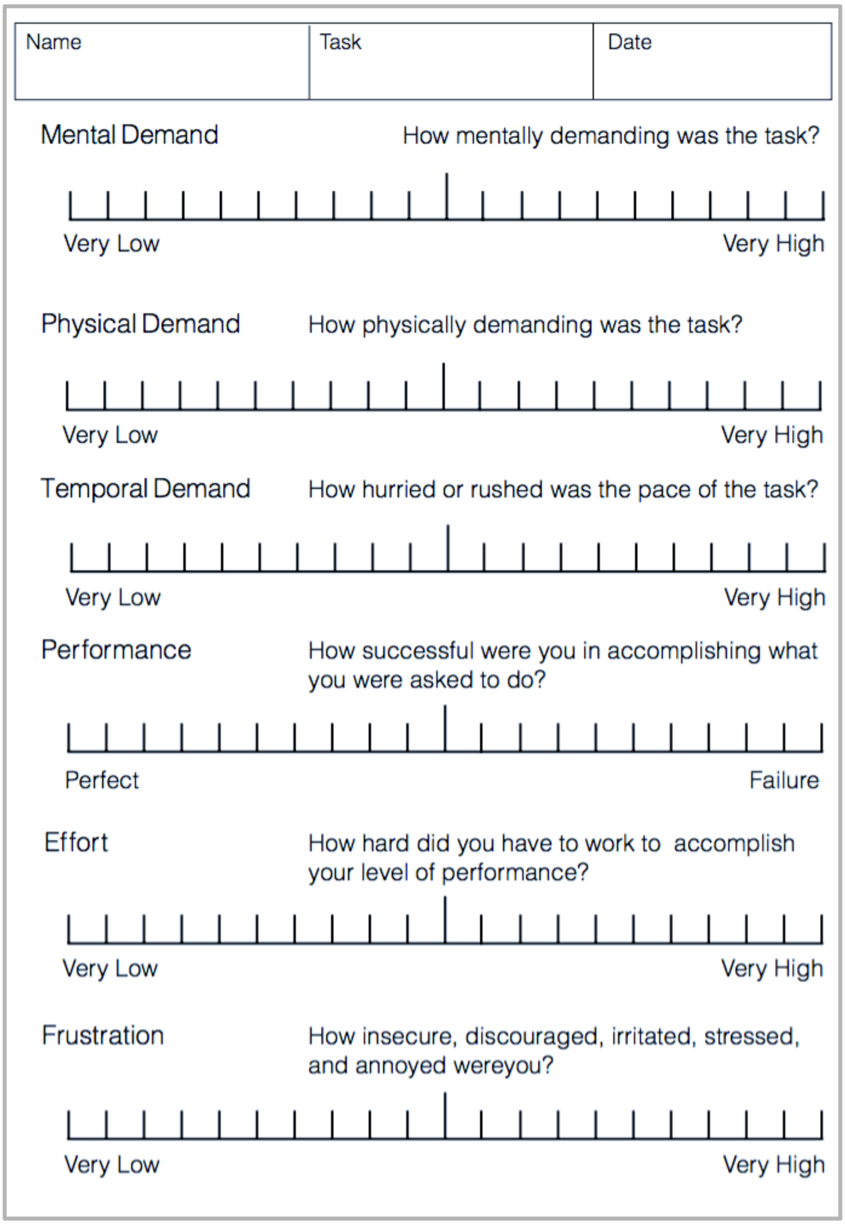
In an attempt to increase the scale sensitivity and to reduce the completion time of the card sorting phase, Luximon and Goonetilleke (2001) proposed some SWAT variants. These variants differ from the original version of the SWAT with respect to the procedure for the weight to be attributed to each scale. This procedure consists of a comparison in pairs between the scales. Luximon and Goonetilleke (2001) test variations that provide a discrete or continuous score in a study in which the participants have to perform arithmetic tasks with a different level of difficulty. Considering emerged results, the researchers suggest adopting a continuous perceived mental workload score, rather than a discrete score. This change was found to provide a more sensitive result than the same assessment made with a discrete score. Another contribution of this study is the suggestion to increase the number of response levels available in discrete scoring to increase the sensitivity of the scale.

National Aeronautics and Space Administration Task Load Index - NASA-TLX

The NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988) is a multidimensional assessment tool that rates perceived workload in order to assess a task, system, or other aspects of performance. NASA-TLX provides both an overall MWL score and a detailed score referring to six sub-dimensional dimensions: mental demand, physical demand, time demand, effort, performance and frustration level.

NASA-TLX was developed in an extensive laboratory research program by Hart and Staveland (1988). Its ability to discriminate mental workload has been demonstrated using a wide variety of tasks and in different operational contexts, such as in flight simulations (Battiste & Bortolussi, 1988; Corwin et al., 1989; Nataupsky & Abbott, 1987), in real contexts (Shively et al., 1987), in air combat (Hill et al., 1988) and using remotely controlled vehicles (Byers et al., 1988). Sawin and Scerbo (1995) used NASA-TLX to analyze the effects of instruction type and readiness for boredom on the performance of supervisory tasks.

Its administration consists of two phases. In the first, respondents are asked to provide an estimate of the perceived mental workload on a scale from 0 to 100, for each response scale. The second phase consists of multiple comparisons between the individual matched scales in order to produce a weighted score. Finally, an algorithm allows us to obtain an overall score.



Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

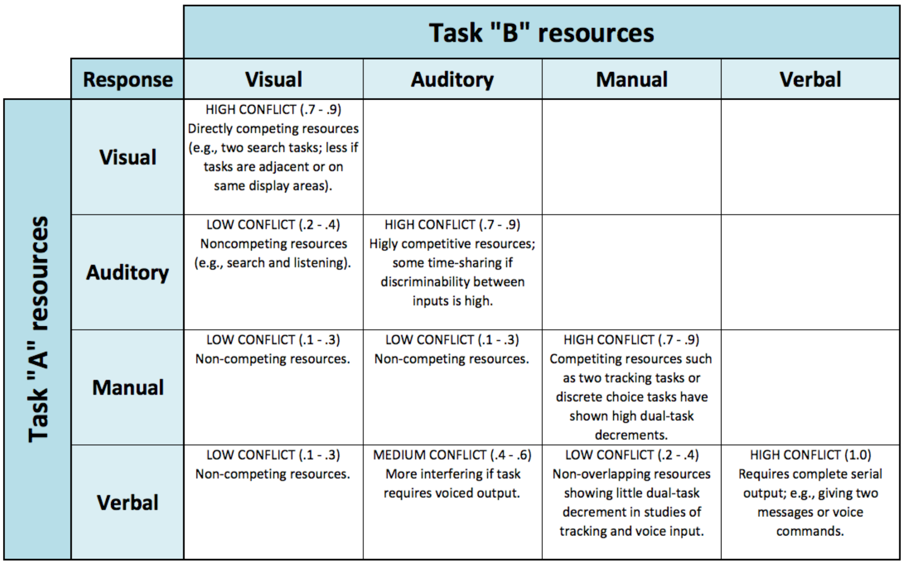
Very Low Very High

Figure 2.3. The NASA-TLX questionnaire.

Workload Index - W/Index

The Workload Index is a tool developed by North and Riley (1989) within the Honeywell Systems and Research Center. The objective of this tool is to predict the mental workload and optimize the design of workstations. The W/Index uses the concept of contention between multiple resources (Wickens, 1984) to calculate the mental workload. This technique compares in an iterative way different work scenarios resulting from different combinations of modalities of stimulus presentation (visual, auditory, manual or verbal channels) so as not to overload the operator. An interesting aspect of this tool is the *"Conflict Matrix"*, used to highlight possible conflicts from tasks that require the operator to use the same *"resource tank"* (Figure 2.4). The added value of this matrix is the possibility to compare the presence of two tasks at the same time.

The W/Index should allow system designers to consider the consequences of some design choices in terms of MWL. The design elements involved are, for example, the physical arrangement of certain functions, the application of automation to specific tasks or the use of various human-machine interface technologies. A limitation of this tool is, in addition to the lack of validation studies, the lack of information compared to the relative values reported by the participants. For example, as reported by the authors themselves (North & Riley, 1989), it is not possible to indicate an upper limit in the W/Index score and this makes it difficult to provide information on the risk of a certain level of MWL.

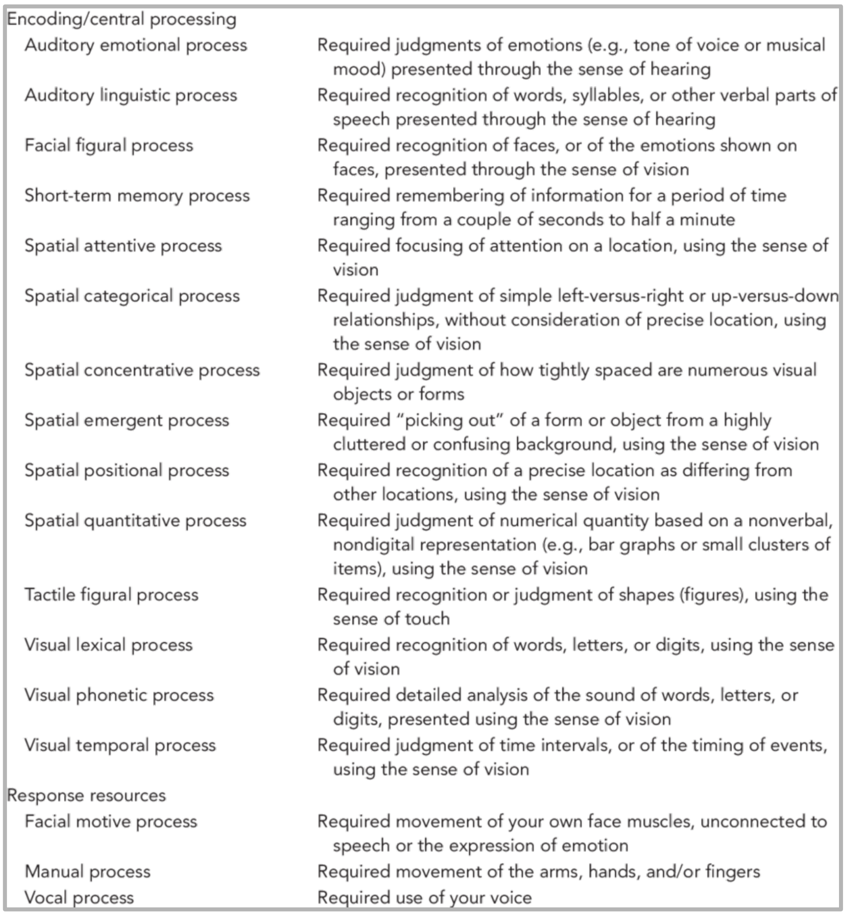


		Task "B" resources			
	Response	Visual	Auditory	Manual	Verbal
Task "A" resources	Visual	HIGH CONFLICT (.7 - .9) Directly competing resources (e.g., two search tasks; less if tasks are adjacent or on same display areas).			
	Auditory	LOW CONFLICT (.2 - .4) Noncompeting resources (e.g., search and listening).	HIGH CONFLICT (.7 - .9) Higly competitive resources; some time-sharing if discriminability between inputs is high.		
	Manual	LOW CONFLICT (.1 - .3) Non-competing resources.	LOW CONFLICT (.1 - .3) Non-competing resources.	HIGH CONFLICT (.7 - .9) Competing resources such as two tracking tasks or discrete choice tasks have shown high dual-task decrements.	
	Verbal	LOW CONFLICT (.1 - .3) Non-competing resources.	MEDIUM CONFLICT (.4 - .6) More interfering if task requires voiced output.	LOW CONFLICT (.2 - .4) Non-overlapping resources showing little dual-task decrement in studies of tracking and voice input.	HIGH CONFLICT (1.0) Requires complete serial output; e.g., giving two messages or voice commands.

Figure 2.4. The Conflict Matrix of the Workload Index.

Multiple Resource Questionnaire - MRQ

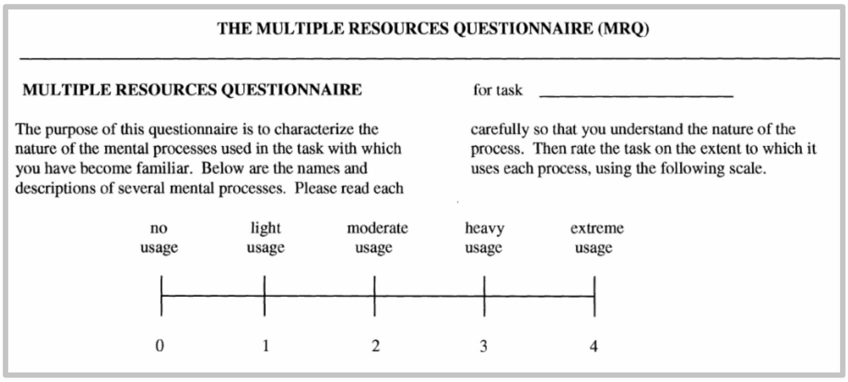
The Multiple Resources Questionnaire (MRQ; Boles & Adair, 2001) is a subjective measure that characterizes the assessment of MWL in seventeen dimensions, expanding Wickens' (1984) model of multiple resources. Like the Workload Index and the Workload Profile, the MRQ items are anchored to the model described by Wickens. In this case there are 17 items and they ask users to evaluate the average amount of use for each "tank" of resources for as long as the task is performed.



Encoding/central processing	
Auditory emotional process	Required judgments of emotions (e.g., tone of voice or musical mood) presented through the sense of hearing
Auditory linguistic process	Required recognition of words, syllables, or other verbal parts of speech presented through the sense of hearing
Facial figural process	Required recognition of faces, or of the emotions shown on faces, presented through the sense of vision
Short-term memory process	Required remembering of information for a period of time ranging from a couple of seconds to half a minute
Spatial attentive process	Required focusing of attention on a location, using the sense of vision
Spatial categorical process	Required judgment of simple left-versus-right or up-versus-down relationships, without consideration of precise location, using the sense of vision
Spatial concentrative process	Required judgment of how tightly spaced are numerous visual objects or forms
Spatial emergent process	Required "picking out" of a form or object from a highly cluttered or confusing background, using the sense of vision
Spatial positional process	Required recognition of a precise location as differing from other locations, using the sense of vision
Spatial quantitative process	Required judgment of numerical quantity based on a nonverbal, nondigital representation (e.g., bar graphs or small clusters of items), using the sense of vision
Tactile figural process	Required recognition or judgment of shapes (figures), using the sense of touch
Visual lexical process	Required recognition of words, letters, or digits, using the sense of vision
Visual phonetic process	Required detailed analysis of the sound of words, letters, or digits, presented using the sense of vision
Visual temporal process	Required judgment of time intervals, or of the timing of events, using the sense of vision
Response resources	
Facial motive process	Required movement of your own face muscles, unconnected to speech or the expression of emotion
Manual process	Required movement of the arms, hands, and/or fingers
Vocal process	Required use of your voice

Figure 2.5. MWL dimensions in the Multiple Resource Questionnaire - MRQ.

During the administration phase, the 17 dimensions investigated by this instrument are presented to the subjects, who are asked to read the description of each process (figure 2.5) and to respond on a 5-point Likert scale indicating the average value on a scale from 0, *"no usage"*, to 100, *"extreme usage"*, in increments of 25 points (figure 2.6).



THE MULTIPLE RESOURCES QUESTIONNAIRE (MRQ)

MULTIPLE RESOURCES QUESTIONNAIRE for task ____________________

The purpose of this questionnaire is to characterize the nature of the mental processes used in the task with which you have become familiar. Below are the names and descriptions of several mental processes. Please read each carefully so that you understand the nature of the process. Then rate the task on the extent to which it uses each process, using the following scale.

Figure 2.6. Presentation of the MRQ and response scale for each of the investigated processes.

2.1.1.3 One-Dimensional Measurements

Cooper Harper rating scale

In 1957, NASA scientist George Cooper presented an instrument with the objective of quantifying the influence of mental workload on performance. This instrument, known as the "Cooper Pilot Opinion Rating Scale", has been used for years in flight tests and fly simulation studies. Later, thanks to the assistance of Robert Harper, the original scale was modified to better evaluate the piloting characteristics of an aircraft. This new assessment was renamed "Cooper-Harper Handling Qualities Rating Scheme" (Harper Jr & Cooper, 1986), and is still widely used as a subjective measure to evaluate the design and performance of an aircraft (Graham et al., 2008). The pilot assesses the flight management of the aircraft on the basis of his ability to control it, the perceived workload and the possibility of achieving specific performance targets. The tool consists in estimating the mental workload following a hierarchical structure with three questions to obtain a discrete score. The first question requires the pilot to indicate whether the situation is controllable. If the answer is yes, the second question is used to estimate the adequacy of the performance based on the workload actually perceived. The pilot may answer whether the mental workload interferes with the performance or not. If the pilot believes he or she can achieve adequate performance with the current workload level, proceed to the third question. The third question requires a judgment regarding the need for support to complete the task at the required performance level. If the pilot responds negatively it means that he needs support to achieve the required performance level. These three questions are linked to descriptors that define the mental workload more analytically. Each descriptor is associated with a value ranging from 1 to 10, where increasing values indicate greater performance impairment due to mental workload (Figure 2.7).

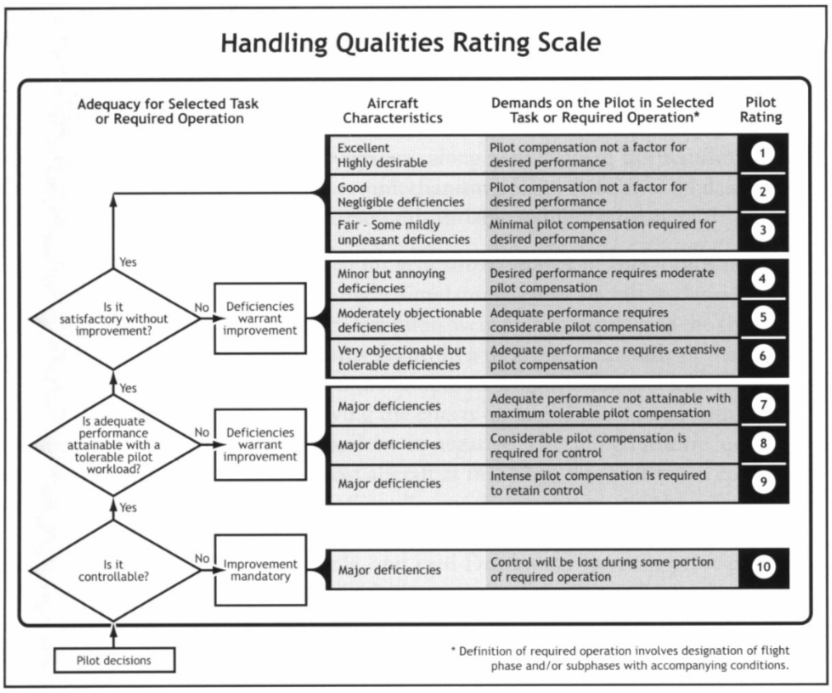

Handling Qualities Rating Scale
Adequacy for Selected Task or Required Operation
Aircraft Characteristics
Demands on the Pilot in Selected Task or Required Operation*
Pilot Rating
Excellent Highly desirable
Pilot compensation not a factor for desired performance
1
Good Negligible deficiencies
Pilot compensation not a factor for desired performance
2
Fair - Some mildly unpleasant deficiencies
Minimal pilot compensation required for desired performance
3
Minor but annoying deficiencies
Desired performance requires moderate pilot compensation
4
Moderately objectionable deficiencies
Adequate performance requires considerable pilot compensation
5
Very objectionable but tolerable deficiencies
Adequate performance requires extensive pilot compensation
6
Major deficiencies
Adequate performance not attainable with maximum tolerable pilot compensation
7
Major deficiencies
Considerable pilot compensation is required for control
8
Major deficiencies
Intense pilot compensation is required to retain control
9
Major deficiencies
Control will be lost during some portion of required operation
10
Yes
Is it satisfactory without improvement?
No
Deficiencies warrant improvement
Yes
Is adequate performance attainable with a tolerable pilot workload?
No
Deficiencies warrant improvement
Yes
Is it controllable?
No
Improvement mandatory
Pilot decisions
* Definition of required operation involves designation of flight phase and/or subphases with accompanying conditions.

Figure 2.7. The Cooper-Harper Rating Scale.

Modified Cooper Harper rating scale - MCH

In 2006, Cummings, Myers and Scott adapted the Cooper-Harper rating scale to the management of Unmanned Aerial Vehicles (UAVs). The main difference in the management of these aircraft is due to the role of automation, which manages most of the flight activities, thus changing the role of the pilot. For this purpose, the Modified Cooper-Harper scale should be more effective in providing guidance on the design and engineering of remote controlled aircraft. The proposed instrument takes the same structure as the MCH but focuses on the display. In the management of UAVs, in fact, the clarity of information presentation is crucial in maintaining the aircraft control. For this reason, the three questions presented in the MCH refer to the ability of the display to convey information effectively (Figure 2.8).

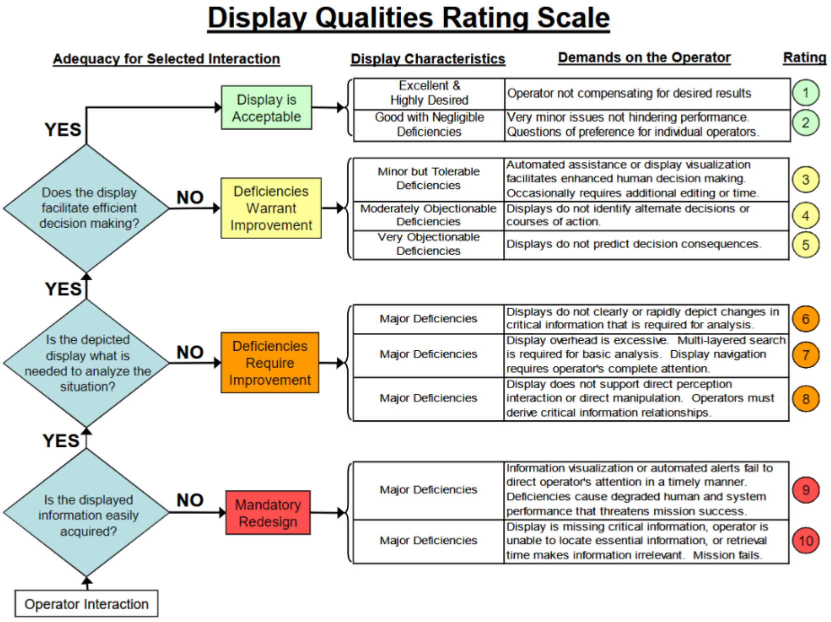

Display Qualities Rating Scale
Adequacy for Selected Interaction
Display Characteristics
Demands on the Operator
Rating
YES
Display is Acceptable
Excellent & Highly Desired
Operator not compensating for desired results
1
Good with Negligible Deficiencies
Very minor issues not hindering performance. Questions of preference for individual operators.
2
Does the display facilitate efficient decision making?
NO
Deficiencies Warrant Improvement
Minor but Tolerable Deficiencies
Automated assistance or display visualization facilitates enhanced human decision making. Occasionally requires additional editing or time.
3
Moderately Objectionable Deficiencies
Displays do not identify alternate decisions or courses of action.
4
Very Objectionable Deficiencies
Displays do not predict decision consequences.
5
YES
Is the depicted display what is needed to analyze the situation?
NO
Deficiencies Require Improvement
Major Deficiencies
Displays do not clearly or rapidly depict changes in critical information that is required for analysis.
6
Major Deficiencies
Display overhead is excessive. Multi-layered search is required for basic analysis. Display navigation requires operator's complete attention.
7
Major Deficiencies
Display does not support direct perception interaction or direct manipulation. Operators must derive critical information relationships.
8
YES
Is the displayed information easily acquired?
NO
Mandatory Redesign
Major Deficiencies
Information visualization or automated alerts fail to direct operator's attention in a timely manner. Deficiencies cause degraded human and system performance that threatens mission success.
9
Major Deficiencies
Display is missing critical information, operator is unable to locate essential information, or retrieval time makes information irrelevant. Mission fails.
10
Operator Interaction

Figure 2.8. Hierarchical structure and descriptors of the Modified Cooper Harper Rating Scale.

The structure of the MCH requires the operator to answer three questions asked in a hierarchical manner. The next question is asked only if the answer to the previous one is yes, otherwise the respondent chooses among the possible descriptors that correspond to a score on an ordinal scale. The highest scores (9 and 10) indicate serious shortcomings in terms of display design and lead to re-design recommendations. These shortcomings, in fact, compromise the acquisition of information that is crucial for a good performance. The scores 6, 7 and 8 indicate important shortcomings in the ability to analyze information and refer to cognitive difficulties imposed by the display. In this case the information is available but difficult to access. Levels 5 to 3 report deficiencies that are not serious but are capable of hindering the decision-making process, increasing the workload of the operator. Finally, the lowest scores on the scale (1 and 2) return a positive evaluation of the interface, which may have negligible defects that do not interfere with the interaction.

Bedford scale

The Bedford scale is a one-dimensional assessment scale designed to identify the mental reserve capacity of the operator during the performance of a certain activity (Roscoe, 1984). The scale was described by Roscoe (1984) as a modification of the Cooper-Harper rating scale with the help of test pilots. The single dimension is evaluated using a hierarchical decision tree that guides the operator through a ten-point evaluation scale, each point of which is accompanied by a descriptor of the associated workload level (Figure 2.9). Paper and pencil are required to make the survey. The operator is asked, on a structure ordered in three levels, if 1) it was possible to complete the task, 2) the workload was tolerable, or 3) the workload was satisfactory without reduction and after this first decision the operator must classify his workload on the respective end points of the evaluation scale (1-10) from “insignificant workload” to “task abandoned”.

This measure shares the limitations of the other subjective measures, in particular the possibility that the respondent is influenced by the procedure of administration of the measure, as the instructions and training of the subjects with the scale are not recorded. In addition, there are no rules for the interpretation of the data. In this regard, Wainwright (1987) suggests that a workload in the lower range (1 to 3) should be considered adequate in all assessments.

The Bedford Scale has been used mainly in applied contexts (Lysaght et al., 1989) due to its ease of use, both in terms of administration and scoring of results. However, there are no studies carried out on the validation of the scale.

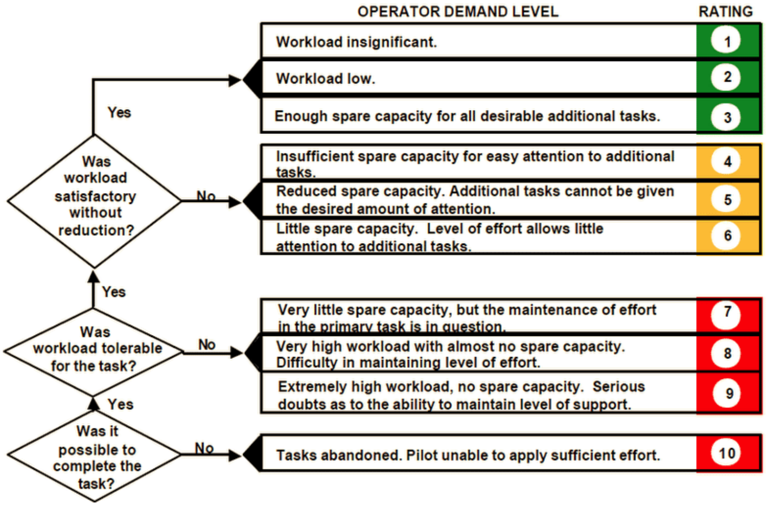

OPERATOR DEMAND LEVEL
RATING
Workload insignificant.
1
Workload low.
2
Enough spare capacity for all desirable additional tasks.
3
Insufficient spare capacity for easy attention to additional tasks.
4
Reduced spare capacity. Additional tasks cannot be given the desired amount of attention.
5
Little spare capacity. Level of effort allows little attention to additional tasks.
6
Very little spare capacity, but the maintenance of effort in the primary task is in question.
7
Very high workload with almost no spare capacity. Difficulty in maintaining level of effort.
8
Extremely high workload, no spare capacity. Serious doubts as to the ability to maintain level of support.
9
Tasks abandoned. Pilot unable to apply sufficient effort.
10
Yes
Was workload satisfactory without reduction?
No
Yes
Was workload tolerable for the task?
No
Yes
Was it possible to complete the task?
No

Figure 2.9. Hierarchical decision tree of the Bedford Scale.

The Bedford Scale is also characterized by a particular vocabulary. It uses the terminology of reserve capacity, emphasizing the information processing dimension of the workload. This is recognized as an essential dimension of workload, and is significant for the Human Factors community, but does not always seem to have the same meaning for everyone (Brennan, 1992).

Air Traffic Workload Input Technique - ATWIT

The Air Traffic Workload Input Technique (ATWIT) is a technique to measure the mental workload of air traffic controllers in "real time". The instrument was proposed by the Technical Center of the Federal Aviation Administration (FAA) for the real-time measurement of the mental workload of air traffic controllers (Stein, 1985). This tool requires respondents to indicate the perceived mental workload on a scale from 1 (very low workload) to 7 (very high workload) via a push-button panel (Workload Assessment Keypad, WAK). The respondents have to press the keypad within a limited period of time after the presentation of an acoustic and visual signal (a tone and lighting).

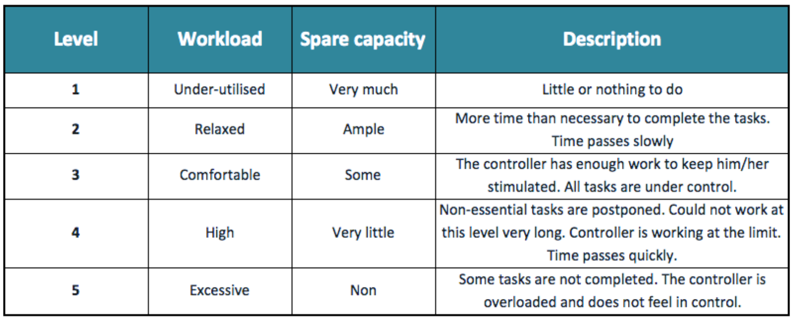
The instrument was validated in a study in which ten air traffic controllers performed a series of one-hour simulations designed to produce a low, moderate and high workload range. The responses of the controllers and observers confirmed the presence of three levels of workload, directly related to the difficulty of the tasks performed.

The ATWIT is an assessment scale designed for air traffic control studies and aims to provide a workload profile that should accurately reflect changes in workload due to changes in air traffic condition. Compared to the NASA-TLX, for example, the ATWIR makes repeated recordings of the perceived mental workload and also records the time needed to respond to each question (Manning et al., 2001).

Instantaneous Self Assessment of workload - ISA

The Instantaneous Self Assessment of Workload (ISA) was developed by the Air Traffic Management Development Centre, National Air Traffic Services (ATMDC; Brennan, 1992). This tool is very easy to use in real-time simulations. By using a five-point assessment scale, the method requires that, every two minutes, the respondent gives an indication of his or her subjective level of workload. The question asked to the respondent is "*How do you evaluate your workload?*" and the frequency with which an answer is recorded allows researchers to draw a workload profile throughout the performance of task. The answer gives an evaluation from 1 to 5 on the degree of effort required from the operator at that precise moment (1 = under-utilised and 5 = excessive, Figure

2.10). This data can be used to compare the workload perceived by operators using different tools or systems.



Level	Workload	Spare capacity	Description
1	Under-utilised	Very much	Little or nothing to do
2	Relaxed	Ample	More time than necessary to complete the tasks. Time passes slowly
3	Comfortable	Some	The controller has enough work to keep him/her stimulated. All tasks are under control.
4	High	Very little	Non-essential tasks are postponed. Could not work at this level very long. Controller is working at the limit. Time passes quickly.
5	Excessive	Non	Some tasks are not completed. The controller is overloaded and does not feel in control.

Figure 2.10. The Instantaneous Self-Assessment of workload scale.

Operators use a small keyboard to indicate their workload, subtracting minimal time from the main activity. It is important that the way in which the workload is recorded does not interfere with the primary task. The measurement is designed to be fast and not intrusive in order to avoid an increment in the user workload due to the scla administration. In addition, the instrument has been designed to facilitate compliance by operators. The recordings obtained from this measurement allow researchers to evaluate the workload levels experienced as a result of the introduction of new equipment, procedures or other system attributes. It also indicates how the workload varies over time. The question asked to the subjects and the scale of answer remain the same, but the conditions of the test or simulation can change. After a certain period of time, the answers of the user can be judged according to what was happening in that moment in order to obtain an indication of the effect of the conditions of the test on the workload perceived by the user.

ISA (Brennan, 1992) is relatively simple in its presentation and application. The instrument seems to be evaluated positively within the area of human factors with regard to air traffic control. The five points on the scale risk favouring a central trend effect and for this reason the scale is sometimes reduced to three points in real practice (Pickup et al., 2005).

Subjective Workload Estimate Rating Scale - AFFTC

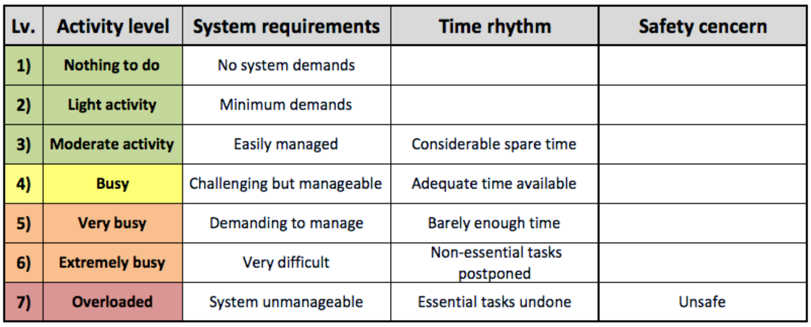
This measure was developed by Ames and George in 1993 within the Air Force Flight Test Center (AFFTC). The two authors (Ames & George, 1993) modified an unpublished scale already proposed by the School of Aerospace Medicine

(SAM). The objective of the tool is an easy and quick evaluation of the mental workload in operational contexts or flight simulations. Even if the measure is presented as one-dimensional, the logic behind its construction is to integrate different dimensions of the workload in a single score. These dimensions are: activity level, system requirements, time pressure and safety of operations. In detail, the dimensions investigated by the Subjective Workload Estimate Rating Scale are:

- *Activity Level:* it can range from having nothing to do to having to handle too many tasks. The respondent's actions are described relative to the scope of the tool and the physical activity becomes more complex as the variety of actions increases and the physical location of the action changes from one place to another. High levels of physical activity can stress muscles, exhaust energy reserves, cause fatigue and tiredness, and eventually lead to total exhaustion;
- *System Requirements:* Demands can range from simple and repetitive to complex and challenging. Difficult tasks can involve detecting stimuli that are difficult to see or hear, requiring extreme concentration to overcome distractions, involving detailed memory or thinking, and requiring important decisions to be made. Tasks may also require precise hand-eye control or coordination between arms and legs. In addition, the work environment may include conditions that make work difficult, such as extreme heat or cold, high levels of humidity, distracting noise or vibration, and poor air quality. The physical condition of the worker can also increase the workload, such as lack of sleep or rest, inadequate food or water intake, or inadequate or unattractive work space.
- *Time Rhythm:* The time available to perform tasks can vary from abundant to non-existent. Inadequate time availability can stress workers by increasing the workload. When time is short, users may need to prioritize multiple tasks with mental priority and act quickly, often resulting in errors and poor performance. Sometimes tasks can be postponed or even completely ignored. The resulting confusion and frustration further increases the workload.
- *Safety concern:* concerns about personal physical safety or the responsibility to protect equipment or supplies from damage increases the

subjective workload. Safety concerns are high when situations are inherently dangerous and life-threatening. Other situations can be dangerous and stressful because the operator cannot see or hear the necessary information, or because the design of the system does not allow an adequate control to the operator.

Each dimension is characterized by descriptors to which is associated a discrete value ranging from 1 to 7, where 7 is the maximum workload level (see figure 2.11).



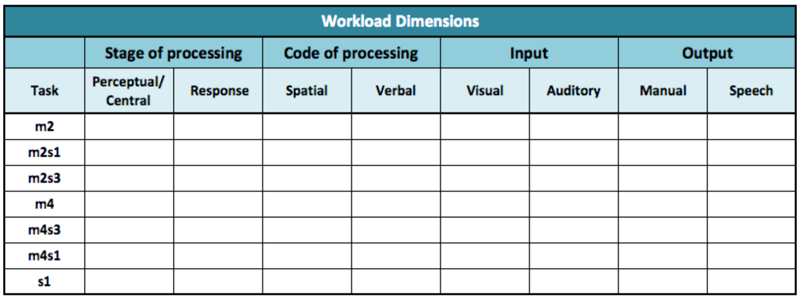
Lv.	Activity level	System requirements	Time rhythm	Safety concern
1)	Nothing to do	No system demands		
2)	Light activity	Minimum demands		
3)	Moderate activity	Easily managed	Considerable spare time	
4)	Busy	Challenging but manageable	Adequate time available	
5)	Very busy	Demanding to manage	Barely enough time	
6)	Extremely busy	Very difficult	Non-essential tasks postponed	
7)	Overloaded	System unmanageable	Essential tasks undone	Unsafe

Figure 2.11. The Subjective Workload Estimate Rating Scale.

In order to verify the psychometric attributes of the scale, the authors carried out a procedure of cross comparisons between levels, following the procedure already described by Vidullch, Ward and Schueren (1991). Subsequently, the authors carried out a further validation study of the steps that make up the scale. This validation required a sample of subjects to order the scale descriptors in order to verify the pertinence of its ordinal dimension.

Workload Profile - WP

Tsang and Velazquez (1996) introduced a method for mental workload assessment that seeks to capitalize on both the high diagnostic capability of dual task procedures and the ease of use of subjective techniques. The WP is based on the multiple resource theory of Wickens (1984). According to the Wickens' theory (Wickens, 1984) the resource reservoirs are independent and depend on the stage at which the elaboration process takes place (coding, processing, response), the modalities involved (visual, auditory), the codes used (spatial, verbal) and the type of response (manual, vocal). The WP (Tsang & Velazquez, 1996) then asks the subjects to indicate the percentage of attentive resources used after carrying out all the tasks under investigation. Subjects at the time of evaluation can consult the definition of each dimension. In each cell of the evaluation sheet, subjects provide a number between 0 and 1 to represent the proportion of attentive resources used in a given dimension for a given task. An evaluation of “0” means that the task has not required the use of a particular dimension; an evaluation of “1” means that the task has required the maximum use of that particular dimension. Individual dimensional assessments are added together for each assignment to provide an assessment of the overall workload (Figure 2.12).

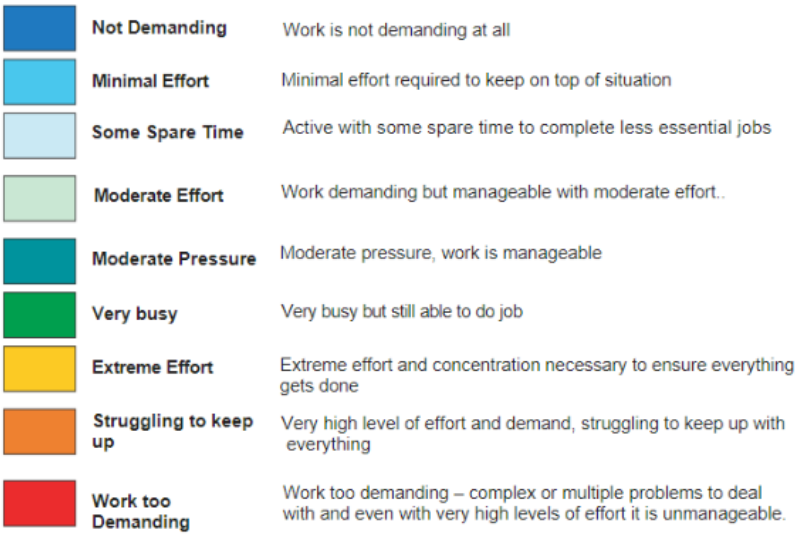


Workload Dimensions								
	Stage of processing		Code of processing		Input		Output	
Task	Perceptual/ Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
m2								
m2s1								
m2s3								
m4								
m4s3								
m4s1								
s1								

Figure 2.12. The Workload Profile structure (Tsang e Velazquez, 1996).

Integrated Workload Scale – IWS

The Integrated Workload Scale is a measure adopted by Pickup, Wilson, Norris, Mitchell and Morrisroe (2005) for the assessment of mental workload in railways. The measure takes its cue from ISA, from which it derives the one-dimensional structure but adopts a 9-point response scale, reducing the tendency to a central evaluation of ISA (Figure 2.13). The tool has proven to be valid, sensitive and usable to evaluate the workload in the railway industry in two separate simulation tests. Its user-friendliness makes it easy to implement and does not seem to interfere significantly with rail work in a simulated environment (Pickup et al., 2005).



🟦	**Not Demanding**	Work is not demanding at all
🟦	**Minimal Effort**	Minimal effort required to keep on top of situation
🟦	**Some Spare Time**	Active with some spare time to complete less essential jobs
🟩	**Moderate Effort**	Work demanding but manageable with moderate effort..
🟦	**Moderate Pressure**	Moderate pressure, work is manageable
🟩	**Very busy**	Very busy but still able to do job
🟨	**Extreme Effort**	Extreme effort and concentration necessary to ensure everything gets done
🟧	**Struggling to keep up**	Very high level of effort and demand, struggling to keep up with everything
🟥	**Work too Demanding**	Work too demanding – complex or multiple problems to deal with and even with very high levels of effort it is unmanageable.

Figure 2.13. The Integrated Workload Scale.

The administration of this instrument is very similar to the one used by Eurocontrol for the ISA (Hering & Coatleven, 1996), i.e. it is based on the use of a keyboard and a nine-point response scale. The operator through the keyboard can express his evaluation of the mental load after hearing an acoustic alarm signal.

2.1.2 Behavioural measures

Behavioural measures are based on the main assumption that the mental workload affects the level of performance of an individual interacting with a system. In the literature (Cain, 2007; O'Donnel & Eggemeier, 1986; Tsang & Vidulich, 2006; Wickens & Hollands, 1999; Wilson & Eggemeier, 2006; Young & Stanton, 2004) performance measures are generally divided into two categories: those based on the performance of the individual at the primary task and those based, according to the “double task paradigm”, on the performance of the individual at a secondary task.

Primary-task performance measures

Mental workload measurements that relate to an individual's performance at the primary task are based on two assumptions: i) individual cognitive resources are limited; ii) when the demands imposed by the task increase the performance inevitably worsens. The most commonly used Behavioural indicators are:

- Reaction times;
- Accuracy;
- Success rate;
- Error rate;
- Number of omissions;
- Task completion time (Tsang & Vidulich, 2006).

There is no measure more discriminating than another. If we consider the individual in everyday life, outside of experimental contexts, this kind of performance is task and context specific (De Waard, 1996). However, since these metrics reflect in a simple and direct way the result of the effort exerted by an operator interacting with a system, they are often used as techniques for assessing mental workload (O'Donnel & Eggemeier, 1986).

According to O’Donnell and Eggemeier (O’Donnel & Eggemeier, 1986), performance during a primary task is an index of the quality of the interaction between man and machine. However, the use of performance indicators for the primary task raises doubts about their ability to discriminate against the mental workload net of other factors that may affect the performance. In fact, two individuals may perform similarly to the same task but experience a different mental workload depending on the expended resources (Gopher and Donchin, 1986). These limitations point out that performance measures at the primary task do not provide a sufficiently reliable estimate of the mental workload.

Secondary-task performance measures

The “*Dual-task paradigm”* has been widely used to limit the reliability problems related to the use of a single task. This methodology requires the individual to perform, simultaneously with the primary task, another task called secondary. The assumption behind the dual-task paradigm is that as the difficulty of the primary task increases, performance at the secondary task worsens. This happens because the individual invests more resources in the primary task and does not have sufficient resources for processing and responding to other stimuli. The evaluation of mental workload is based on the analysis of the individual’s performance in the secondary task: if the performance at the secondary task is poor, a high mental workload is assumed (O’Donnel & Eggemeier, 1986). Finally, by comparing the performance obtained by the individual during the execution of the primary task presented individually with the performance at the primary task administered at the same time as a secondary task, it is possible to obtain information on the residual capacity of the operator (De Waard 1996; Kahneman, 1973; Rubio et al., 2004; Wiebe, Roberts & Behrend, 2010).

The choice of the secondary task must be made with care and accuracy. It is important to consider the nature of the primary task and the variables being studied. The secondary task should, specifically, be sensitive to changes in the primary task, and therefore have a certain “fit” with the characteristics of the primary task. According to Wickens (1984), the secondary task should require the individual to use the same type of resources as the primary task.

2.1.3 Physiological measures

Physiological measures have been widely used to estimate the mental workload (see Charles & Nixon, 2019). These measures have the advantage of providing objective and continuous information on the cognitive load experienced by a user. Several authors (Brookhuis & De Waard, 2010; De Waard, 1996; Noyes & Bruneau, 2007; Rubio et al., 2004; Wiebe et al., 2010) highlighted that the relation between physiological measures and MWL is indirect. For this reason it is advisable to combine physiological and Behavioural measures to maximize the reliability of the assessments. When using physiological measures it is very important to reduce their intrusiveness. Such instruments should not affect or interfere with the natural execution of the task.

The most commonly used physiological measures in the assessment of MWL refer to cardiovascular activity, brain activity and eye movements of the operator.

Cardiovascular measures

The most commonly used cardiac measures refer to electrocardiogram (EKG), blood pressure, blood oxygen concentration and heart rate (Heart rate, HR) (Wilson et al., 2004). Some studies (see Charles & Nixon, 2019, Lean & Shan, 2012) have found an increase in heart rate and a decrease in heart rate variability (HRV) associated with an increase in mental workload. However, the use of these measures has shown several limitations regarding the lack of sensitivity and diagnostic capability. Heart rate is influenced by multiple cognitive and physiological processes. Therefore it is risky to attribute a change in heart rate exclusively to increases or decreases in mental workload (O'Donnel & Eggemeier, 1986).

Neurophysiological measures

The main neurophysiological measures used for the mental workload assessment refer to the analysis of electroencephalogram (EEG), event-related potentials (ERP), magnetic resonance imaging (MR) or functional magnetic resonance imaging (fMR).

ERP components are generally identified by their polarity, which can be negative, "N", or positive, "P", and by a number indicating the minimum recorded latency time from the moment the event eliciting the sequence occurs. The P300 is the most widely used ERP in the study of cognitive processes in Human-Computer Interaction. The P300 records a positive deflection 300 milliseconds after the appearance of the stimulus with greater amplitude in the front-central areas (Donchin et al., 1986; Parasuraman, 1990). The amplitude of P300 is generally an index of the amount of cognitive resources invested by an individual for the processing of a particular stimulus. Numerous studies (Israel et al., 1980; Käthner, et al., 2014; Kutas et al., 1977; McCarthy & Donchin, 1981; Ragot, 1984) have tried to evaluate the diagnostic capacity of the P300 as a measure of mental workload. Using the dual-task paradigm it has been noticed how the amplitude of the P300 associated with the secondary task is reduced in correspondence with an increase in the difficulty of the primary task that, necessarily, saturates the resources available to the individual. The P300 is rather sensitive to factors that have effects on verbal/spatial and visual/acoustic processing. However, the same studies have shown that the P300 is insensitive to factors that influence the motor processing imposed by the various tasks. The major limitations of neurophysiological measurements concern the sensitivity, diagnostic capacity, and especially the intrusiveness of measuring instruments that limit their use in real work contexts. However, these measurements are useful indicators to be used as a concurrent measure in experimental validation studies.

3. Eye tracking in usability and mental workload research

Eye tracking is a method that is gaining more and more popularity in the scientific community for the evaluation of usability and mental workload (Bergstrom & Schall, 2014; Goldberg & Wichansky. 2003; Jacob & Karn, 2003; Majaranta & Bulling, 2014; Marquart, Cabrall, & de Winter, 2015; May et al., 1990; Nielsen & Pernice, 2010; Pan et al., 2004; Poole & Ball, 2006). The analysis of eye movements provides objective information on the behavior of the individual avoiding possible distortions related to subjective metrics (such as self-report questionnaires and interviews).

In the next paragraph a brief description of the human eye system and of the main eye-tracking techniques will be provided. Subsequently the main usability and MWL ocular metrics will be described.

3.1. Eye movements: typology and characteristics

A synthetic description of the path that leads to the elaboration of a visual stimulus could be this: the light penetrates inside the eye through the pupil, the image is inverted in the crystalline lens, after which it is projected into the back of the eyeball, i.e. into the retina (Lens, Nemeth & Ledford, 2008).

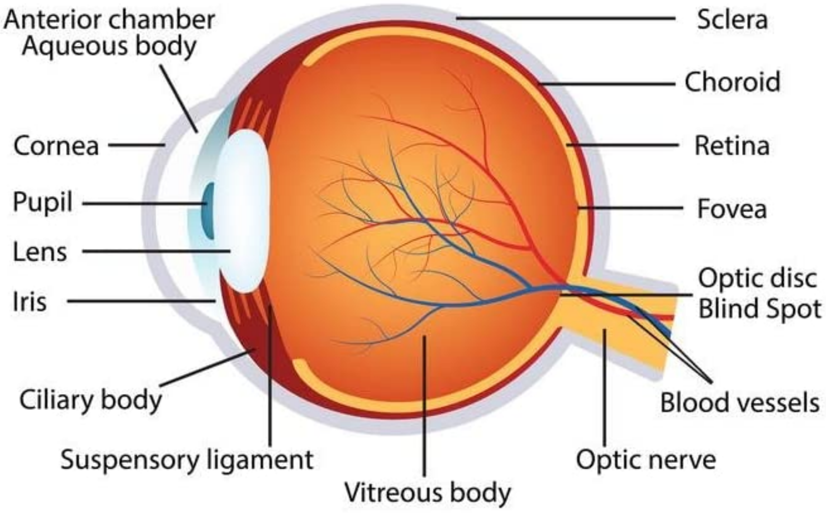


Figure 3.1. Anatomy of the eye.

The retina consists of light-sensitive cells, cones and rods, which transduce light into electrical signals to be sent through the optic nerve to the visual cortex for information processing. Cones allow a detailed view of the visual scene and to distinguish the various colours. Rods are more sensitive to variations in light, movement and depth than to the details of stimuli and allow individuals to see in low light conditions. The distribution of cones and rods in the retina is not uniform. In fact, the cones are concentrated in the centre of the retina, in a small area called fovea, able to cover only 2° of the visual field, while the rods are more present in the "periphery" of the retina. For this reason, visual acuity, namely *"the ability of the eye to resolve and perceive fine details of an object"* (Cline, Hofstetter & Griffin, 1996), is highest in the centre of the retina and decreases rapidly towards its periphery.

The fine characteristics of a visual stimulus can only be extracted through a foveal vision and, on a functional level, it follows that individuals must necessarily move their eyes to get a detailed view of the different elements present in the environment.

Generally, the eyes make rapid movements, called saccades, followed by short stops, called fixations. Saccades and fixations are commonly considered the basic elements of eye behavior. The sequence generated by alternating saccades and fixations is called "*scanpath*".

The aim of the saccades is to bring the elements of interest in the foveal area of the retina, they last on average between 30 and 80 milliseconds and rarely proceed from one point to another along the shortest possible segment, they can in fact follow different shapes or curves. An individual performs about 3-4 saccades per second, during which he is not able to perceive any visual stimulus, this phenomenon has been defined as "*saccadic suppression"* (Ishida & Ikeda 1989; Wolverton & Zola, 1983). When the eyes follow a target that moves along the field of vision can make movements slower than the saccades, these movements are called "smooth pursuits" and vary depending on the speed of the target (Holmqvist et al., 2011).

The stimuli present in the visual field of the individual reach the foveal part of the retina during fixations, which can last from 60 milliseconds to several seconds. Some authors (Just & Carpenter, 1980) believe there is a positive relationship between fixations and degree of attention. However, this relationship is not always considered valid. In fact, although an individual can keep his eyes fixed on a certain stimulus his attention could be turned elsewhere (Anderson, Bothell & Douglass, 2004).

Although the term fixation is commonly used the eye is never completely stationary. In fact, there are three different types of micro-movements: “tremors” or “nystagmuses”, “microsaccades” and “drifts” (Martinez-Conde, Macknik, & Hubel, 2004). Nystagmuses are small movements, involuntary in most cases, probably due to inaccuracies in muscle control. Drifts are movements that slowly distract the eye from the centre of fixation, for this reason, the function of microsaccades is to quickly return the eye to its original position. Table 3.1 shows the average values in terms of duration (expressed in milliseconds), amplitude and speed (expressed in degrees or minutes, where 1° = 60') related to the most used measures for the study of eye movements in psychology, cognitive science, ergonomics and neuroscience.

Table 3.1. Eye movements and their average values (from Holmqvist, Nystrom, Andersson, Dewhurst, Jarodzka & Van de Weijer, 2011, p. 114).

Eye movement	Duration (ms)	Amplitude	Velocity
Fixation	200-300	-	-
Saccade	30-80	4-20°	300-500°/s
Smooth pursuits	-	-	10-30°/s
Microsaccade	10-30	10-40'	15-50°/s
Tremor	-	>1'	20'/s (peak)
Drift	200-1000	1-60'	6-25'/s

3.2 Eye Tracking

The term eye tracking refers to the use of appropriate techniques and tools for the identification of an individual’s eye movements. Eye tracking allows to detect and analyze data related to “what” a subject looks at in his visual field and “how” during the performance of a task.

Yarbus (1967) was among the first to analyze in detail the ocular paths obtained through eye-tracking techniques. His famous experiment on the visual exploration of Repin’s painting “The Unexpected Visitor” can be considered a milestone in the study of eye movements. He noted that the participants in his experiment explored the painting differently depending on the task assigned to them. This was reflected in differences in the visual exploration paths generated by the participants.

Over the years, different methods for measuring eye movements have been developed. Technological development has allowed the replacement of the first instruments, uncomfortable and inaccurate, with less invasive devices characterized by high sensitivity and precision in the recording of eye behavior.

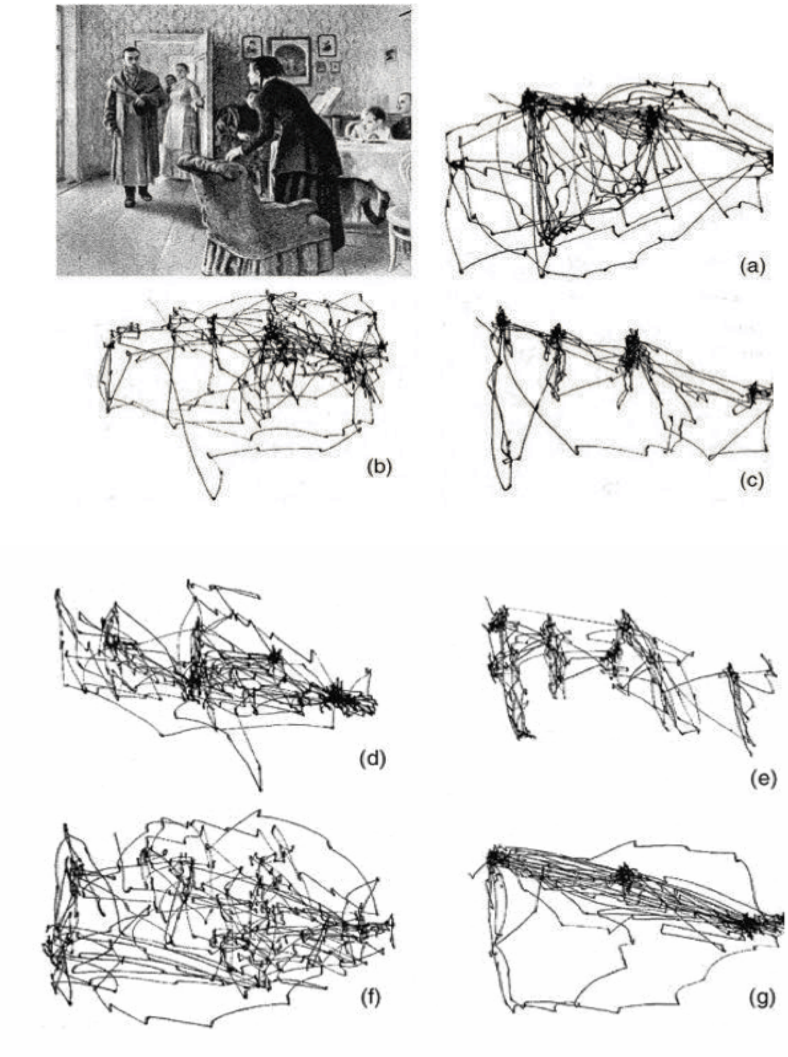


Figure 3.2. The Yarbus experiment: the visual exploration of an image changes according to the users' goal; a) free exploration; b) examines the social condition of the family; c) assigns an age to the subjects; d) tries to understand what they were doing before the visitor's arrival; e) remembers the clothes worn by the subjects; f) remembers the position of objects and people; g) establishes how long the unexpected visitor has been away.

3.2.1 Recording techniques

Over the years, different techniques for recording eye movements have been developed: electro-oculography (EOG), photo-oculography (POG) or video-oculography (VOG), galvanometric or "scleral coil" technique (scleral contact lenses/search coil), infrared oculography (combined pupil-corneal reflection) (Duchowski, 2017).

Electro-oculography is a technique based on the measurement of electrical potential variations associated with the movement of the eyeball. These variations are measured through four electrodes placed just above, below, left and right of the subject's eye. This technique, used since the ‘60s, lacks accuracy in the detection of movements, however, has the advantage of being cost-efficient, and is also the only one applicable to study eye behavior during sleep.

The galvanometric or “scleral coil” technique uses a contact lens that covers the cornea and sclera to which is connected a “pedicle” through which the lens sends data related to ocular activity to a mechanical or optical device, such as a coil that measures the electromagnetic variations. Although it is a very precise method it has the obvious disadvantage of being too intrusive.

The photo/video-oculography allows through sequences of shots or video footage to measure specific characteristics of the eyes during their movement (such as the shape of the pupil, the edge that separates the sclera and iris and the corneal reflexes caused by one or more light sources, usually infrared). Controlling the stimuli presented to a subject at a given time, and the direction of the gaze, the technique allows us to make assumptions about visual behaviour. However, even this technique can analyze only the ocular movement in itself, based on the position of the head, which must be held fixed through a chin rest.

Finally, infrared oculography is based on the cornea's capability to reflect the infrared light. This technique can record through a camera with CCD (Charge Coupled Device) sensor the corneal and pupil reflexes generated by an infrared light source. In general, data sampling takes place at a speed that varies between 30 and 2000 Hz, depending on the device used, and with an accuracy between 1/2° and 2° of the field of view. The calibration procedure is the first step to recording eye movements with this technique. It consists of matching a specific number of points presented to the subject with its corneal reflexes (also known as Purkinje reflections). Thus, corneal reflexes allow researchers to identify the exact position of the pupil and to derive the direction of the gaze. Infrared oculography has the advantage of being not intrusive and of providing a precise estimation of the gaze direction, also offering compensation for head movements (Goldberg & Wichansky, 2003).

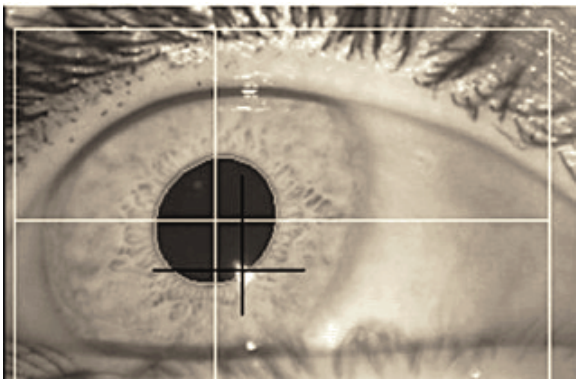


Figure 3.3. Infrared Oculography: the detection of eye movements is based on the identification of the position of the pupil (intersection of white lines) and the reflections generated by infrared light on the cornea (intersection of black lines).

There are mainly three types of Eye trackers that use the infrared oculography technique:

1) *Monitor-based:* generally they consist of 17'' LCD monitors (or higher) that integrate the eye movement detection device. This type of eye-tracker is the most widespread. It records the direction of the individual's gaze in reference to the screen and for this reason it is the most suitable for the study of desktop application interfaces and for the direct control of the computer in the assistive field.

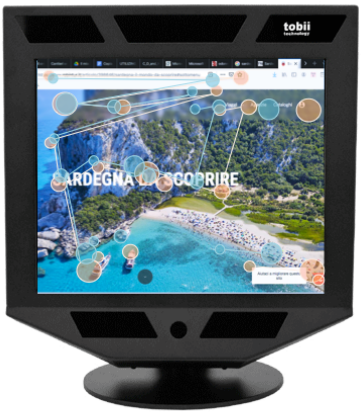


Figure 3.4. Monitor-based Eye Tracker

2) *Mixed*: these eye-trackers do not have their own monitor, but only infrared LED emitters and a video camera to identify the position of the pupil and the reflections produced by infrared light on the cornea. A mixed eye-tracker has several advantages, in fact it can be used with any screen (after an initial calibration phase) and allows to measure the direction of the gaze even outside a laboratory context.

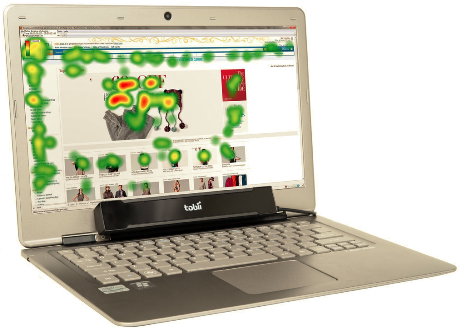


Figure 3.5. Mixed Eye-Tracker

3) *Wearable or head-mounted:* they include those eye tracking systems that must be worn by the user. In the past, such devices consisted of real "helmets" to be placed on the user's head (with the disadvantage of being very invasive, and requiring generally long and complex calibration procedures). Nowadays, the wearable systems look like hats or glasses on which the infrared emitters and the video camera for recording are fixed. These instruments allow detecting eye movements without altering the ecological validity of the task; it is also possible to detect the eye behavior of individuals in different circumstances of daily life.

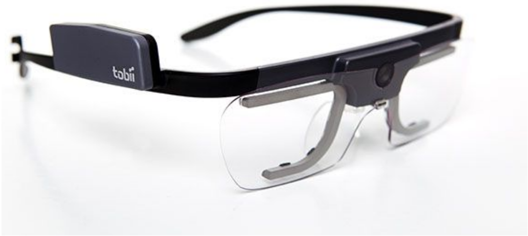


Figure 3.6. Wearable Eye Tracker

3.2.2 Advantages and disadvantages of eye tracking

The accuracy of the recordings, the low intrusiveness and the ecological validity are today considered the greatest advantages of eye tracking.

Eye-trackers on the market are able to provide data regarding eye position, pupil diameter, and distance of the subject from the device. It’s not difficult to imagine the use of this technology into everyday devices such as personal computers, tablets or smartphones in the near future.

In spite of the progress achieved so far, some limits related to the accuracy of the recordings and the commercial availability of eye-tracking devices still need to be overcome. A first limit is the noise signals generated by movements of the subject’s head, blinks, external light or other factors that mislead the recording device. In fact, devices that use infrared light can have problems in detecting the eye position of users wearing glasses, or users with a particular anatomical shape of the eyes (eg: almond-shaped eyes).

During the performance of a task the recordings of eye movements can also be distorted by two types of errors, one due to poor accuracy of the instrument that involves a dispersion of points (gaze points) around the real fixation maintained by the subject (variability error), and one due to poor accuracy that involves moving the average position of the gaze points from the real fixation executed by the subject (systematic error). Another limit is represented by the high price of the devices on the market. This barrier in fact prevents the purchase to universities or research companies that can not have a specific budget (Goldberg & Wichansky, 2003).

3.3 Eye tracking and usability

In the reference literature the main metrics used to assess usability refer to the number, duration and frequency of fixations, the number, amplitude and speed of saccades, the proportion of time spent within a specific area of interest (AOI). Further metrics used to assess usability refer to the scanpath analysis, the number of fixations within each area of interest and the transition between different areas of interest.

Bojko (2013) distinguishes between two categories of indicators and metrics, which recall the concepts of “*attraction*” and “*performance*”. Attraction metrics are useful to assess how much an element is able to capture the user's attention regardless of whether the user actively looks for it with his or her eyes. Performance metrics are useful for measuring the visibility of an area or object. These concepts are in turn decomposable into more specific categories, which refer to the concepts of “*noticeability*”, “*interest*” and “*emotional stimulus”* for what concerns the metrics of attraction and the concepts of *“cognitive overload*”, “*findability*” and “*recognizability*” for what concerns the performance metrics.

Table 3.2. Eye-tracking metrics classification according to Bojko (2013).

Dimension	Indicator	Metrics
Attraction	**Target noticeability**	Percentage of participants looking at an area of interest (AOI)
		Number of gaze fixations prior to the first fixation of an area of interest
		Time to first-fixation
	Interest	Total fixations number
		Total fixation time
		Percentage of time spent looking at an area of interest compared to total fixation time
Performance	**Target findability**	Percentage of participants looking at the target area of interest
		Number of gaze fixations prior to the first fixation of the target AOI
		Time to first-fixation
	Target recognizability	Number of fixations on the area of interest before the item is selected
		Time elapsed between the first fixation and the moment the item is selected
	Cognitive overload	Average fixation duration

One of the very first studies to include ocular metrics in usability assessments was conducted by Paul Fitts' working group (Fitts, Jones & Milton, 1950) in 1947. The authors recorded through fixed cameras the eye movements of 40 Air Force pilots during landing procedures. The recommendations suggested by Fitts and his collaborators are still considered valid, although the results of their studies are influenced by large individual differences in the participants' eye behavior. On the basis of the results obtained, Fitts and colleagues indicated that: the frequency of fixations on a particular area is an indication of its relevance for the user; the duration of the fixations is directly proportional to the difficulty of information processing; transitions of the fixations between nearby areas of interest, and therefore a lower saccadic amplitude, indicate a correct arrangement of the information in the individual's visual field. Over time, these results have been confirmed in numerous other studies.

Goldberg and Kotval (1999) analyzed the eye movements of 12 subjects during interaction with an interface presented in two different graphical layouts. In particular, the same interface was presented to the participants in an “optimal” version and in a “poor” version with respect to the degree of optimization of the function menu layout in the control panel. In the “optimal” condition the menu has been arranged in order to group similar functions (editing, drawing, text) according to the principle that individuals tend to expect neighbouring elements to be connected by some common, physical or conceptual, feature (Wickens & Carswell, 1995). In the condition of “poor” optimization the menu has been arranged randomly. The analysis of eye movements showed a higher number of saccades, a higher irregularity of the scanpath and a higher number of fixations in the “poor” optimization condition than in the “optimal” condition. In addition, in the “poor” condition the ratio of fixations on the area of interest to total fixations was significantly lower, indicating a low efficiency of the exploration strategy.

Byrne and collaborators (Byrne et al., 1999) studied the visual research strategies of exploring vertical menus in a study involving 11 participants. The menus used for the experimental sessions differed from each other in the number of items. In particular, participants were asked to identify a target stimulus within menus consisting of three, six and nine items. The results showed that the time taken by participants to identify the target stimulus (time to first fixation) increases as the number of items in the menu increases, and varies according to the position that the target stimulus occupies in the list of items. A correct arrangement of the items within the menu is associated with a shorter time to identify the target, and also promotes a more effective information search strategy.

Goldberg and his collaborators (Goldberg et al., 2002) analyzed the eye patterns of 7 subjects. The participants had to perform six different tasks, of varying difficulty, on a digital interface organized in different thematic areas and with different functionalities. The assigned tasks, specifically, concerned actions such as “customize a thematic area”, “hide a thematic area”, “find a certain content”, or “logout from your personal page”; each task could require 2 to 7 actions (mouse clicks) to be completed. The results showed how the subjects’ eye patterns varied depending on the difficulty of the task. In fact, more complex tasks were associated with a larger scanpath width (i.e. the length of the inspected area) and a larger saccadic width. The authors attributed this evidence to differences in the mental representations of the subjects regarding the arrangement of windows and functions in the interface. According to the authors, in fact, the increased scanpath length could be indicative of an inefficient search strategy, due to a not optimal arrangement of windows and interface features. In the same study, the authors also observed the so-called “left-to-right bias”, i.e., the fact that subjects start to explore a web page starting from the top left area, a phenomenon that was also found in other studies (Chatterjee, Southwood & Basilico, 1999). However, this bias depends on the culture of individuals. In fact, in Eastern cultures, which have a “right to left” reading system, the opposite phenomenon is observed, so that the exploration of the page starts from the top right area (Tylén et al., 2010). These phenomena give clear indications on how to organize the most important contents in a web page according to the culture of origin of the users.

A research published in 2005 by Pool and colleagues (Pool, Ball & Phillips, 2005) reports a study in which the authors involved 30 participants in a visual research task. The participants had the assignment to identify among different navigation paths the one related to the web page previously shown by the investigator.

The methodology included six different experimental conditions in which the information architecture (top-down and bottom-up) and the number of elements of the navigation path (one, two or three elements) were manipulated. The paths with top-down structure first had the page name and then the more specific elements (e.g.: website name/section name/content name). Conversely, paths with a bottom-up structure would have the content name first and then the more generic elements (e.g.: content name/section name/website name). The results showed no significant differences in fixation times between the navigation paths organized according to a top-down logic and the navigation paths organized according to a bottom-up logic. However, it was found that the target navigation paths received more fixations than the navigation paths considered distractors. This emphasizes that the contents of great importance for the achievement of a task can be associated with more fixations by the individual.

Other empirical evidence validating the use of eye movements as usability indices can be found in the work of Cowen and collaborators (Cowen, Ball, & Delin, 2002). They involved 70 subjects in the execution of two tasks (searching for information, buying an item online) on four different websites, characterized by a different degree of usability (due to a different organization and arrangement of content). Following the analysis of eye movements it emerged how the distribution of the fixations and the width and direction of the saccades can be used as usability indices. In fact, the websites with a greater degree of usability, also confirmed by a better performance obtained by the subjects in the assigned tasks, showed a greater grouping of fixations in certain areas of the page, indicating, following the authors, a more efficient search strategy. On the contrary, when browsing the less usable websites, the subjects showed a more dispersed pattern of fixations and a greater number of sudden changes of direction between one fixation and the next, probably an indicator of confusion and loss of the user, as the arrangement of the interface elements does not meet his expectations.

The relationship between usability and eye behavior was also verified in a study conducted by Habuchi, Kitajima and Takeuchi (2008). The authors asked 11 participants to perform information search tasks within four different websites. The websites were characterized by a different complexity of the information architecture and a different degree of usability due to the ambiguity of menu labels and to the hyperlinks operation. The results showed that websites characterized by a lower degree of usability were associated with a greater number and longer duration of fixations compared to those recorded during the exploration of the most usable websites.

Again, Ehmke and Wilson (2007), in a study involving 19 participants, observed that some parameters of eye movements are related to specific usability problems. In their study, they assigned participants some information search tasks to be performed at two different websites. During the performance of the tasks, a number of usability problems emerged at both of the websites analyzed. The work of the authors was to verify, after the recordings, the eye movements associated with the detected usability problems. For ease of reference, the results of the study by Ehmke and Wilson (2007) are reported in Table 3.3.

Table 3.3. Usability issues and related eye patterns: Ehmke and Wilson's study results (2007).

Usability Problem	Consequences for the user	Eye movements
Missing or difficult to find information	The user does not find the expected content within the page he has decided to visit.	● No fixation on "target" elements; ● Short-term fixations; ● Increased scanpath dispersion.
Confused or ambiguous functional elements	The user does not distinguish the functions of different elements of the web page (external/internal links).	● Short-term fixations; ● Increased scanpath dispersion.
Information overload	The user perceives that the page has too much content and experiences difficulties in recognizing and reading all of it.	● More fixations; ● Short-term fixations, ● Greater saccadic amplitude.
Poor visibility of the system status (failure to recognize the error)	The user does not perceive that the system has reported an error.	● No fixation on the "target" element; ● Longer fixations on other interface elements.
Mode of interaction with the system unclear	The input fields are not clearly labeled to support the user in entering data. The user experiences confusion.	● No fixation on "target" elements; ● Short term fixations; ● Increased scanpath dispersion.
Poor correspondence between the organization of functions and the mental model of the user	The user experiences difficulties in locating some functions due to the lack of logic in the grouping of interface elements.	● More fixations; ● Increased scanpath dispersion; ● No fixations on the interface area where an unclear grouping is presented.
Poor correspondence between the language used and the mental model of the user	The user, although he has reached the "target" web page, does not recognize the meaning of the information reported and therefore abandons the web page.	● Short fixations on the "target" information. ● More regressions on unclear elements; ● Greater dispersion of the scanpath;

More recently, Wang's working group (Wang et al., 2018) conducted a study in which subjective evaluations (expressed through questionnaires) and objective usability assessments derived from the analysis of the performance and eye movements of 35 university students were compared. In particular, participants were assigned seven different tasks (search for information, perform login operations, fill in forms) to be carried out within a website. The tasks assessed as the most difficult by the participants (through a self-assessment questionnaire) were found to be associated with a lower success rate and a longer task completion time. The analysis of the ocular behavior of the subjects revealed patterns of visual exploration characterized by a greater number of fixations and a longer duration of fixations for the tasks perceived more difficult by the subjects.

In summary, the various studies examined found a correlation between users' eye movements and perceived usability. However, it is important to emphasize that the choice and interpretation of metrics should be flexible. In fact, it always depends on the researcher's goals. If the goal is to understand how a banner ad can capture the attention of future customers, more fixations on it (and a longer duration of fixations) will generally be a good thing; conversely, if the goal of the study is to investigate the findability of a specific content on a web page, more fixations will be associated with poor usability of the website, as it is likely to indicate greater difficulty in understanding the terminology used or greater complexity of the research. Table 3.4. summarises the main metrics used and their interpretation in usability evaluations.

3.4 Eye tracking and mental workload

Cognitive processes such as reading, visual search, and problem solving can be studied based on the analysis of individuals' eye behavior (Kahneman, Beatty, & Pollack, 1967; Maier et al., 2014; Rayner, K. & Pollatsek, A., 1989; Rayner, 1998).

Many studies have previously researched the relationship between MWL and eye movements. The most used ocular metrics in MWL research generally refer to blink rate, changes in pupil diameter, saccadic amplitude, duration of fixations and scanpath analysis.

Blink rate

Several studies have tried to identify the relationship between eye blinks and mental workload. In the medical field is defined "blink" a *"rapid and momentary closure of the eyelids, voluntary or involuntary, which occurs as a reaction to a certain stimulus or in order to cleanse the conjunctival portion of the ocular* *globe"* (Knop et al., 2011).

Table 3.4. Ocular metrics in usability evaluations.

Metric		Interpretation in usability studies
Fixations	**Total fixation number**	The total number of fixations is inversely related to the efficiency of information search within a website (Goldberg & Kotval, 1999; Kotval & Goldberg, 1998; Ehmke & Wilson, 2007; Habuchi, Kitajima & Takeuchi, 2008; Wang et al., 2018). A higher number of fixations indicates a less efficient search due, probably, to a wrong arrangement of the interface elements.
	Average fixation duration	Longer duration of fixations generally indicates difficulties in processing information (Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Habuchi, Kitajima & Takeuchi, 2008; Wang et al., 2018).
	Fixation rate	The frequency of fixation on a certain element (Area of Interest - AOI) of the interface reflects its importance to the user. Important elements are fixed more frequently by the user (Fitts, Jones & Milton, 1950; Poole et al., 2005; Ehmke & Wilson, 2007).
	Spatial density of fixation	A greater grouping of fixations in certain areas of the interface could indicate a more efficient search strategy. In contrast, a more dispersed fixation pattern is associated with an ineffective search strategy, and is a potential indicator of user confusion and loss, as the arrangement of interface elements does not match the user's expectations (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007).
	Time to 1st fixation	This measure refers to the amount of time elapsed before the user performs a fixation on a target area. It is used to estimate the difficulty in finding a certain information or functionality within an interface (Byrne et al., 1999; Goldberg, 2003).
Saccades	**Total saccades number**	The total number of saccades is negatively correlated with the efficiency of information search (Goldberg & Kotval, 1999; Kotval & Goldberg, 1998). A higher number of saccades indicates an inefficient search strategy due, probably, to an incorrect arrangement of the interface elements.
	Average saccadic amplitude	A wider range of saccades indicates difficulties and less efficient search strategies probably due to an incorrect arrangement of the interface elements (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007; Fitts, Jones & Milton, 1950; Goldberg et al., 2002).
	Rapid changes in saccades direction	A change greater than 90° from the previous saccade is an indicator of a rapid change of direction in the exploration of the visual scene. Such a change could be indicative of a sudden change of "target" by the user, or a poor optimization of the interface components that do not reflect the user's mental model and expectations (Cowen, Ball, & Delin, 2002).
	Regressions	During the reading of a text, "regressions" are saccades which are directed towards parts of the text already read.A higher number of regressions is indicative of difficulties in text processing (Rayner & Pollatsek, 1989). In usability studies the presence of regressions may show a poor correspondence between the language used and the mental model of the user (Ehmke & Wilson, 2007).
Scanpath	**Scanpath amplitude**	A larger scanpath amplitude may indicate an inefficient search strategy due, probably, to a wrong arrangement of the interface elements (Goldberg et al., 2002).
	Spatial density of the scanpath	A low spatial density of the scanpath is associated with efficient search strategies (Goldberg & Kotval, 1999).
	Scanpath regularity	A high irregularity of the scanpath is indicative of inefficient search strategies probably due to a wrong arrangement of the interface elements (Goldberg & Kotval, 1999).
	Transition between different Areas of interest	Similar scanpaths, in terms of spatial width and density, may vary depending on the exploration path followed by a user. Generally, transitions between nearby areas of interest indicate an efficient arrangement of information, while transitions between distant areas of interest indicate an incorrect arrangement of

		interface elements (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007; Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Poole et al., 2005).

There are three basic types of blinks: reflexes, voluntary and endogenous. Endogenous blinks are distinguished from other types of blink because their occurrence is not related to any "determinant" external stimulus. For this reason blinks can be used as indicators of the mental workload induced by external elements, as the performance of a specific task (Stern, Walrath & Goldstein, 1984; Stern, Boyer & Schroeder, 1994).

Human factor research has focused on the study of different metrics derived from endogenous blinks, such as the blink rate, the blink amplitude and the blink closure duration.

Morris and Miller (1996) conducted an experimental study with 10 pilots. During the study the researchers recorded the blink rate, their total duration and blink closure duration in relation to changes in performance, fatigue and MWL. Specifically, the task assigned to the subjects consisted of a flight simulation lasting 4 hours and 30 minutes, without breaks. During the execution of the task the participants were required to perform several flight maneuvers characterized by a different level of complexity and a different degree of attention demand. The authors observed an increase in workload and fatigue in relation to the passage of time, substantiated by the execution of a greater number of errors. With regard to the ocular behavior of the subjects, the authors observed that an increase in the frequency of error, and therefore an increase in mental workload and fatigue, was associated with a decrease in the amplitude of blinks and a longer eyelid closing time (i.e. an eyelid closing time greater than 500 ms is an indicator of fatigue).

Another study, conducted by di di Veltman and Gaillard (1996), analyzed the eye movements of 14 subjects during a flight simulation. During the experimental session the mental workload was manipulated through the introduction of a secondary task (an acoustic recognition task, the Continuous Auditory Memory Task, CMT) and through the introduction of subtasks of different complexity. The flight simulation was organized according to five moments: 1) rest; 2) flight; 3) flight and secondary task; 4) landing; 5) post-landing. The results showed how during the "landing" phase (characterized by a higher mental workload) the blink rate was lower compared to the other phases. The blink duration was instead

greater in the “rest” phase, characterized by a lower mental workload experienced by the subjects.
Again, Veltman and Gaillard (1998) in a similar study involved 12 participants who were assigned two flight simulation tasks characterized by different levels of difficulty. The first task presented a single condition “to pursue a target stimulus”, and was considered easy; the second task was instead constituted by four different levels of increasing difficulty. The results have confirmed the existence of an inverse relationship between blink rate and mental workload.

Similar results emerged from a study conducted by Backs, Ryan and Wilson (1994). The authors recorded several physiological parameters, including eye movements, of 12 participants during a monitoring task. The experimental setting included different cognitive load conditions based on the variation in the number and type of the target stimuli and on the presence/absence of disturbance signals during the task. At the end of the experimental sessions the researchers found a decrease in the blinks rate linked to a greater cognitive load experienced by the participants, also confirmed by the analysis of cardiac and respiratory activity.

Hankins and Wilson (1998) involved 10 pilots in an experiment characterized by three different flight conditions: in the first condition, called Visual Flight Rules (VFR), pilots could use visual information outside the cockpit to determine the altitude and verify their position; in the second condition, called Instrument Flight Rules (IFR), pilots had no external visual information and could use only the information provided by the cockpit instruments; the third condition consisted of a high-speed IFR flight segment. The IFR condition imposed a higher cognitive load on the participants, as it required a higher investment of resources in monitoring and interpreting the status of the system. In this condition in fact the subjects could not have feedback from the environment outside the cockpit.
The analysis of the ocular movements of the subjects showed a variation of the blink rate depending on the experimental condition. In fact, IFR conditions showed a significant reduction of blinks rate compared to VFR condition.

In summary, although these studies have shown a clear relationship between blinks and mental workload, this metric can only be used for an assessment of the workload as a whole. It has in fact a poor diagnostic ability to detect the different types (cognitive, physical, temporal) of mental workload experienced by individuals.

Pupillary diameter

The pupil is located in the center of the eye and has the function of optimizing the amount of light that reaches the retina allowing a detailed perception of stimuli. For this reason the pupil diameter varies according to different elements such as the external lighting conditions and the distance of visual stimuli. The technique of measuring variations in pupil diameter is commonly known as "*pupillometry*".

Several studies have observed an association between increased pupil diameter and increased mental workload under controlled light conditions (Hess & Polt, 1964; Juris & Velden., 1997; Marshall, 2002; Nakayama et al., 2002, Takahashi, Rayman & Dynlacht, 2000).

Kahneman (1973), Beatty and Lucero-Wagoner (2000) found that variations in pupil diameter due to cognitive processing (variations up to 0.5 mm of pupil diameter) are significantly different from the variations that occur in response to changes in luminosity (variations from 2 to 8 mm of pupil diameter). The variations due to cognitive processing of a stimulus have been defined as "task-evoked pupillary response" (TEPR). Arithmetic tasks, verbal comprehension tasks, mnemonic tasks, vigilance tasks, or visual perception tasks are all tasks that involve cognitive processing and that cause a variation in pupil diameter (Beatty & Lucero-Wagoner, 2000).

A limitation in the use of this metric is that TEPRs do not occur consistently and homogeneously. For this reason it is necessary to record and calculate an average of multiple TEPRs to obtain a reliable estimate of the mental workload.

Hess and Polt (1964) were among the first researchers to include the measurement of pupil diameter as an index of cognitive load. The authors involved 5 participants in the execution of some logical and arithmetic tasks characterized by different levels of difficulty. The results showed that during the most difficult tasks there was an increase in pupil diameter of the participants.

Bradshaw (1968) involved 6 subjects in a memory task characterized by two levels of difficulty: in the “easy” condition the subjects had to remember if the first and the last element of a string of letters were equal after having seen it for few seconds; in the “hard” condition the succession of the stimuli was faster and there was a greater quantity of letters for each string. In addition, participants had to remember if more “chunks” of letters were equal to each other. Analyzing the ocular data, the author observed an increase in the pupil diameter of the participants related to the increase in the complexity of the task and the frequency of presentation of the stimuli.

Recarte and Nunes (2002, 2003) investigated the variations in the mental workload of 24 subjects during a driving task. The authors manipulated the difficulty of the task by introducing several secondary, visual and auditory tasks. The results of their studies showed that an increase in workload was related to a significant increase in pupil diameter.

Iqbal and collaborators (2004, 2005) conducted several studies involving participants in planning, reading and comprehension tasks, mathematical reasoning and visual research. Each task included an “easy” and a “hard” condition. The authors recorded several measures such as task completion time, percentage change in pupil size (PCPS), average percentage change in pupil size (APCPS) and subjective evaluations of perceived difficulty. The results showed how the percentage variation in pupil size (calculated by dividing the difference between the pupil size recorded at a specific time and the baseline pupil size by the reference pupil size) is positively correlated to the MWL experienced by the subjects during the execution of the task.

Palinko’s working group (2010) conducted an experiment involving 32 participants in a driving simulation. The mental workload of the participants was manipulated by introducing a secondary task (e.g. talking to a passenger). The authors have introduced a new measure for the calculation of pupil diameter changes, namely the “mean pupil diameter change rate” (MPDCR). This metric is useful to monitor pupil changes along a continuum and then to compare different time series during the experiment. The results showed that in situations of high mental workload, where participants were asked to speak or think, there was a greater dilation of the pupil and a greater mean pupil diameter change.

A similar study was conducted more recently by Kun and collaborators (2013). The authors investigated the mental workload in relation to a driving task, the difficulty of which was manipulated through the introduction of a conversation task with a computer. Measurements on pupil diameter variations, specifically the TEPR, were compared for two different moments of the task: 1) just before the computer expressed a sentence and 2) just before the participant's response. It was found that the pupil diameter was significantly larger at time 2 than at time 1 in 69% of the conversations, but this effect was weak. The authors explain that in some cases the TEPR may have been influenced by the pupillary reflection of light, emphasizing the negative effects of light reflections for the validity of the pupil diameter measurement.

Saccadic movements

The function of saccades is to convey the elements of interest to the foveal area of the retina. Saccades last on average between 30 and 80 milliseconds and rarely proceed from one point to another along the shortest possible segment. They can in fact follow different shapes or curves. The speed of the saccades is proportional to the amplitude of eye movement, is measured in degrees and oscillates between 300 and 500 degrees per second. Saccades and visual attention are closely related to each other (Hoffman, 1998; Pashler & Sutherland, 1998). In the study of mental workload, the saccades were mostly analyzed in piloting or air traffic control tasks.

Krebs and collaborators (Krebs, Wingert & Cunningham, 1977), during a flight simulation task, found a decrease in the saccadic amplitude of the participants in relation to an increase in mental workload (manipulated through the introduction of turbulence phases). The authors' interpretation is that under high mental workload conditions the participants focus on a restricted portion of the cockpit interface, decreasing the visual exploration.

In another study based on a flight simulation task, Katoh (1997) found that the saccadic amplitude varied according to the type of activity: when the participant had to concentrate exclusively on the instruments available in the cockpit, the saccades had a smaller amplitude; on the contrary, when the participant also used context elements to perform the task, the saccades were wider.

The working group of May (May et al., 1990) conducted four different experiments in which a total of 35 subjects were involved in visual exploration tasks characterized by three different levels of difficulty: low, medium, high. The results of the studies converge in finding an inversely proportional relationship between saccadic amplitude and mental workload, also supported by data on the performance of the participants. Specifically, the participants' saccadic amplitude was significantly lower in the high mental workload conditions.

Some innovative studies conducted by Di Stasi and collaborators (Di Stasi et al., 2009; Di Stasi et al., 2010a, 2010b; Di Stasi et al., 2011; Di Stasi et al., 2016), investigated the relationship between peak velocity (PV) of saccades and mental workload. The theoretical assumption of their experiments is that, although the PV and the duration of saccades increase as the amplitude of the saccade increases (Bahill, Clark & Stark, 1975), the PV is independent of the saccadic duration (Becker, 1989). This allows the use of PV as an independent metric in cognitive studies.

These authors have found in their experiments that tasks characterized by a high mental workload (manipulated both in relation to visual-spatial aspects and effort imposed on subjects) are associated with a significant reduction in the speed of saccadic movements.

Fixations

The fixations constitute the only moment in which the stimuli present in the visual field of the individual are able to reach the fovea. For this reason the function of fixations is to allow a detailed view of visual stimuli. The fixations can last, on average, from 60 milliseconds to 300 milliseconds, however, it is also possible to find fixations with a longer duration.

The main metrics derived from fixations refer to their number, duration and frequency.

Rayner's review work (1998) reports various evidence that correlates the duration

of fixations with the difficulty of processing and understanding texts or solving arithmetic problems. Some researchers (Findlay & Kapoula, 1992; Moffitt, 1980) have found a relationship between the duration of fixations and the difficulty of visual processing.
The studies of Fitts and colleagues (Fitts, Jones & Milton, 1950), previously listed in relation to usability assessments, can be considered the first studies that associated the individuals' eye behavior to their mental workload, defining some criteria that are still valid today. In his experiments Fitts found that there was a positive correlation between the number and duration of fixations and the difficulty of processing and the mental workload experienced by participants.

In a study by Tole and collaborators (Tole et al., 1982) it was found that the ocular behavior of a sample of participants engaged in flight simulation tasks was influenced by the introduction of an auditory-verbal secondary task. More specifically, the addition of the secondary task, related to an increase in mental workload, led to an increase in the duration of the participants' fixations.

Callan (1998) compared the eye movements of 16 pilots engaged in a flight simulation characterized by "reduced", "normal" and "high" mental workload segments. The author noted that the high mental workload flight segments resulted in a decrease in the performance of the participants associated with longer duration and more fixations.

Goldberg and Kotval (1999) found similar results in a study comparing the eye patterns of 12 participants. The participants performed tasks under two conditions: one condition required the use of "optimized" software according to logical criteria of functionality grouping; the other condition required the use of an "not optimized" version of the software. The results showed that subjects performed more fixations in the not optimized version of the interface, which constrained them to review their exploration strategies.

Harbluk, Noy and Eizenman (2002), evaluated the impact of secondary tasks on the behavior of 21 participants in a road driving experiment. Participants were asked to drive for eight hours in an urban context, during which they were asked

to perform some secondary tasks of arithmetic reasoning with a different level of difficulty. Each participant was subjected to three experimental conditions: 1) driving without a secondary task; 2) driving with an easy secondary task; 3) driving with a difficult secondary task.
At the end of the experimental sessions the authors compared the ocular data with a subjective measure of mental workload, i.e. the subjects' answers to the NASA-TLX questionnaire. The data analysis showed a decrease in saccades, and a longer duration of fixations under conditions of high mental workload, a data confirmed also by the subjective perceptions of the participants.

The nature of the secondary task (auditory-verbal vs. visuospatial) affects an individual's eye movements differently. Recarte and Nunes (2000) compared the effects caused by the introduction of secondary tasks of different nature on the eye patterns of 12 individuals engaged in a road driving task. Although the results showed that the addition of a secondary task is related to an increase in mental workload, different effects on the duration of the fixations emerge depending on the nature of the task. Secondary tasks of a visual-spatial nature have in fact led to an increase in the duration of the fixations in conditions of high mental workload. However, this did not occur when the nature of secondary tasks was auditory-verbal.

Van Orden and colleagues (Van Orden et al., 2001) found changes in individual eye behavior related to manipulations of task difficulty and mental workload. They recorded the eye movements of 11 participants engaged in a visual-spatial task whose difficulty was manipulated based on the number of target stimuli to which participants had to pay attention during the experiment. The results showed no significant differences with regard to the average duration of fixations in different experimental conditions, however it was found an increase in the frequency of "long" fixations (more than 500 ms) in the most difficult condition, characterized by the simultaneous presence of multiple target stimuli in the participants' visual field.

De Greef and collaborators (de Greef et al., 2009) recorded the eye movements of 18 individuals involved in a monitoring task characterized by three different levels of mental workload: 1) underload; 2) normal-load; 3) overload. The results showed a significant and discriminatory effect of mental workload manipulation on the participants fixation time. Specifically, in the overload condition the authors found an increase in the duration of the fixations and a decrease in the performance of the participants.

Scanpath

As described above, the whole sequence of saccades and fixations recorded during a visual exploration task is called “scanpath”. In the ergonomic field it is generally analyzed both from a qualitative point of view (for example, by analyzing the area, extension or shape) and from a quantitative point of view (by means of a mathematical analysis) in order to evaluate the mental workload of an individual. Another common distinction in the use of this metric is the technique used for its interpretation. In fact, there is a distinction between techniques that analyze the scanpath in relation to the transitions that occur between different areas of interest (Fitts, Jones & Milton, 1950; Tole et al., 1983), and techniques that analyze the scanpath globally, referring to the whole visual field of the individual (Di Nocera, Camilli & Terenzi, 2007).

The pioneering study of Fitts (Fitts, Jones & Milton, 1950) carried out in the aviation field also obtained interesting results with regard to transitions between areas of interest. In their study the researchers found that transitions between non-contiguous and distant areas of interest were associated with less efficient research strategies and greater cognitive load of the participants.

The research groups of Tole (Tole et al., 1983) and Harris (Harris, Glover & Spady Jr, 1986) conducted two different studies in the aerospace sector introducing an innovative technique for the analysis of transitions between areas of interest. They introduced the concept of “entropy” in the analysis of eye movements. In particular, this concept applied to visual exploration strategies indicates the degree of randomness that is recorded in the succession of saccades and fixations between two or more areas of interest. In fact, according to the authors, the degree of stereotypicality of this sequence may vary depending on the cognitive load experienced by the individual. The studies of Tole and Harris have observed the ocular behavior of some pilots engaged in monitoring tasks characterized by a variable level of cognitive demands imposed on the participants. The authors have manipulated the demands imposed by the task by introducing secondary tasks of auditory-verbal nature, thus creating "low workload" and "high workload" conditions. In the high workload conditions the authors found a decrease in performance and an increased mental workload. With regard to the analysis of the participants' eye movements, Tole and Harris first of all attributed different areas of interest to the different instruments of the cockpit. Afterwards, they observed the visual exploration strategy used by the participants to move from one area of interest to another during the experimental session. The results showed that the participants' scanpath tended to disorder in low workload conditions, while it became more stereotypical (i.e. less random) in high mental workload conditions.

The Tole and Harris studies, however, have several limitations. One of them is that the effects observed cannot be generalized to all types of tasks and domains, such as the specificity of the application field (aerospace). Moreover, results of more recent studies (Kruizinga, Mulder & de Waard, 2006) that have applied the concept of entropy to the analysis of eye movements have found a diametrically opposed pattern, highlighting the need for further research.

Di Nocera's research group Di Nocera, Camilli & Terenzi, 2007; Di Nocera, Ranvaud & Pasquali, 2015) introduced the use of an algorithm for the analysis of the scanpath called Nearest Neighbour Index (NNI).

The Nearest Neighbour Index (NNI) is an algorithm developed by Clark and Evans (1954) in the geostatistic field. NNI provides information about the average distance among points and about their spatial distribution. Di Nocera's working group (Di Nocera, Camilli & Terenzi, 2006; Di Nocera, Camilli & Terenzi, 2007) applied this algorithm to eye movements analysis and proposed a global interpretation of scanpath as indicator of MWL. Thanks to NNI it is possible to compare the average distance among the fixations that an individual has done during the execution of a specific task. The result is expressed by a single value that can vary between 0 (maximum clustering) and 2.1491 (strictly regular hexagonal pattern). NNI values close to 1 indicate that the distribution of fixations is not different from a random distribution, NNI values greater than 1 indicate a dispersion of the fixation pattern while NNI values less than 1 indicates a grouping of fixations. The NNI can be estimated for very small periods (1 minute), providing a continuous measurement (time series) of user behavior.

To estimate the index the first step is to calculate the nearest neighbor distance or d(NN):

$$d(NN) = \sum_{i=1}^{N} \left[min \frac{(dij)}{N} \right]$$

where min (dij) is the distance between each point and the nearest point and N is the number of points in the distribution.

The second element of the equation is obtained by calculating the average random distance or d(ran); this value would correspond to the value of d(NN) if the distribution of points were completely random:

$$d(ran) = 0.5\sqrt{\frac{A}{N}}$$

where A is the area of the polygon defined by the most extreme fixations and N is the number of points. Finally, the NNI value is calculated by dividing the nearest neighbor distance, d(NN), for the average random distance, d(ran):

$$NNI = \frac{d(NN)}{d(ran)}$$

The validity of this algorithm as a measure of mental workload was confirmed in a methodological study (Camilli, Terenzi & Di Nocera, 2007) that showed a consistency of the NNI with both subjective (NASA-TLX score) and physiological (amplitude of the P300 component of event-related potentials) measures. One of the advantages was the possibility of providing “online” information that can not be obtained otherwise. Furthermore, Camilli, Terenzi & Di Nocera (2008) also demonstrated the diagnostic sensitivity of NNI. In fact, depending on the type of the task demand, it is possible to expect differential effects on the NNI: while an increase in the visuo-spatial demand determines a clusterization of the fixations pattern (NNI values are therefore less than 1), an increase in the temporal demand determines a greater dispersion of fixations (NNI values are therefore greater than 1). This difference in the distribution of fixations pattern can be explained, at functional and behavioural level, with the need to maximize the stimuli detection when a task impose an high temporal request to the user and with the need to increase the visuo-spatial resources involvement when the task is characterized by high complexity of visual and spatial elements.

To conclude, the various studies examined show how it is possible to derive information about the mental workload perceived by an individual based on the analysis of his eye movements. However, although these metrics have the undisputed advantage of measuring MWL in real time and without changing the nature of the task, researchers must try to mitigate the various technical problems related to the recording of eye movements (sampling rate of recording devices, problems related to the brightness of environments or stimuli, etc.). Table 3.5. summarises the main metrics used and their interpretation in mental workload assessments.

3.5 Conclusion

This chapter describes the main ocular metrics used for usability and mental workload assessments. The two constructs play a fundamental role in the design and evaluation of the interaction between an individual and a system.

The pervasive diffusion of digital technologies has been accompanied by the birth of a new scientific discipline known as “Human-Computer Interaction” (HCI). This discipline is characterized by transversality (it involves several fields: medicine and health care, security systems, control systems, automotive, communications, etc.) and interdisciplinarity (it is increased by the contribution of several disciplines: engineering, computer science, cognitive psychology, sociology, etc.). The main objective of HCI is to improve human-machine interaction through the implementation of systems that respect the real characteristics of the end user.

Table 3.5. Ocular metrics in MWL evaluations.

Metric		Interpretation in mental workload studies
Eye blinks	**Blink rate**	Blink rate is inversely related to user mental workload (Backs et al., 1994; Brooking et al., 1996; Hankins & Wilson 1998; Veltman & Gaillard, 1996,1998)
	Blink duration	A shorter blink duration (i.e., longer eyelid closing time) is associated with a greater mental workload, a decrease in the amplitude of blinks and a longer blink closure duration (Morris & Miller, 1996; Veltman & Gaillard, 1996).
	Blink amplitude	A decrease in the amplitude of winks indicates a greater mental workload (Morris & Miller, 1996).
Pupillary diameter	**Pupillary diameter variation**	An increase in pupil diameter variation (recorded under controlled light conditions) is positively correlated with an increase in mental workload (Hess & Polt, 1964; Bradshaw, 1968; Iqbal et al., 2004, 2005; Juris et al., 1977; Kun et al., 2013; Nakayama et al., 2002, Palinko, 2010; Recarte & Nunes, 2002, 2003; Takahashi et al., 2000).
Saccades	**Average saccadic amplitude**	The amplitude of saccadic movements decreases as the mental workload increases (Katoh, 1997; Krebs et al., 1977; May et al., 1990).
	Saccades velocity	The velocity of saccadic movements (Peak Velocity) decreases as the mental workload experienced by the individual increases (Di Stasi et al., 2009; Di Stasi et al., 2010a, 2010b; Di Stasi et al., 2011; Di Stasi et al., 2016).
Fixations	**Total fixation number**	A greater number of fixations is indicative of greater processing difficulty and mental workload (Callan, 1998; Findlay & Kapoula, 1992; Fitts et al., 1950; Goldberg & Kotval, 1999; Moffitt ,1980).
	Average fixation duration	A longer duration of fixations is commonly associated with an increased mental workload (Bunecke, 1987; Callan, 1998; De Greef et al., 2009; Ephrath et al., 1980; Fitts et al., 1950; Harbluk & Noy, 2002; Tole et al., 1982).
	Fixation rate	The frequency of fixation on a specific component (AOI) of the interface reflects its importance to the user. Important elements for the execution of the task are generally fixed more frequently than non-important elements (Fitts et al., 1950). An increase in the frequency of long-term fixations is associated with an increase in mental workload (Van Orden et al., 2001).
Scanpath	**Transition between AOIs**	Transitions between non-contiguous and distant areas of interest are associated with less efficient search strategies and increased mental workload (Fitts et al., 1950; Goldberg & Kotval, 1999).
	Entropy	More stereotypical exploration strategies are indicative of high mental workload (Harris et al., 1986; Tole et al., 1983).
	Nearest neighbor Index	The pattern of eye movements, under conditions of high mental workload, is distributed differently depending on the nature of the task: an increase in visual-spatial demand determines a concentration of the fixation pattern - lower values of the NNI; an increase in temporal demand determines a greater dispersion of the scanpath - higher values of the NNI (Camilli et al., 2007; Camilli et al., 2008; Di Nocera et al., 2007).

The design of human-machine interaction must be based on the principle of “reducing the cognitive load imposed on the user”, so as to allow users to use the “machine” to achieve their goals with effectiveness, efficiency and satisfaction. However, although many studies have been conducted on the cognitive processes involved in human-machine interaction, due to the different fields of application of mental workload (aviation, safety, etc.) and usability (communications, e-commerce, etc.), the two phenomena have been studied separately, leaving open the investigation on their relationship, interaction and integration.

Despite this apparent differentiation with regard to the areas of application, there are many points in common between usability and mental workload. There are similarities both in interface design and in the evaluation of the individual-interface interaction. An example is provided by the objective metrics used to assess mental workload and usability, such as measures related to the performance of the individual and measures derived from the analysis of eye behavior and individual visual exploration strategies. The analysis of eye movements is a promising technique that can return objective and real-time information regarding the interaction of a user with a system, allowing to reach conclusions both on the degree of usability of the system itself and on the mental workload experienced by the user during the use of the system.

The most commonly used metrics refer to the analysis of basic eye movements, saccades and fixations, or to the entire path of visual exploration (scanpath). In particular, in the studies examined, the interpretation of metrics such as the number and average duration of fixations, the amplitude of saccades and the analysis of transitions between certain areas of interest seems to go in the same direction both for usability aspects and for aspects related to the investment and saturation of individual cognitive resources. A practical example of the overlap between usability and mental workload can be found in the fact that, during the interaction between an individual and a system, more and longer fixations are associated with poor usability and high mental workload (Callan, 1998; Ehmke & Wilson, 2007; Findlay & Kapoula, 1992; Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Habuchi, Kitajima & Takeuchi, 2008; Kotval & Goldberg, 1998; Moffitt, 1980; Wang et al., 2018). Another example derives from the interpretation given to the analysis of transitions between areas of interest, as transitions between non-contiguous and distant areas of interest are associated with less efficient research strategies, poor perceived usability and increased mental workload (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007; Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Poole et al., 2005).

Table 3.6. Ocular metrics in usability and mental workload assessments.

Metric		Interpretation in usability studies	Interpretation in mental workload studies
Fixations	**Total fixation number**	The total number of fixations is inversely related to the efficiency of information search within a website (Goldberg & Kotval, 1999; Kotval & Goldberg, 1998; Ehmke & Wilson, 2007; Habuchi, Kitajima & Takeuchi, 2008; Wang et al., 2018). A higher number of fixations indicates a less efficient search due, probably, to a wrong arrangement of the interface elements.	A greater number of fixations is indicative of greater processing difficulty and mental workload (Callan, 1998; Findlay & Kapoula, 1992; Fitts et al., 1950; Goldberg & Kotval, 1999; Moffitt ,1980).
	Average fixation duration	Longer duration of fixations generally indicates difficulties in processing information (Fitts et al., 1950; Goldberg & Kotval, 1999; Habuchi, Kitajima & Takeuchi, 2008; Wang et al., 2018).	A longer duration of fixations is commonly associated with an increased mental workload (Bunecke, 1987; Callan, 1998; De Greef et al., 2009; Ephrath et al., 1980; Fitts et al., 1950; Harbluk & Noy, 2002; Tole et al., 1982).
Saccades	**Average saccadic amplitude**	A wider range of saccades indicates difficulties and less efficient search strategies probably due to an incorrect arrangement of the interface elements (Fitts et al., 1950; Goldberg et al., 2002; Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007).	The amplitude of saccadic movements decreases as the mental workload increases (Katoh, 1997; Krebs et al., 1977; May et al., 1990).
Scanpath	**Transition between AOIs**	Transitions between nearby areas of interest indicate an efficient arrangement of information, while transitions between distant areas of interest indicate an incorrect arrangement of interface elements (Fitts et al., 1950; Goldberg & Kotval, 1999; Cowen, Ball, & Delin, 2002; Poole et al., 2005; Ehmke & Wilson, 2007).	Transitions between non-contiguous and distant areas of interest are associated with less efficient search strategies and increased mental workload (Fitts et al., 1950; Goldberg & Kotval, 1999).

4. Experimental studies

4.1 Study 1

The objective of this study was to validate a new questionnaire for a quick and reliable measure of web usability, the “Simple Outlook on Usability & Promotion” (SOUP). This instrument is a short version of the questionnaire Usability System Evaluation 2.0 (Us.E. 2.0) proposed in 2013 by Di Nocera (Di Nocera, 2013). The SOUP summarises the 19 items of Us.E. 2.0 in only three items regarding handling, satisfaction, and attractiveness. The SOUP administration, scoring, and interpretation of scores follow the same procedures as NPS (Reichheld, 2003). The item structure of SOUP uses the “word-of-mouth” paradigm asking the user the probability with which he would recommend the use of a specific website to friends or colleagues for each usability dimension. The response scale used for the three items is the same used in the NPS (Reichheld, 2003), a Likert scale with 11 intervals, where 0 corresponds to “not at all likely” and 10 indicates “Extremely likely”.

Usability dimension	Item
Handling	With specific reference to the ease of browsing (for example, moving between pages of the website without getting lost, recognising hyperlinks, finding the info you were looking for), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Satisfaction	With specific reference to the satisfaction of your needs (for example, finding information, reaching your goals), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Attractiveness	With specific reference to the aesthetic features (for example, pleasantness of the graphics, colors, images), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely

Figure 4.1.1. The SOUP questionnaire.

As well as for the NPS (Reichheld, 2003), depending on the score that is given to the questions, three categories of people can be distinguished:

1. Promoters = respondents giving a 9 or 10 score;
2. Passives = respondents giving a 7 or 8 score;
3. Detractors = respondents giving a 0 to 6 score.

The SOUP aims to be a more straightforward tool compared with the Us.E. 2.0, with a simpler structure and scoring procedure. On the other side, it should be able to provide more detailed information compared to the NPS (Reichheld, 2003).

Methods and Materials

Sample

This study involved 866 volunteer participants among the teaching staff of Sapienza University of Rome (382 female subjects; mean age 54.1; dev. st. = 8.9; 484 male subjects; mean age 54.2; dev. st. = 8,9). Participants belonged to four different academic positions: “Associate Professor” (N = 356), “Assistant professor” (n = 272), “Full Professor” (N = 166), “Temporary assistant professor” (n = 72), as reported in table 4.1.1.

Table 4.1.1. Sample composition by academic position.

Academic positions	Sample size	%
Assistant professor	272	31.4%
Full professor	166	19.2%
Associate professor	356	41.1%
Temporary assistant professor	72	8.3%
Total	**866**	**100%**

Web platform

The management web platform “InfoStud” was used as experimental material. InfoStud is a management web platform used by students and teachers of Sapienza University of Rome to manage their careers’ data. In the present study, we focus on the interface used by professors to:

1. modify their personal information and access data;
2. consult data on university roles and projects;
3. download reports on their courses;

4. consult the number and personal details of students registered for examinations;
5. manage examinations sessions for active courses (verbalisation).

The 90.5% of participants reported the “verbalisation” function as the most used.

Usability questionnaires

Subjective measures of perceived usability were collected using the following scales:

- Net Promoter Score® (NPS: Reichheld, 2003; Reichheld & Covey, 2006);
- Usability Evaluation 2.0: (Us.E. 2.0: Di Nocera, 2013);
- System Usability Scale - SUS (Brooke, 1986)
- SOUP (Di Nocera et al., in press).

Procedure

Researchers asked participants to answer the different usability questionnaires based on their experience with the InfoStud platform. The questionnaires were remotely administered. The order of administration of the various scales was randomised.

Data Analyses and Results

Pearson’s r coefficient

The analysis of Pearson’s r coefficient showed significant correlations between the SOUP scores and the other questionnaires’ scores. Handling (H_SOUP), Satisfaction (S_SOUP), and Attractiveness (A_SOUP) scales were compared individually with the scores obtained in the SUS (Brooke, 1986), the NPS (Reichheld, 2003) and the Us.E. 2.0 questionnaires (in the corresponding dimensions; Di Nocera, 2013). The results confirmed positive correlations between the scales of the SOUP and the criterion variables, as summarised in Table 4.1.2.

*Table 4.1.2. Correlation matrix between variables (*p < .05).*

	H_Us.E. 2.0	S_Us.E. 2.0	A_Us.E. 2.0	SUS	NPS	H_SOUP	S_SOUP	A_SOUP
H_Us.E. 2.0		0.77*	0.33*	0.82*	0.70*	0.74*	0.70*	0.52*
S_Us.E. 2.0			0.26*	0.71*	0.66*	0.65*	0.67*	0.46*
A_Us.E. 2.0				0.29*	0.38*	0.40*	0.37*	0.64*
SUS					0.74*	0.75*	0.73*	0.52*
NPS						0.89*	0.87*	0.67*
H_SOUP							0.86*	0.68*
S_SOUP								0.65*
A_SOUP								

Variance Analysis (ANOVA)

The scores obtained by the participants in the SUS (Brooke, 1986), the NPS (Reichheld, 2003) and the Us.E. 2.0 (Di Nocera, 2013) questionnaires were used as dependent variables in different ANOVA designs using the categorical SOUP's subscales as the factor (Promoters, Passives, Detractors).

Comparison between SOUP and SUS

The results of the ANOVA conducted using the SUS questionnaire scores as a dependent variable showed a significant effect of the different categories of the SOUP in all dimensions. Specifically, with regard to the dimensions "Handling" (r = .76; $p < .05$; $F_{2,863} = 311,46$, $p < .01$, figure 4.1.2.), "Satisfaction" (r = .73; $p <. 05$; $F_{2,863} = 318,14$, $p < .01$, figure 4.1.3.) and "Attractiveness" (r = .52; $p < .05$; $F_{2,863} = 64.041$, $p < .01$, figure 4.1.4). The "promoters" category presented higher SUS scores. In comparison, the "detractors" category presented lower SUS scores. The post-hoc analysis performed with the Duncan test showed significant differences between all categories for all three SOUP's dimensions.

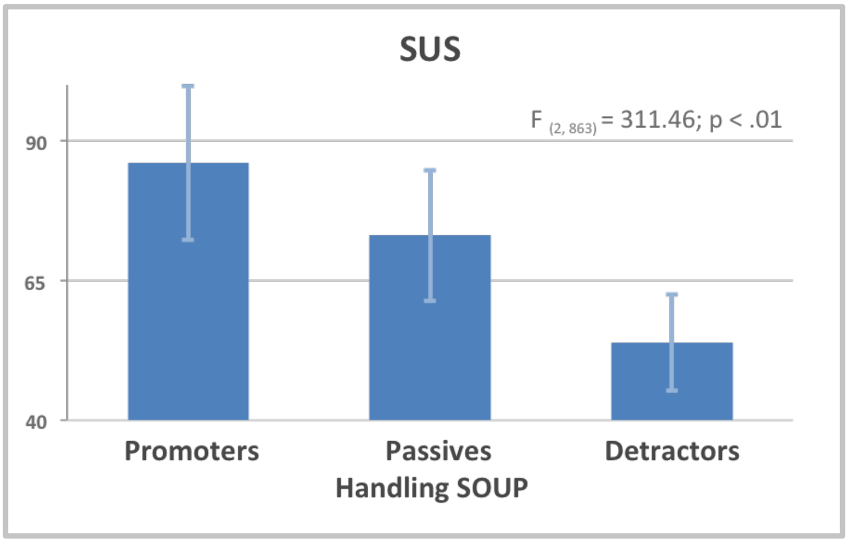


Figure 4.1.2. SUS scores per SOUP user categories - Handling.

The error bar denotes a confidence interval of 0.95.

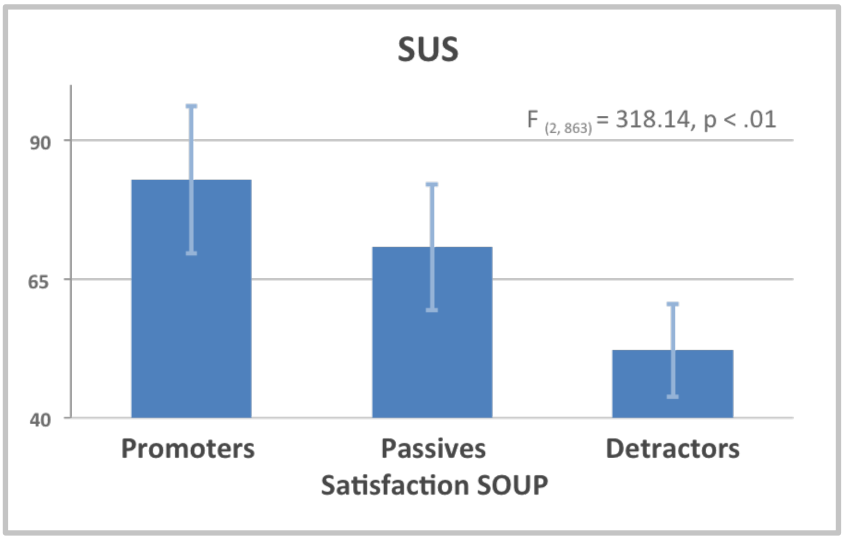


Figure 4.1.3. SUS scores per SOUP user categories - Satisfaction.

The error bar denotes a confidence interval of 0.95.

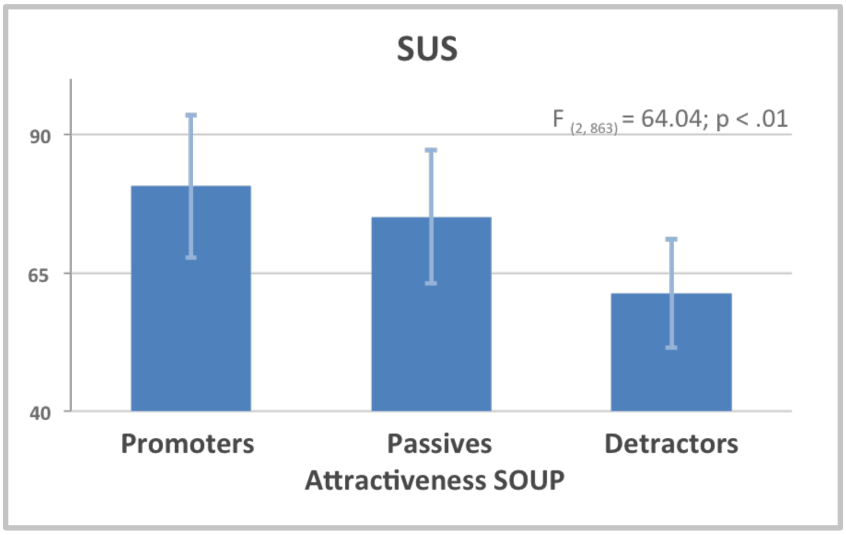


Figure 4.1.4. SUS scores per SOUP user categories - Attractiveness.
The error bar denotes a confidence interval of 0.95.

Comparison between SOUP and NPS

The results of ANOVA conducted using the NPS scores as a dependent variable showed a significant effect of the different categories of the SOUP in all dimensions. Specifically, with regard to the dimensions “Handling” (r = .89; p < .05; $F_{2,863}$ = 481.3, p < .01, figure 4.1.5.), “Satisfaction” (r = .87; p < .05; $F_{2,863}$ = 513.18, p < .01, figure 4.1.6.) and “Attractiveness” (r = .67; p < .05; $F_{2,863}$ = 100.19, p < .01, figure 4.1.7). The “promoters” category presented higher NPS scores. In comparison, the “detractors” category presented lower NPS scores. The post-hoc analysis performed with the Duncan tests showed significant differences between all categories for all three SOUP’s dimensions.

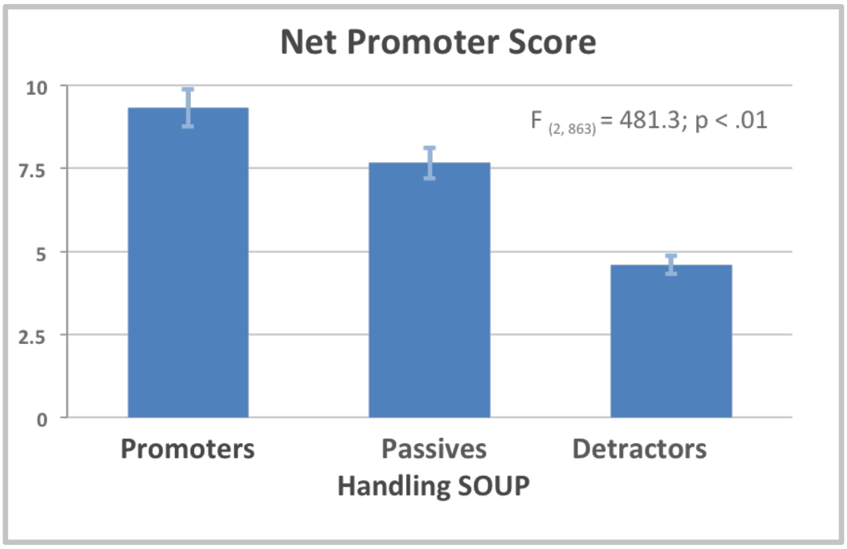


Figure 4.1.5. NPS scores per SOUP user categories - Handling.

The error bar denotes a confidence interval of 0.95.

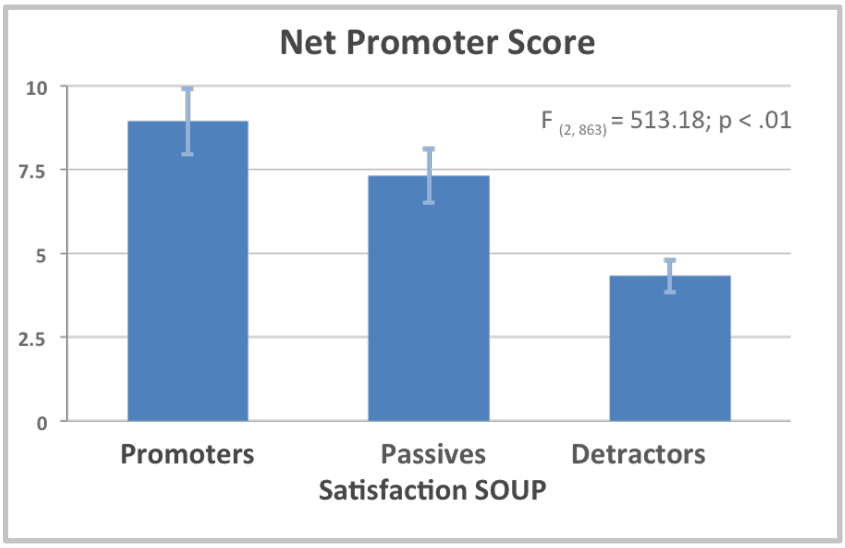


Figure 4.1.6. NPS scores per SOUP user categories - Satisfaction.

The error bar denotes a confidence interval of 0.95.

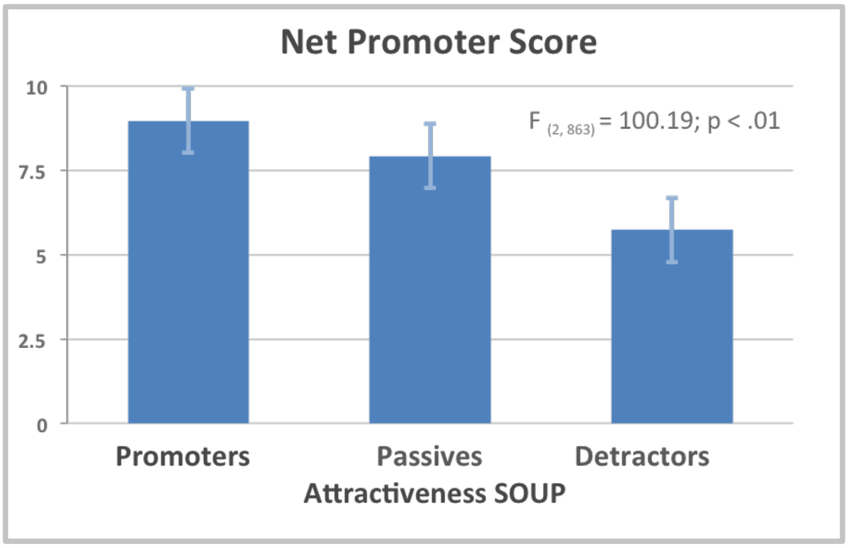


Figure 4.1.7. NPS scores per SOUP user categories - Attractiveness.
The error bar denotes a confidence interval of 0.95.

Comparison between SOUP and Us.E. 2.0

The results of ANOVA conducted using the size scores of the questionnaire Us.E. 2.0 as a dependent variable showed a significant effect of the different categories of the SOUP in all dimensions. Specifically, with regard to the dimensions "Handling" (r = .74; $p < .05$; $F_{2,863} = 324.28$, $p < .01$, figure 4.1.8.), "Satisfaction" (r = .67; $p < .05$; $F_{2,863} = 242.59$, $p < .01$, figure 4.1.9.) and "Attractiveness" (r = .64; $p < .05$; $F_{2,863} = 94.074$, $p < .01$, figure 4.1.10.). The "promoters" category presented higher Us.E. scores. In comparison, the "detractors" category presented lower Us.E. scores in the same usability dimension. The post-hoc analysis performed with the Duncan tests showed significant differences between all categories for all three SOUP's dimensions.

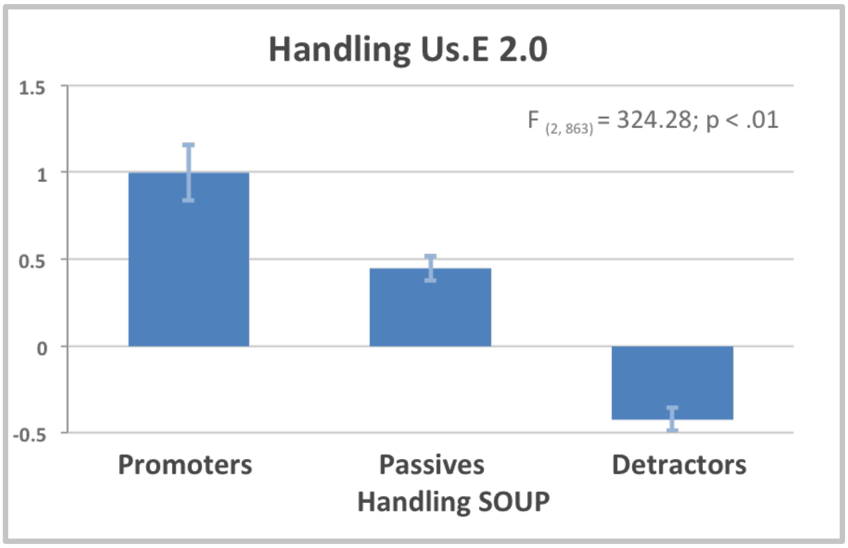


Figure 4.1.8. Us.E. 2.0 scores per SOUP user categories - Handling.

The error bar denotes a confidence interval of 0.95.

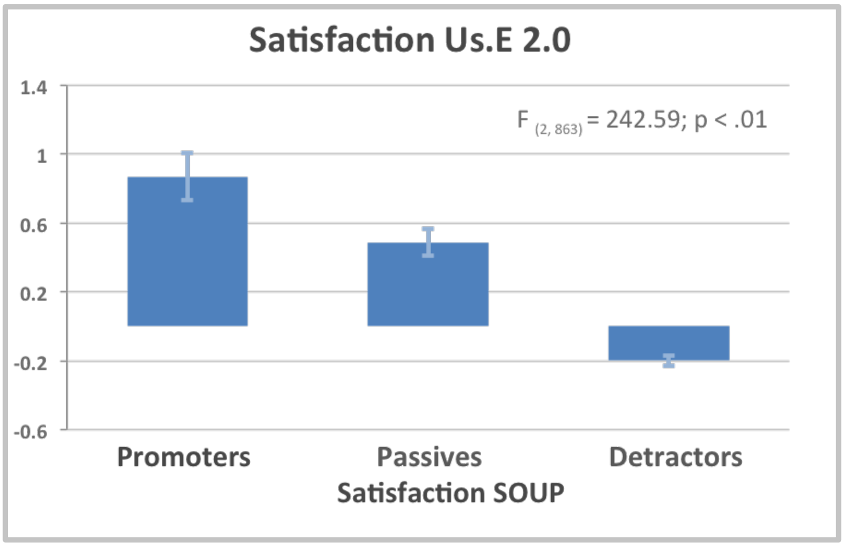


Figure 4.1.9. Us.E. 2.0 scores per SOUP user categories - Satisfaction.

The error bar denotes a confidence interval of 0.95.

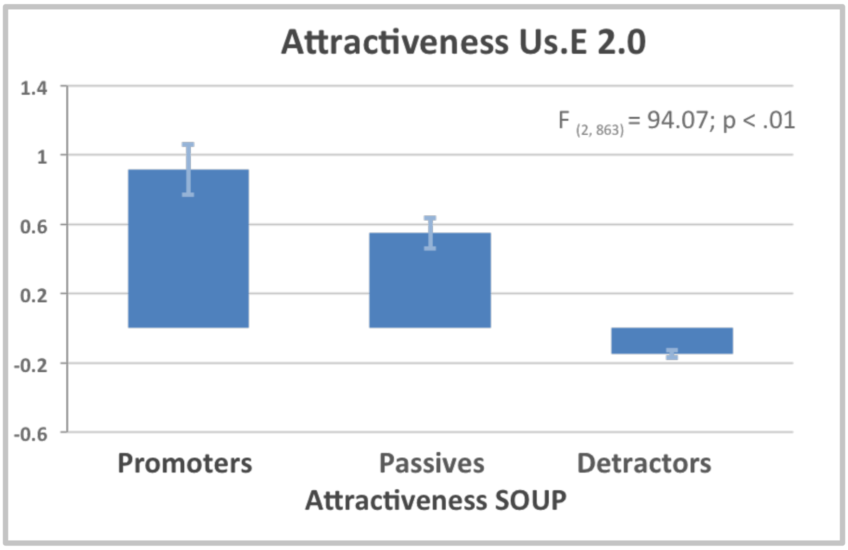


Figure 4.1.10. Us.E. 2.0 scores per SOUP user categories - Attractiveness.
The error bar denotes a confidence interval of 0.95.

Differences between academic positions

To verify the existence of any differences in usability ratings due to the participants' academic position, the scores obtained by the participants in the SUS (Brooke, 1986), the NPS (Reichheld, 2003), the Us.E. 2.0 (Di Nocera, 2013) and the SOUP questionnaires were used as dependent variables in different ANOVA designs in which the categorical variable "Academic position" was used as the factor.

Academic positions in relation to the SUS questionnaire

The result of the ANOVA between SUS scores (as dependent variable) and "Academic position" (as the factor) revealed no significant differences ($F_{3,862}$ = 1.20, $p > .05$).

Academic positions in relation to the NPS questionnaire

The result of the ANOVA between the NPS scores (as dependent variable) and "Academic position" (as the factor) revealed no significant differences ($F_{3,862}$ = 2.34, $p > .05$). The analysis showed a lower score from the "Temporary assistant professor" position, as can be seen from figure 4.1.11.

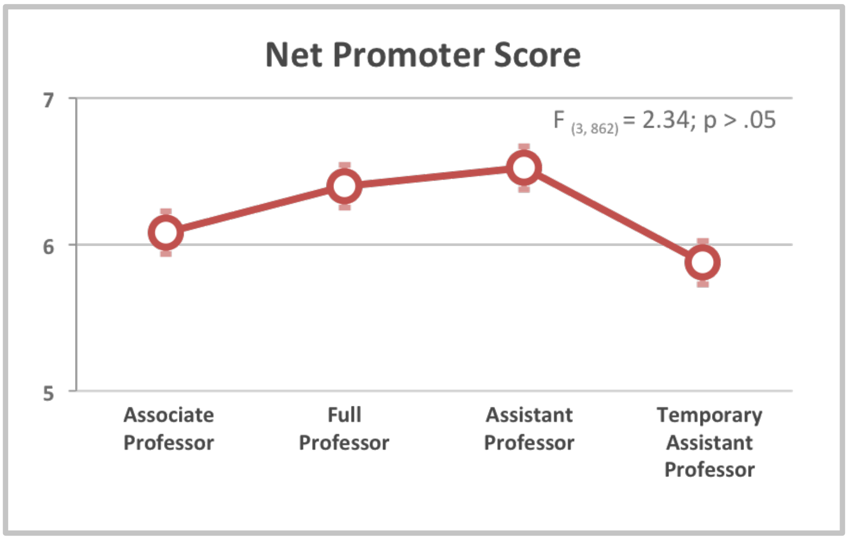


Figure 4.1.11. NPS scores per academic position.

The error bar denotes a confidence interval of 0.95.

Academic positions in relation to the Us.E. 2.0 questionnaire

Three independent ANOVAs were conducted using the Academic Position as the independent variable and the three dimensions of the Us.E. 2.0 questionnaire (Di Nocera, 2013) as the dependent variables. The analysis conducted on the Handling dimension revealed a significant effect ($F_{3,862}$ = 3.80, p<.01, Figure 4.1.12.). Duncan post hoc test showed significant differences between academic positions, specifically between "Temporary Associate professor", "Associate professor", "Full Professor" and "Associate Professor". The analyses conducted on the Satisfaction and Attractiveness dimensions respectively were not statistically significant.

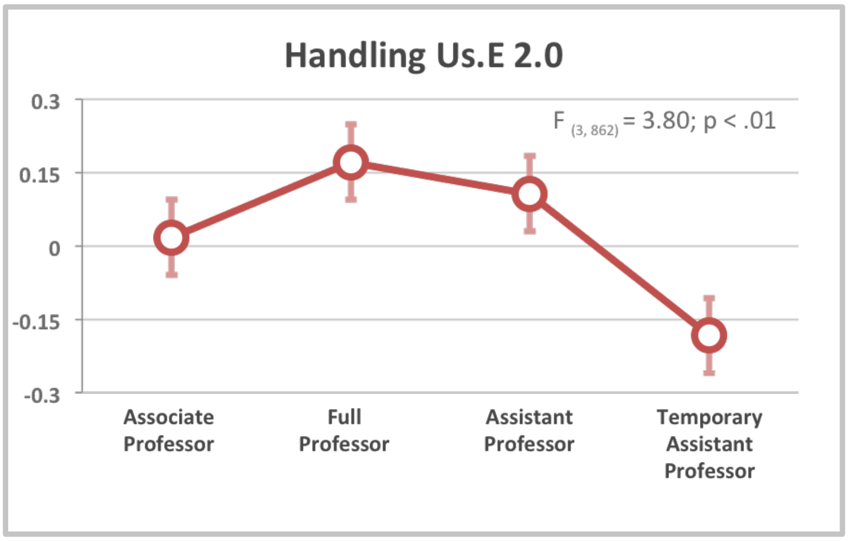


Figure 4.1.12. Comparisons of scores in Us.E. 2.0 for the Handling dimension according to the academic position. The error bar denotes a confidence interval of 0.95.

*Table 4.1.3. Pairwise comparison among Us.E. 2.0 Handling dimension and academic position (*p < 0.5).*

Academic position	Assistant professor	Temporary Assistant professor	Full professor	Associate professor
Assistant professor		.01*	.49	.34
Temporary Assistant professor			.01*	.03*
Full professor				.12
Associate professor				

Academic positions in relation to the SOUP questionnaire

Three independent ANOVAs were conducted using the Academic Position as the independent variable and the three dimensions of the SOUP questionnaire as the dependent variables. No significant differences emerged on the dimensions of “Handling” ($F_{3,862}$ = 2.49, *p* = .06) and “Satisfaction” ($F_{3,862}$ = 1.86, *p* =.13), while the differences for the “Attractive” dimension was statistically significant ($F_{3,862}$ = 2.92, *p* < .05, figure 4.1.13.).

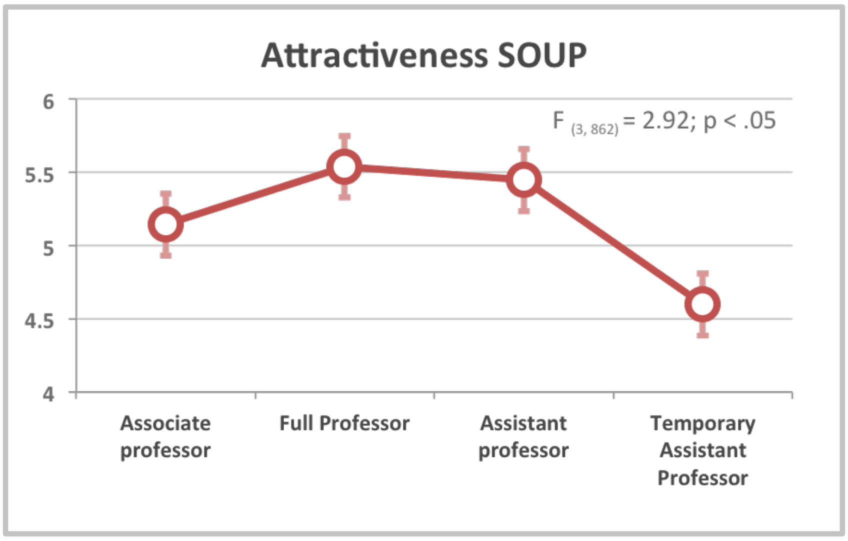


Figure 4.1.13. Comparisons of scores in the SOUP for the Attractive dimension by academic position. The error bar denotes a confidence interval of 0.95.

Duncan post hoc test confirmed a significant difference between the "Temporary Assistant professor" and "Full Professor" positions. Indeed, lower scores in terms of attractiveness in the SOUP questionnaire would appear to be associated with subjects categorised as "Temporary Assistant professor" (Table 4.1.4.).

*Table 4.1.4. Pairwise comparison among Attractive dimension and academic position (*p< 0.5).*

Academic position	Assistant professor	Temporary Assistant professor	Full professor	Associate professor
Assistant professor		.01*	.76	.31
Temporary Assistant professor			.01*	.01*
Full professor				.21
Associate professor				

Discussion Study 1

The objective of this study was to validate the SOUP questionnaire, a new efficient and reliable tool for web usability evaluation. A large number of participants (N = 866) were involved in this study. In order to evaluate the validity of the SOUP, a correlation analysis (*Pearson's r*) was carried out. SOUP's scores were compared with scores obtained from other questionnaires chosen among the most popular in the scientific literature: the SUS (Brooke, 1986), the NPS (Reichheld, 2003), and the Us.E. 2.0 (Di Nocera, 2013). Participants evaluated a management web platform of "Sapienza University of Rome". All participants were familiar with the platform as they used it more or less frequently to carry out different activities (classes and examinations management).

The statistical test showed positive and significant correlations between the SOUP scores and the other validated questionnaires, supporting a good convergent validity by the SOUP. In order to verify the validity of the SOUP, variance analyses were also carried out. Following the NPS approach, based on the SOUP scores the participants were classified as Promoters, Passives and Detractors. The results showed an association between the subjects classified as promoters and high scores in the SUS questionnaires (Brooke, 1986), the NPS (Reichheld, 2003) and the Us.E. 2.0 (Di Nocera, 2013). In particular, subjects classified as "Promoters" in the SOUP tend to provide more positive usability ratings in the other questionnaires while subjects classified as "Detractors" in the SOUP tend to provide more negative usability ratings in the other questionnaires.

Other analyses were carried out to verify any differences in usability assessments due to academic position.

As far as the academic position is concerned, in almost all the questionnaires, there is a difference in evaluations expressed by "Temporary Assistant professor", who gave significantly lower scores than the other participants. A possible explanation for this result can be found in the nature of this unusual academic position which involves less interaction with the examined web platform. As a result, it is possible that these subjects have less experience in using the experimental platform, and this could explain a difference in terms of usability assessment. Despite this, the concordance between the questionnaires supports the convergent validity of the SOUP.

In conclusion, the results have shown that the SOUP could be a valid tool for the evaluation of web usability. The advantages of using SOUP could be many. First of all, it provides a usability measure based on three dimensions. Unlike other tools that return global evaluations, the use of SOUP can give more detailed information. Secondly, SOUP can be used for evaluating interfaces in different areas. Its simplicity of administration and coding makes it an adaptable tool for different websites or interfaces in general. Furthermore, the short amount of time and effort required by users to complete it is a definite advantage for conducting large-scale research.

In the final analysis, some limitations may have influenced the results. A first limitation concerns the lack of real experimental tasks. Each participant provided a personal evaluation based on their daily use of the InfoStud platform. Designing an experimental setting by manipulating variables related to success or aesthetics could provide important information on the ability of the scale to discriminate between different dimensions of usability. Furthermore, it is not yet possible to generalise these results to other areas of web usability or other sectors such as Public Administrations, Portals and Communities, Companies and Services. Considering that specific benchmarks are missing, is still impossible to compare the SOUP usability evaluations with specific reference data.

Therefore, future studies should include new measurements that take into account the results and limitations of this preliminary study.

4.2 Study 2

The objective of this study was to include in the SOUP questionnaire an item capable of estimating the user mental workload. For this purpose, a between-subjects study has been designed. The usability and mental workload evaluations expressed by two groups of participants were compared. The groups of participants carried out some research tasks within websites characterised by a different design layout.

Participants

Fifty-one volunteers participated in this study (35 females, average age = 36; st.dev. = 9). 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 had a normal or corrected-to-normal vision. All the subjects declared to use the Internet every day. This study was performed in accordance with the Helsinki Declaration of the World Medical Association.

Materials and method

Two large italian Public Administration's websites (whose identity we are not allowed to disclose) were selected after a heuristic evaluation of their Information Architecture (IA) structures aimed at exploring the number of menu levels and their related categories. The websites were characterized by two different versions: a "well-designed" version and a "poorly designed" version based on the application of usability guidelines. The well-designed versions were similar in the design and interaction features (i.e. menu, colours, aesthetic) but different in terms of information architecture complexity. The poorly designed versions were similar both in design and interaction features (menus, colors, aesthetics) and in information architecture complexity. The Table 4.2.1. Resumes data related to the IA of the selected websites. The abbreviation "W-D" will indicate the well-designed versions, while the abbreviation "P-D" will indicate the poorly designed versions.

Table 4.2.1. 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 - W-D	6	31	171	79	0	0	287	**Low**
2 - W-D	7	76	256	195	9	0	543	**High**
1 - P-D	7	48	30	31	55	72	243	**Low**
2 - P-D	10	86	171	10	0	0	277	**Low**

Three equivalent research tasks have been proposed for each website. Subjects had to search specific information in different areas of the websites (e.g. "Administration", "Open Data", "Downloads").

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*: task completion time is usually used to measure the efficiency of a system (Albert & Tullis, 2013). In this study we considered task completion time as the amount of time the user needs to complete all the assigned tasks.

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 & Covey, 2006);
- Usability Evaluation 2.0: (Us.E. 2.0: Di Nocera, 2013);
- System Usability Scale (SUS - Brooke, 1986);
- Simple Outlook on Usability and Promotion (SOUP - Di Nocera, in press);
- NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988);
- NASA-TLX (Adapted Version): with the aim of adding one or two items to the SOUP questionnaire to evaluate MWL, we have adapted the NASA-TLX questionnaire following the NPS (Reichheld, 2003) structure. For each MWL size investigated by NASA-TLX we designed an item that leveraged the “word of mouth” paradigm, asking participants how likely they would recommend the use of a specific website to friends or colleagues. The resulting questionnaire is illustrated below:

Table 4.2.2. The adapted version of the NASA-TLX.

MWL dimension	Item
Mental Demand	With specific reference to the mental activity required to interact with this website (for example, understand its architecture and organization of information, read the texts, find specific information), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Physical Demand	With specific reference to the physical effort required to interact with this website (for example, number of actions required, number of clicks on links, need to continuously move your eyes on the interface, etc.), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Temporal Demand	With specific reference to the time needed to achieve the objectives of browsing this website (too fast or too slow), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Overall Performance	With specific reference to achieving your navigation goals on this website (for example, finding the information you were looking for), how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Effort	With specific reference to the level of physical and mental effort required to interact with this website, how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely
Frustration Level	With specific reference to how irritated, stressed and annoyed you felt rather than relaxed and smug while browsing this website, how likely would you recommend it to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely

Moreover, in order to include in the SOUP questionnaire an item capable of estimating the mental workload, the following item was add at the SOUP scale: "*With specific reference to how hard you had to try (e.g., maintaining focus, thinking, making decisions, etc.) to find the information you were looking for, how likely would you recommend this site to a friend or colleague?*".

Table 4.2.3. The SOUP Mental Workload item.

SOUP dimension	Item
SOUP_MWL	With specific reference to how hard you had to try (e.g., maintaining focus, thinking, making decisions, etc.) to find the information you were looking for, how likely would you recommend this site to a friend or colleague? 0 1 2 3 4 5 6 7 8 9 10 Not at all likely Extremely likely

Procedure

Participants were divided into two groups: group "A" performed the experiment in "Well-designed" websites while group "B" performed the experiment in "Poorly-designed" websites. They performed the entire test remotely in a single session of about 45 minutes. With the aim of avoiding effects related to the order of presentation of the stimuli, the websites and the tasks were randomly assigned to the participants. The researchers moderate the experimental sessions through a platform for screen, audio and video sharing. All participants used their personal computer (notebook or desktop computer). Specifically, each session included:

1. Familiarisation with the first website under investigation: a free navigation session in which the participants observed the website structure (duration: approximately 2 minutes);
2. Performing the assigned tasks for the first website (duration: approximately 15 minutes);
3. Answering to questionnaires for the first website: after completing all the three tasks, participants answered the questionnaires concerning perceived usability and mental workload scales (duration: approximately 5 minutes);
4. Familiarisation with the second website under investigation: a free navigation session in which the participants observed the website structure (duration: approximately 2 minutes);
5. Performing the assigned tasks for the second website (duration: approximately 15 minutes);
6. Answering to questionnaires for the second website: after completing all the three tasks, participants answered the questionnaires concerning perceived usability and mental workload scales (duration: approximately 5 minutes).

Before the beginning of each session, the participants were invited to answer a questionnaire of a personal nature aimed at collecting information such as gender, age, education, occupation and frequency and mode of Internet use. In addition, the degree of familiarity of the participants with the site under investigation was investigated. Twenty-five participants completed the experimental session for the “Well-Designed” condition, while twenty-six participants completed the experimental session for the “Poorly Designed” condition.

Data Analyses and Results

Pearson's r coefficient

Pearson's r coefficient analysis showed significant correlations between SOUP scores with the SUS (Brooke, 1986) and the NPS (Reichheld, 2003) for both websites. The SOUP Handling scale achieved a statistically significant Pearson's r correlation with both the SUS (Brooke, 1986) and the NPS (Reichheld, 2003) on both websites. The correlation with the SUS (Brooke, 1986) was .85 for website 1 and .78 for website 2, while the correlation index with the NPS (Reichheld, 2003) was .81 for website 1 and .73 for website 2. The SOUP Satisfaction scale achieved a statistically significant correlation with both the SUS (Brooke, 1986) and the NPS (Reichheld, 2003) on both websites. With the SUS (Brooke, 1986), the correlation was .69 for website 1 and .69 for website 2; with the NPS (Reichheld, 2003), the correlation index was .75 for website 1 and .83 for website 2. The SOUP Attractiveness scale achieved a statistically significant correlation with both the SUS (Brooke, 1986) and the NPS (Reichheld, 2003) on both websites. With the SUS (Brooke, 1986), the correlation was .58 for website 1 and .54 for website 2; with the NPS (Reichheld, 2003), the correlation index was .56 for website 1 and .65 for website 2. The results are summarised in Table 4.2.4.

*Table 4.2.4. Correlation matrix between SOUP dimensions and the other usability questionnaires (*p < .05).*

Correlation between SOUP dimensions and the other usability questionnaires		
Website 1		
SOUP	SUS	NPS
SOUP_H	.85 *	.81 *
SOUP_S	.69 *	.75 *
SOUP_A	.58 *	.56 *
Website 2		
SOUP	SUS	NPS
SOUP_H	.78 *	.73 *
SOUP_S	.69 *	.83 *
SOUP_A	.54 *	.65 *

The analysis of Pearson's *r* coefficient showed significant correlations between the NASA-TLX scores, the SOUP Mental Workload item and the NASA-TLX adapted version. The results confirmed a negative correlation between the NASA-TLX scores, the SOUP Mental Workload item and the NASA-TLX adapted version, as summarised in Table 4.2.5.

*Table 4.2.5. Correlation matrix between NASA-TLX and the other mental workload items (*p < .05).*

Correlation between NASA-TLX and the other mental workload items							
Website 1							
	SOUP MWL	TLX-A Mental Demand	TLX-A Physical Demand	TLX-A Temporal Demand	TLX-A Performance	TLX-A Effort	TLX-A Frustration
NASA-TLX	- .69*	- .55 *	- .52 *	- .71 *	- .77 *	- .70 *	- .70 *
Website 2							
	SOUP MWL	TLX-A Mental Demand	TLX-A Physical Demand	TLX-A Temporal Demand	TLX-A Performance	TLX-A Effort	TLX-A Frustration
NASA-TLX	- .47*	- .41 *	- .54 *	- .39 *	- .55 *	- .46 *	- .64 *

The analysis of Pearson's r coefficient showed significant correlations between the SOUP Mental Workload item and the NASA-TLX adapted version. The results confirmed positive correlations between the two scales, as summarised in Table 4.2.6.

*Table 4.2.6. Correlation between the SOUP Mental Workload item and the NASA-TLX adapted version (*p < .05).*

Correlation between the SOUP Mental Workload item and the NASA-TLX adapted version						
Website 1						
	TLX-A Mental Demand	TLX-A Physical Demand	TLX-A Temporal Demand	TLX-A Performance	TLX-A Effort	TLX-A Frustration
SOUP MWL	.72 *	.67 *	.84 *	.75 *	.84 *	.83 *
Website 2						
	TLX-A Mental Demand	TLX-A Physical Demand	TLX-A Temporal Demand	TLX-A Performance	TLX-A Effort	TLX-A Frustration
SOUP MWL	.75 *	.62 *	.71 *	.63 *	.73 *	.70 *

Variance Analysis (ANOVA) – Usability questionnaires

The scores obtained by the participants in the SUS (Brooke, 1986), the NPS (Reichheld, 2003) and SOUP questionnaires were used as dependent variables in different ANOVA designs using the design condition (i.e. "Well-Designed" and "Poorly Designed") as a factor.

The results of the ANOVAs conducted using the SUS (Brooke, 1986) showed a significant effect of the "Design condition" on the Attractiveness scale scores for the website 1 ($F_{(1,49)} = 7.59$; $p < .05$; see figure 4.2.1.). Specifically, the "Well-Designed" websites are associated with higher scores in the SUS questionnaire, while the "Poorly-Designed" websites are associated with lower scores.

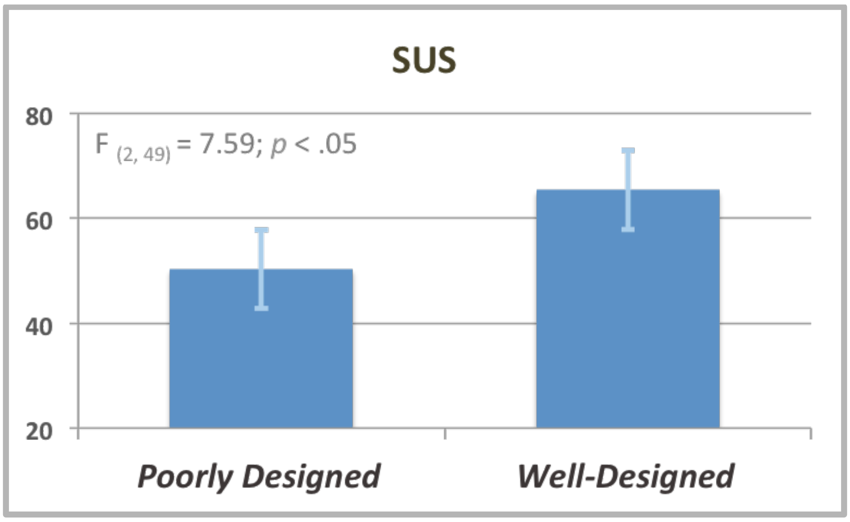


Figure 4.2.3. SUS scores per design condition (Website 1). The error bar denotes a confidence interval of 0.95.

The results of the ANOVAs conducted using the "Handling" dimension of the SOUP questionnaire showed a significant effect of the "Design condition" on the Attractiveness scale scores for the website 1 ($F_{(1,49)} = 5.03$; $p < .05$; see figure 4.2.2.). Specifically, the "Well-Designed" websites are associated with higher scores in the "Handling" dimension of the SOUP questionnaire, while the "Poorly-Designed" websites are associated with lower scores in the "Handling" dimension.

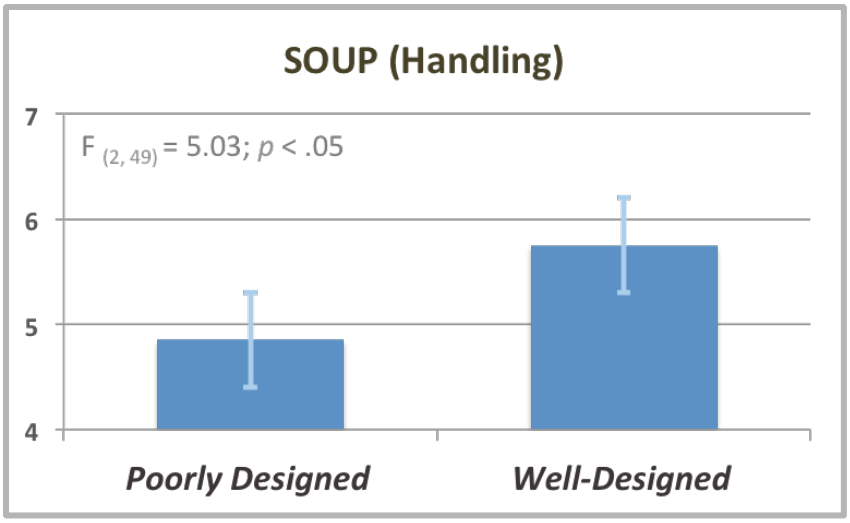


Figure 4.2.3. SOUP Handling dimension scores per design condition (Website 1). The error bar denotes a confidence interval of 0.95.

The results of the ANOVAs conducted using the "Attractiveness" dimension of the SOUP questionnaire showed a significant effect of the "Design condition" on the Attractiveness scale scores for the website 1 ($F_{(1,49)} = 27.02$; $p < .01$; see figure

4.2.3.). Specifically, the “Well-Designed” websites are associated with higher scores in the “Attractiveness” dimension of the SOUP questionnaire, while the “Poorly-Designed” websites are associated with lower scores in the “Attractiveness” dimension.

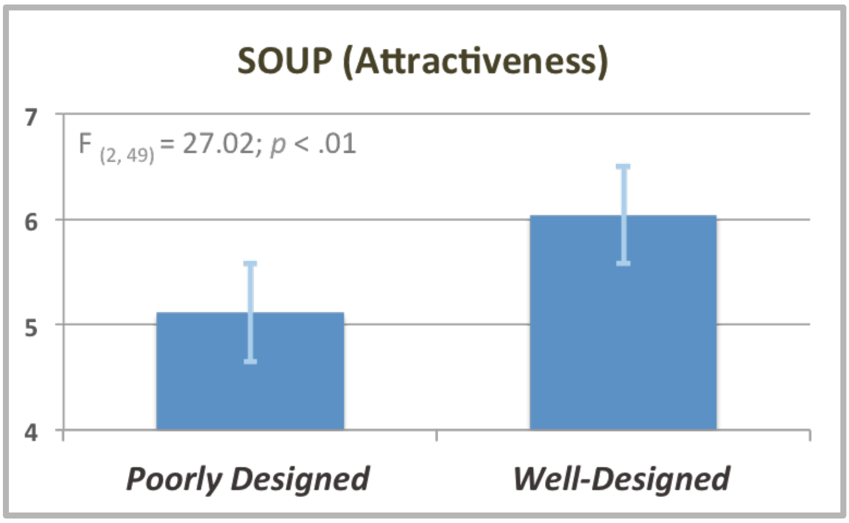


Figure 4.2.3. SOUP Attractiveness dimension scores per design condition (Website 1). The error bar denotes a confidence interval of 0.95.

The results of the ANOVAs conducted using the NPS and the “Satisfaction” dimensions of the SOUP questionnaire didn’t show significant effects.

Variance Analysis (ANOVA) – Mental workload questionnaires

The scores obtained by the participants in the NASA-TLX (Hart & Staveland, 1988), in the NASA-TLX- Adapted and in the SOUP MWL dimension were used as dependent variables in different ANOVA designs using the design condition (i.e. “Well-Designed” and “Poorly Designed”) as a factor. The results of the ANOVAs didn’t show significant effects.

Variance Analysis (ANOVA) – Performance measures

The success rate and the time-on-task obtained by the participants in the assigned tasks were used as dependent variables in different ANOVA designs using the design condition “(i.e. “Well-Designed” and “Poorly Designed”) as a factor.
The results of the ANOVAs didn’t show significant effects.

Discussion Study 2

The objective of this study was to include in the SOUP questionnaire an item capable of estimating the user mental workload. Moreover, this study aimed to evaluate the SOUP's capability to discriminate between different dimensions of usability such as handling, satisfaction and attractiveness. The goal of including a subjective measure for estimating mental workload is inspired by the hypothesis that the two constructs should be integrated in the study and evaluation of digital interfaces. Historically, mental workload is a construct studied only in highly complex operating systems. Consequently, subjective measures of mental workload are marked by a very specific lexicon. In addition, the administration is geared toward a reference population of experts. Constructing a measure capable of estimating mental workload with reference to "common" users and interfaces requires lexical and structural adaptation. For this purpose the subjective evaluations of usability and mental workload expressed by two groups of subjects were compared. Fifty-one participants carried out some research tasks within websites characterized by a different design layout (i.e. "Well-Designed" VS "Poorly Designed").

In order to evaluate the validity of the SOUP, a correlation analysis (*Pearson's r*) was carried out with scores obtained from other questionnaires chosen among the most popular in the scientific literature: the SUS (Brooke, 1986) and the NPS (Reichheld, 2003). The statistical test showed positive and significant correlations between the SOUP scores and the other questionnaires, supporting a good convergent validity by the SOUP. In order to verify the SOUP's capability to discriminate between different dimensions of usability, variance analyses were also carried out. The scores of the SUS, the NPS and the different SOUP dimensions (Handling, Satisfaction, Attractiveness) were compared for the "Well-Designed" and the "Poorly-Designed" websites. The results showed no differences between the "Well-Designed" and the "Poorly Designed" websites for the SUS, the NPS and the "Handling" and "Satisfaction" SOUP dimensions. A significant difference emerged between the "Well-Designed" and the "Poorly-Designed" condition for the "Attractiveness" SOUP dimension for the website 1. This result underlines the capability of the SOUP to discriminate, net of other factors, for the usability issues related to the aesthetic attributes of a website. Moreover, this result takes on more significance if we consider that no differences emerged between the “Well-Designed” and the “Poorly Designed” condition with regard to the mental workload experienced by the participants and their performance at the assigned tasks. This result suggests that some of the most widely used usability questionnaires are "unbalanced". In fact, they correctly detect aspects of usability related to manageability and satisfaction, while leaving out the aesthetic dimension. The added value of the SOUP could be to provide a multidimensional assessment, rather than a global score, as is the case for the SUS (Brooke, 1986) and the NPS (Reichheld, 2003).

In conclusion, this preliminary study has shown that the SOUP could be a valid tool for the evaluation of web usability dimensions. Moreover, the correlation analysis (*Pearson's r*) carried out between the NASA-TLX (Hart & Staveland, 1988) scores, the SOUP Mental Workload item and the NASA-TLX adapted version, have shown that there are good indications for including an evaluation of the MWL within the SOUP. Specifically, the SOUP item designed to assess mental workload has a negative and significant correlation with the criterion measure (NASA-TLX, Hart & Staveland, 1988). The correlation is negative because it asks an inverse question with respect to the criterion measure and, therefore, indicates a consistent response between the two measures. This result is absolutely encouraging, as it would suggest the possibility of synthesizing the six different scales represented by the NASA-TLX into a single item (Hart & Staveland, 1988).

In the final analysis, some limitations may have influenced the results. The size of the experimental sample is the major limitation of this study. The results obtained, although promising, cannot be generalised. Considering that the experimental design is of the “between-subjects” type, it would be necessary to include many more participants to obtain reliable results. Future studies should include new measurements that take into account the results and limitations of this preliminary study.

4.3 Study 3[2]

The main objective of the study was: i) to evaluate the mental workload associated with browsing websites with different levels of complexity and ii) to understand its effects on perceived usability. Our hypothesis was that a greater complexity of the information architecture structure would be related to higher mental workload and poor usability evaluations. Three large Italian Public Administration's websites (whose identity we are not allowed to disclose) were selected after a heuristic evaluation of their IA structures aimed at exploring the number of menu levels and their related categories. All websites were similar in the design and interaction features (i.e. menu, colours, aesthetic) but different in terms of information architecture complexity. From the less to the more complex, the identified websites will be indicated as website 1, website 2, website 3 hereinafter.

Table 4.3.1. 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

Participants

Twenty volunteers participated in this study (7 females, average age = 57; st.dev. = 6). 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 had a normal or corrected-to-normal vision. All the subjects declared to use the Internet every day. This study was performed in accordance with the Helsinki Declaration of the World Medical Association.

[2] This study refers to the following publication: Serra, G., De Falco, F., Maggi, P., Forsi, R., Cocco, A., Gaudino, G., ... & Di Nocera, F. The role of mental workload in determining the relation between website complexity and usability: an eye-tracking study.

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 it was installed below a 22'' desktop computer screen

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: task completion time is usually used to measure the efficiency of a system (Albert & Tullis, 2013). In this study we considered task completion time as the amount of time the user needs to complete all the assigned tasks.

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 & Covey, 2006);
- Usability Evaluation 2.0: (Us.E. 2.0: Di Nocera, 2013);
- Simple Outlook on Usability and Promotion (SOUP: Di Nocera, in press);
- NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988).

Procedure

Five equivalent research tasks have been proposed for each website. The tasks included similar activities for each website, e.g. downloading a form, obtaining information about a service, consulting a table of data. The tasks were designed 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.Subjects had to search specific information in different areas of the websites (e.g. “Administration”, “Open Data”, “Downloads”). Prior knowledge of websites was investigated by asking participants if they had ever browsed the selected websites. All the participants stated 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 from each other, in order to limit effects related to fatigue and task duration. Moreover, with the aim of avoiding 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 observed the website structure (duration: approx. 5 minutes);
2. Eye-tracker calibration: participants were positioned at a distance of about 60 cm from a 22" screen, they performed a dynamic 9-point calibration, the calibration always started from the centre of the screen (duration: about 3 minutes);
3. Performing the assigned tasks: participants in each session performed five research tasks on one of the target sites. The starting fixation point for each task was the centre of the screen. 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: approximately 30 minutes);
4. Answering to questionnaires: after completing all the five tasks, participants answered several questionnaires concerning their personal information (e.g. gender, age, educational qualification, employment, the frequency of internet use), perceived usability and mental workload scales (duration: approximately 10 minutes).

Data analysis and results

Success rate, completion time, perceived complexity, NPS, Us.E. 2.0 (Handling, Satisfaction, Attractiveness), NASA-TLX, and NNI scores were analysed in repeated-measure ANOVA designs using Complexity (website 1 vs. website 2 vs.

website 3) as the repeated factor.

Success rate was significantly different between websites ($F_{2,36} = 3.04$; $p < .05$). Duncan post-hoc testing showed that the success rate for the high-complexity website (website 3) was significantly lower than the other two (see Figure 4.3.1.).

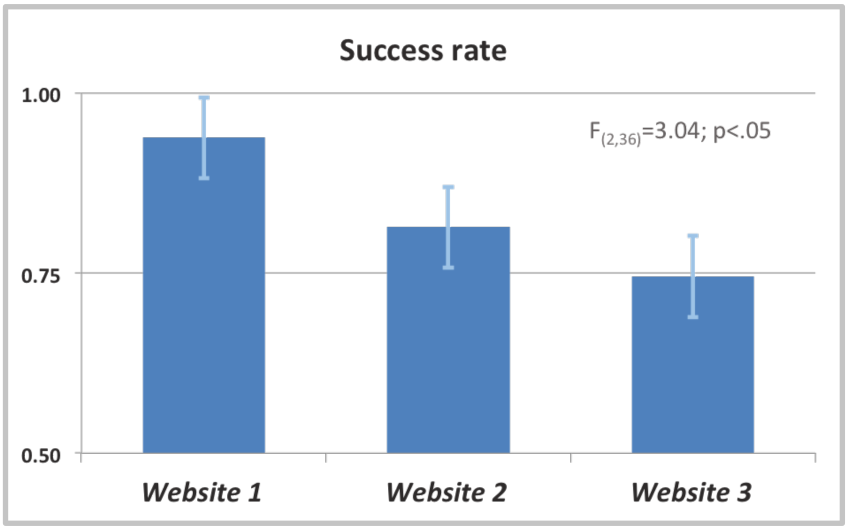


Figure 4.3.1. Success rate per website

*Table 4.3.2. Pairwise comparison among Success Rate and website (*p < 0.5).*

Success Rate	Website 1	Website 2	Website 3
Website 1			
Website 2	.12		
Website 3	.02*	.39	

Completion time (seconds) was significantly different between websites ($F_{2,36} = 4.02$; $p < .05$). Duncan post-hoc testing showed that completion time for the low-complexity website (website 1) was significantly faster than the other two (see Figure 4.3.2.).

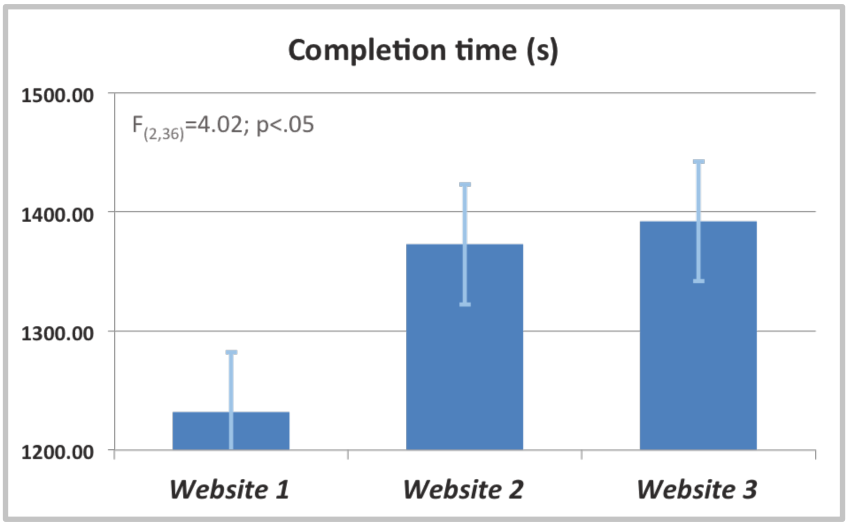


Figure 4.3.2. Completion time (average) per website.

*Table 4.3.3. Pairwise comparison among Completion Time and website (*p < 0.5).*

Completion time	Website 1	Website 2	Website 3
Website 1			
Website 2	.02*		
Website 3	.01*	.75	

Perceived complexity was significantly different between websites ($F_{2,36} = 3.92$; $p < .05$). Duncan post-hoc testing showed that the perceived complexity of the low-complexity website (website 1) was significantly lower than the other two (see Figure 4.3.3.).

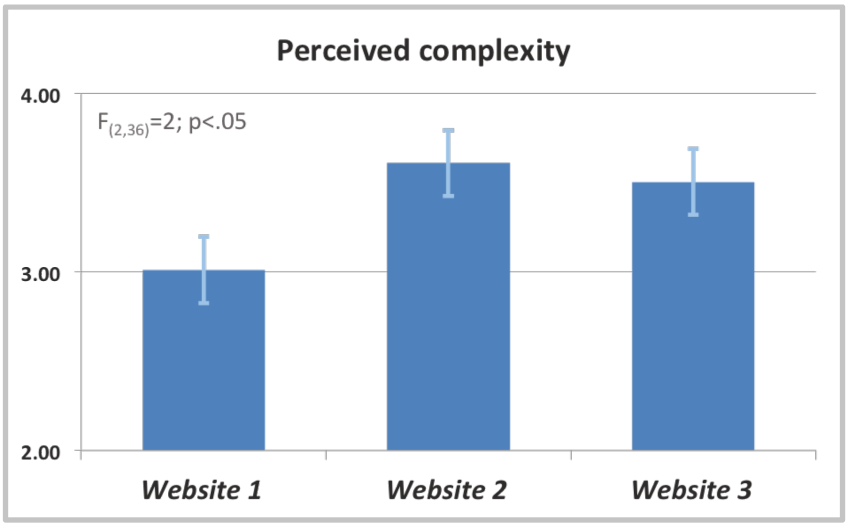


Figure 4.3.3. Perceived complexity per website.

*Table 4.3.4. Pairwise comparison among Perceived complexity and website (*p < 0.5).*

Perceived complexity	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.03*	.64	

NPS score was significantly different between websites ($F_{2,36} = 4.52$; $p < .05$). Duncan post-hoc testing showed that the proportion of the low-complexity website (website 1) was significantly higher than the other two (see Figure 4.3.4.).

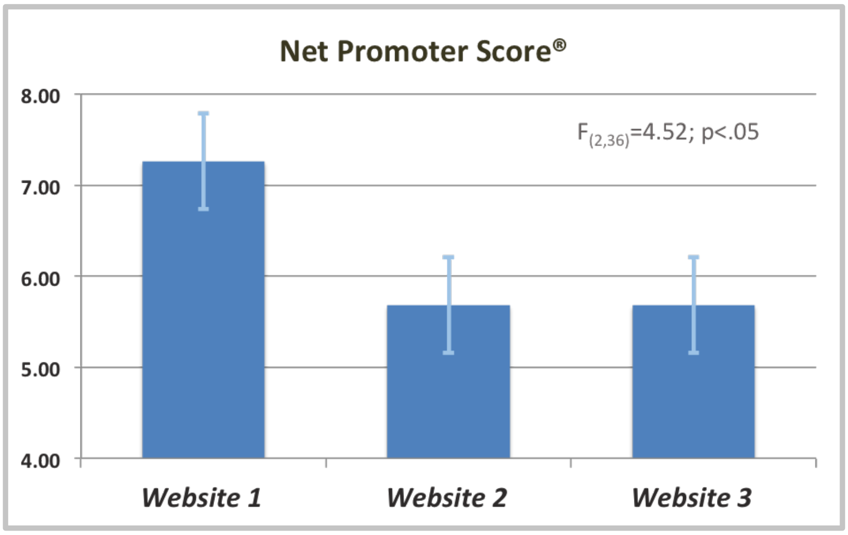


Figure 4.3.4. Net promoter Score per website

*Table 4.3.5. Pairwise comparison among Net Promoter Score and website (*p < 0.5).*

NPS	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.01*	.78	

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" ($F_{2,36} = 6.80$, $p < .01$) and "Satisfaction" ($F_{2,36} = 3.45$, $p < .05$). No significant differences were found for the dimension "Attractiveness" (see Figure 4.3.5.).

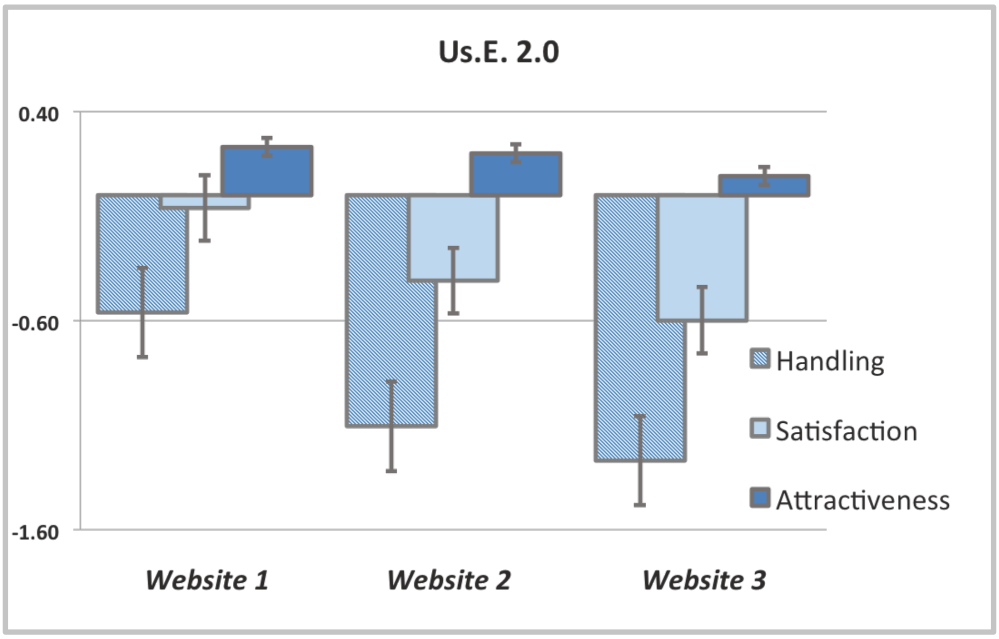


Figure 4.3.5. Us.E. 2.0 scores per website.

*Table 4.3.6. Pairwise comparison among Us.E. 2.0 (Handling dimension) and website (*p < 0.5).*

Handling (Us.E.)	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.01*	.42	

*Table 4.3.7. Pairwise comparison among Us.E. 2.0 (Satisfaction dimension) and website (*p < 0.5).*

Satisfaction (Us.E.)	Website 1	Website 2	Website 3
Website 1			
Website 2	.10		
Website 3	.01*	.36	

SOUP 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" ($F_{2,36}$ = 5.70, *p* < .01) and "Satisfaction" ($F_{2,36}$ = 3.45, *p* < .05). No significant differences were found for the dimension "Attractiveness" (see Figure 4.3.6.).

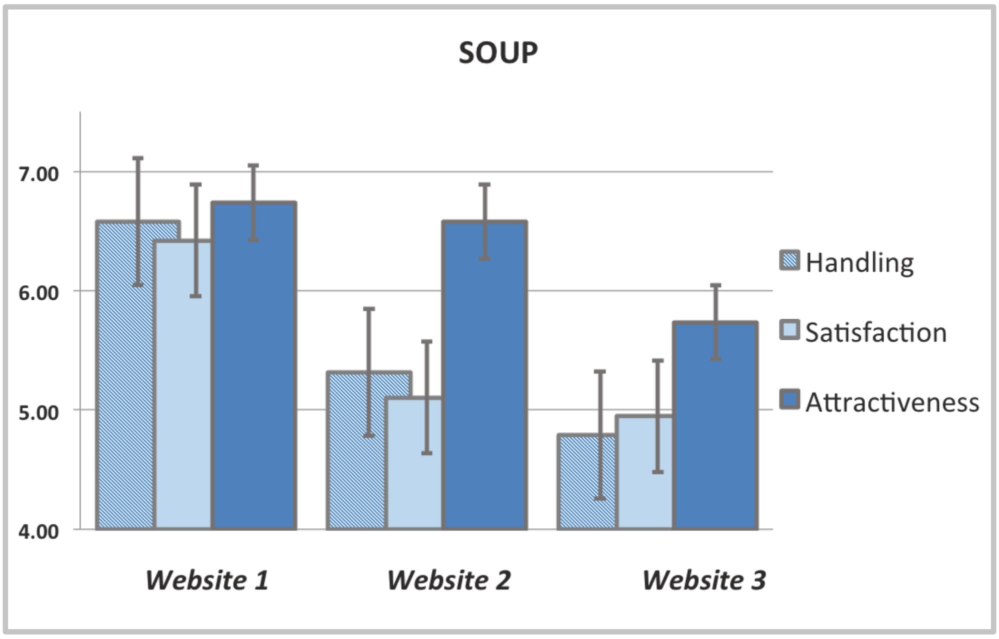


Figure 4.3.6. SOUP scores per website.

*Table 4.3.8. Pairwise comparison among SOUP (Handling dimension) and website (*p < 0.5).*

Handling (SOUP)	Website 1	Website 2	Website 3
Website 1			
Website 2	.02*		
Website 3	.01*	.34	

*Table 4.3.9. Pairwise comparison among SOUP (Satisfaction dimension) and website (*p < 0.5).*

Satisfaction (SOUP)	Website 1	Website 2	Website 3
Website 1			
Website 2	.04*		
Website 3	.02*	.79	

NASA-TLX score was significantly different between websites ($F_{2,36} = 7.38$; $p < .01$). Duncan post-hoc testing showed that perceived workload for the low-complexity website (website 1) was significantly lower than the other two (see Figure 4.3.7.).

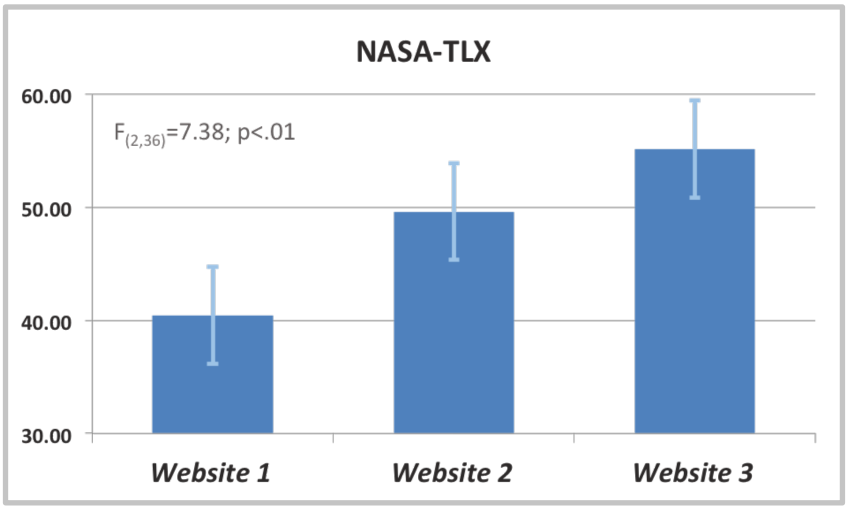


Figure 4.3.7. NASA-TLX scores per website.

*Table 4.3.10. Pairwise comparison among NASA-TLX and website (*p < 0.5).*

NASA-TLX	Website 1	Website 2	Website 3
Website 1			
Website 2	.02*		
Website 3	.01*	.18	

The Nearest Neighbour Index was significantly different between websites ($F_{2,36}$ = 6.41; p < .01). Duncan post-hoc testing showed that the fixation pattern of the medium- and high-complexity websites (websites 2 and 3) were significantly more clustered than the low-complexity website (see Figure 4.3.8.).

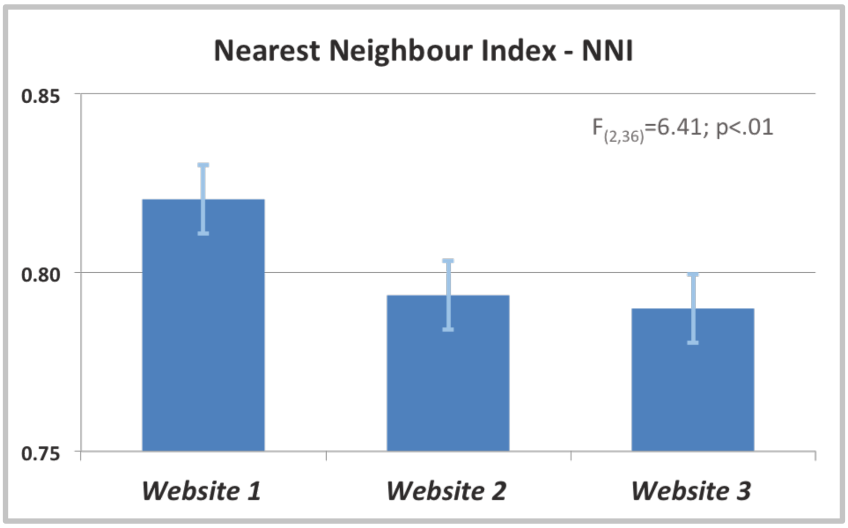


Figure 4.3.8. Nearest Neighbour Index per website.

*Table 4.3.11. Pairwise comparison among NNI and website (*p < 0.5).*

NNI	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.01*	.78	

Discussion Study 3

Results indicate a correspondence between usability measures and mental workload measures. Specifically, the websites associated with lower levels of mental workload (assessed by both objective and subjective measures) received more positive usability evaluations and were associated with a greater success rate in the assigned search tasks.

The analysis of the performance measures showed that the success rate was higher and completion time shorter for the low-complexity website than the other two. These results are also confirmed by the analysis of "perceived complexity" scores that were lower for the low-complexity website than the other two.

Regarding perceived usability, NPS scores were statistically significantly higher for website 1, while lower values were reported for the websites 2 and 3. Consistent results can be found in the usability evaluations expressed using the Us.E. 2.0 and the SOUP scales.

The analysis showed statistically higher values for the website 1 in both the "(Mental) Handling" and "Satisfaction" scales, while lower values were reported for the websites 2 and 3. The "(Mental) Handling" scale, which measures the interaction with the structure of the website (e.g. information architecture, layout) was found as the most problematic by the participants. The "Attractiveness" scale did not show any difference between websites, highlighting that the aesthetic evaluations were not influenced by the quality of the user experience.

On the side of perceived MWL, the analysis of NASA-TLX questionnaire showed lower scores for the low-complexity website than the other two, confirming, in line with perceived usability results, that participants experienced low MWL while browsing the less complex website and high MWL while browsing the more complex website. This result is supported by the analysis of ocular behaviour.

The analysis of eye movements showed statistically significantly higher NNI values for website 1, while lower values were reported for the websites 2 and 3. According to Clark and Evans (1954), NNI values close to 1 indicate that the distribution of fixations is not different from a random distribution, NNI values greater than 1 indicate a dispersion of the fixation pattern, while NNI values less than 1 indicate a grouping of fixations. With that in mind, NNI analysis highlights a less clustered fixations pattern for website 1 and, on the contrary, a more clustered fixations pattern for the websites 2 and 3.

Camilli, Terenzi and Di Nocera (2008) found that when a task imposes a high visual-spatial demand on the user - as in the case of searching for information on a web page - a greater grouping of fixations (i.e.: smaller NNI values) is associated with a greater mental workload experienced by the user. Therefore, it is correct to say that subjects involved in this study have certainly experienced a greater mental workload while browsing the higher-complexity websites.

In conclusion, the results indicated consistency between usability and mental workload measures. Specifically, the websites associated with lower levels of mental workload (assessed by both objective and subjective measures) received more positive usability evaluations and were associated with a greater success rate in the assigned search tasks.

4.4 Study 4

According to the latest reports published by Audiweb (Audiweb, 2019), the use of the Internet through mobile devices is the main mode of access to the Internet in Italy. In fact, in 2019 about 29.3 million users (average per day) have browsed through mobile devices (smartphones and tablets). The exclusive use of mobile devices to access the Internet has exceeded the number of users who access the Internet through desktop computers.

With that in mind, the aim of this study was to replicate the "Study 3" on mobile devices. The selected websites have in fact a "responsive web design". Responsive web design (RWD) is a web development approach that creates dynamic changes to the appearance of a website, depending on the screen size and orientation of the device being used to view it. RWD is one approach to the problem of designing for the multitude of devices available to users, ranging from smartphones to desktop monitors. Specifically, a responsive website is characterized by the same information architecture and by the same layout and contents in both the desktop and mobile versions, except for the input and scrolling modes of the page which are obviously based on the touch of the users on the screen.

Participants

Twenty volunteers participated in Study 4 (7 females, average age = 55.8; st.dev. = 4.6). All of them were native Italian speakers, were naïve as to the aims, the expected outcomes, and the methodology of the experiment, and had a normal or corrected-to-normal vision. All the subjects declared to use the Internet every day both through a desktop computer, laptop and smartphone. This study was performed in accordance with the Helsinki Declaration of the World Medical Association.

Material and Method

The X2-30 eye-tracking system (Tobii, Sweden) was used to record eye movements during the interaction with the websites. . In this study it was installed below an 7'' tablet screen.

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 & Covey, 2006);
- Usability Evaluation 2.0: (Us.E. 2.0: Di Nocera, 2013);
- Simple Outlook on Usability and Promotion (SOUP - Di Nocera, in press);
- System Usability Scale (SUS - Brooke, 1996).
- NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988).

Procedure

The same five research tasks used in "Study 3" have been proposed for each website. Participants performed the entire test in three separate sessions. The single sessions were performed at about 15 days apart from each other, in order to limit effects related to fatigue and task duration. Moreover, with the aim of avoiding 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 observed the website structure (duration: approx. 5 minutes);

2. Eye-tracker calibration: participants were positioned at a distance of about 60 cm from a 7" tablet, they performed a dynamic 9-point calibration, the calibration always started from the centre of the screen (duration: about 3 minutes);
3. Performing the assigned tasks: participants in each session performed five research tasks on one of the target sites. The starting fixation point for each task was the centre of the screen. 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: approximately 30 minutes);
4. Answering to questionnaires: after completing all the five tasks, participants answered several questionnaires concerning their personal information (e.g. gender, age, educational qualification, employment, the frequency of internet use), perceived usability and mental workload scales (duration: approximately 10 minutes).

Data analysis and results

Success rate, completion time, perceived complexity, NPS, Us.E. 2.0 (Handling, Satisfaction, Attractiveness), SUS, NASA-TLX, and NNI scores were analysed in repeated-measure ANOVA designs using Complexity (website 1 vs. website 2 vs. website 3) as the repeated factor.

Success rate was significantly different between websites ($F_{2,30} = 5.58$; $p < .01$). Duncan post-hoc testing showed that the success rate for website 2 was significantly lower than the website 1 (see Figure 4.4.1.).

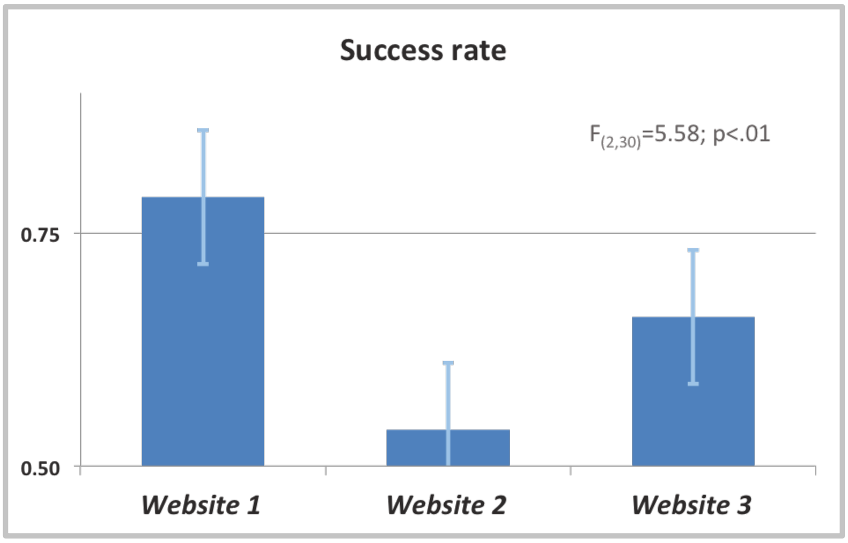


Figure 4.4.1. Success rate per website.

*Table 4.4.1. Pairwise comparison among Success Rate and website (*p < 0.5).*

Success Rate	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.09	.11	

Completion time (seconds) was significantly different between websites ($F_{2,30}$ = 9.00; p < .01). Duncan post-hoc testing showed that completion time for the low-complexity website (website 1) was significantly faster than the other two (see Figure 4.4.2.).

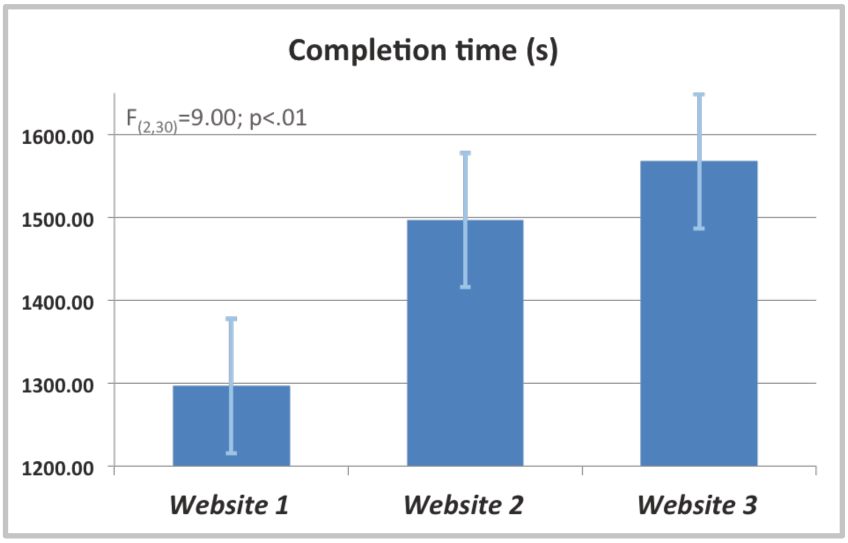


Figure 4.4.2. Completion time (average) per website.

Table 4.4.2. Pairwise comparison among Completion Time and website (p < 0.5).*

Completion time	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.29	.01*	

Perceived complexity was significantly different between websites ($F_{2,30}$ = 6.20; *p* < .01). Duncan post-hoc testing showed that the perceived complexity of website 2 was significantly higher than the other two (see Figure 4.4.3.).

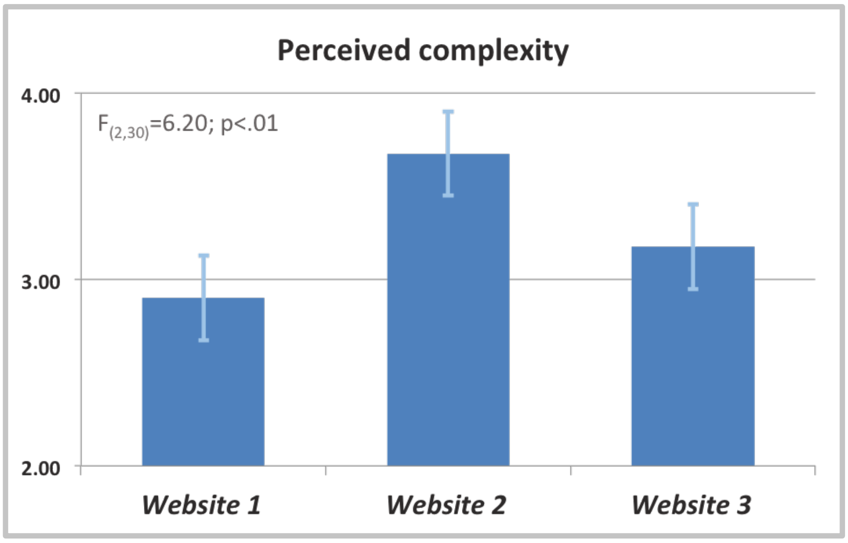


Figure 4.4.3. Perceived complexity per website.

*Table 4.4.3. Pairwise comparison among Perceived complexity and website (*p < 0.5).*

Perceived complexity	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.22	.03*	

NPS score was significantly different between websites ($F_{2,30}$ = 6.78; *p* < .01). Duncan post-hoc testing showed that scores of website 2 were significantly lower than the other two (see Figure 4.4.4.).

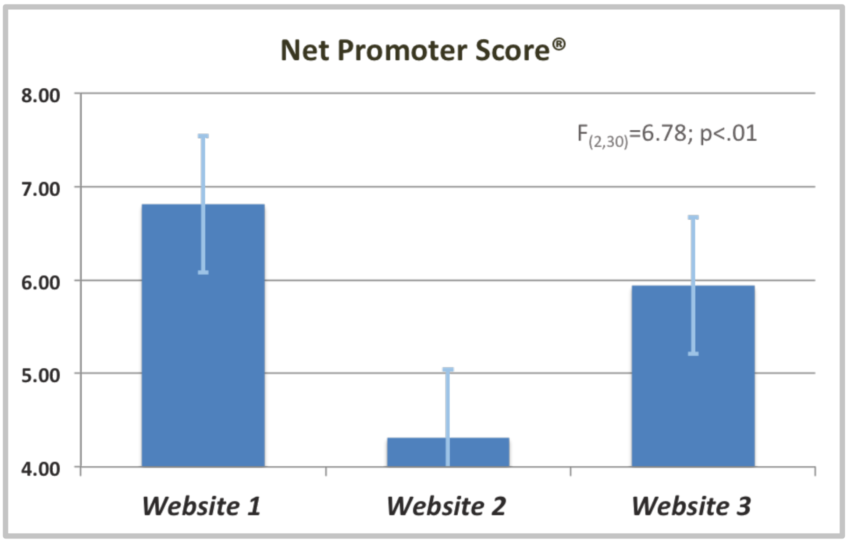


Figure 4.4.4. Net Promoter Score per website.

*Table 4.4.4. Pairwise comparison among Net Promoter Score and website (*p < 0.5).*

NPS	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.21	.02*	

Us.E. 2.0 scores were significantly different between websites. Specifically, Duncan post-hoc testing showed that scores for the website 2 was significantly lower than the other two for the dimensions (Mental) “Handling” ($F_{2,30} = 6.26$, $p < .01$) and “Satisfaction” ($F_{2,30} = 7.24$, $p < .01$). No significant differences were found for the dimension “Attractiveness” (see Figure 4.4.5.).

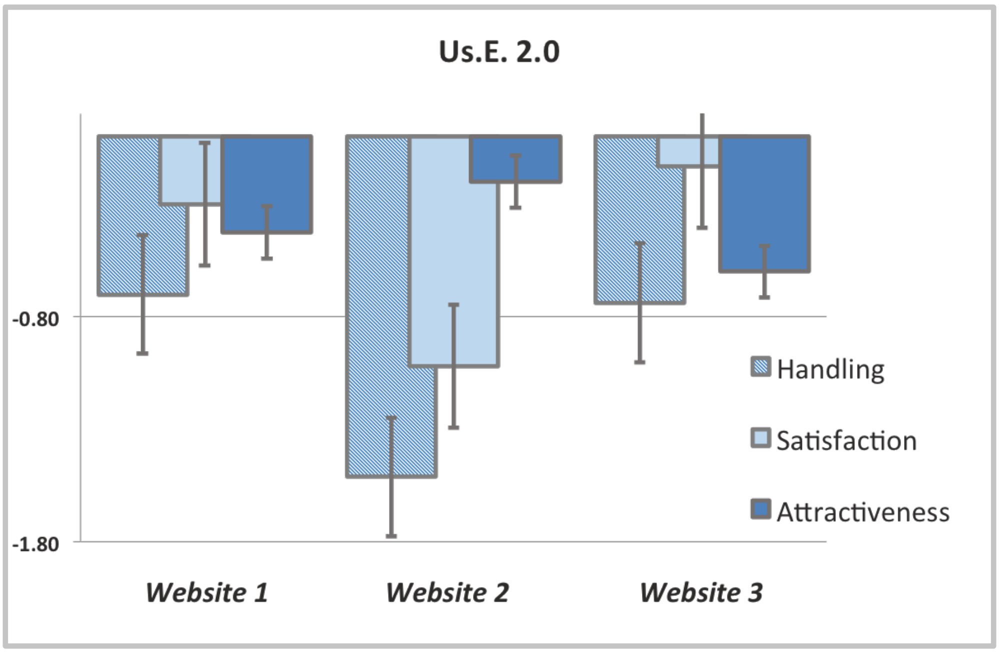


Figure 4.4.5. Us.E. 2.0 scores per website.

*Table 4.4.5. Pairwise comparison among Us.E. 2.0 (Handling dimension) and website (*p < 0.5).*

Handling (Us.E.)	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.88	.01*	

*Table 4.4.6. Pairwise comparison among Us.E. 2.0 (Satisfaction dimension) and website (*p < 0.5).*

Satisfaction (Us.E.)	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.50	.01*	

SOUP scores were significantly different between websites. Specifically, Duncan post-hoc testing showed that scores for the website 2 was significantly lower than the other two for the dimensions (Mental) "Handling" ($F_{2,30} = 6.27$, $p < .01$) and "Satisfaction" ($F_{2,30} = 6.83$, $p < .01$). No significant differences were found for the dimension "Attractiveness" (see Figure 4.4.6.).

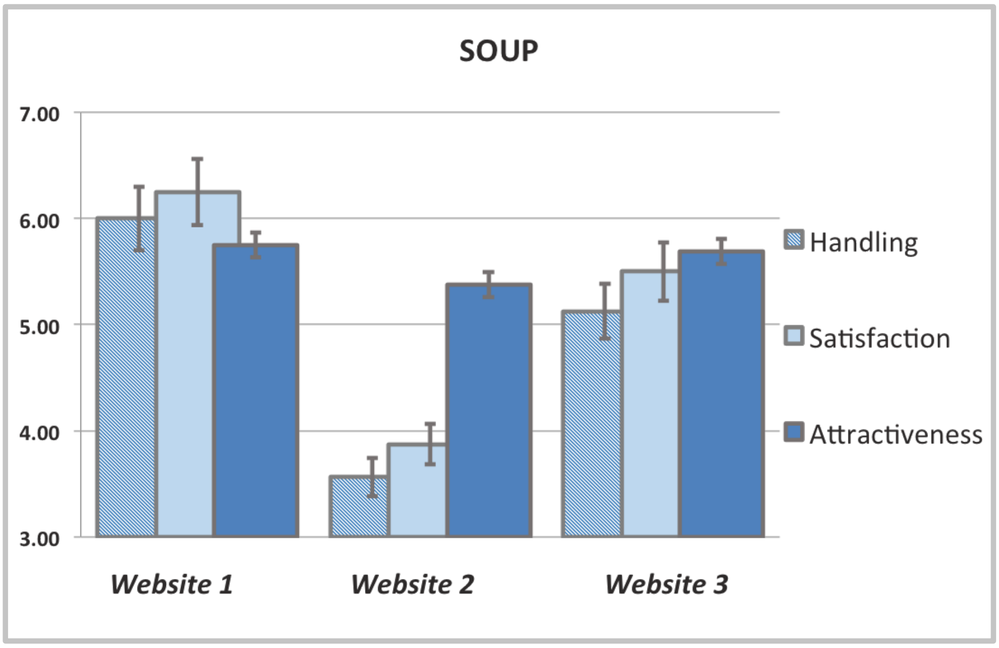


Figure 4.4.6. SOUP scores per website

*Table 4.4.7. Pairwise comparison among SOUP (Handling dimension) and website (*p < 0.5).*

Handling (SOUP)	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.21	.03*	

*Table 4.4.8. Pairwise comparison among SOUP (Satisfaction dimension) and website (*p < 0.5).*

Satisfaction (SOUP)	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.26	.01*	

SUS score was significantly different between websites ($F_{2,30} = 3.93$; $p < .05$). Duncan post-hoc testing showed that scores for website 2 were significantly lower than the other two (see Figure 4.4.7.).

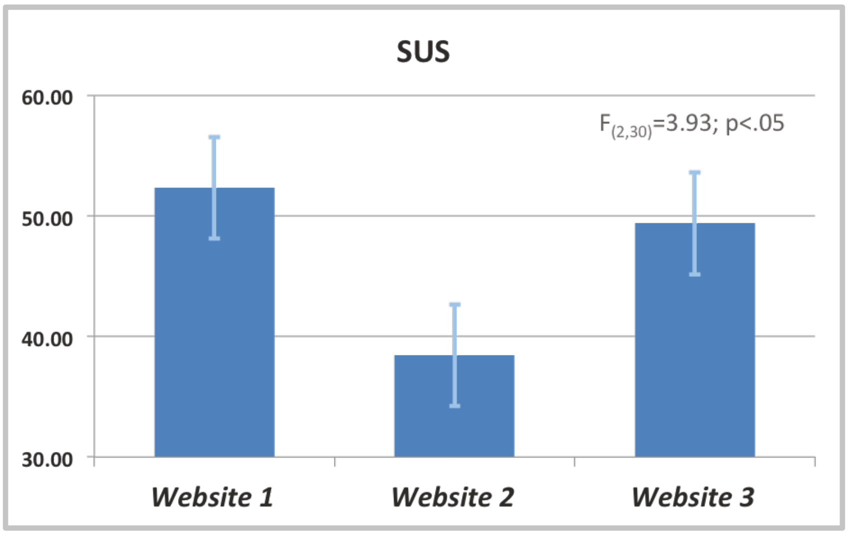


Figure 4.4.7. SUS scores per website.

*Table 4.4.9. Pairwise comparison among SUS and website (*p < 0.5).*

SUS	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.57	.04*	

NASA-TLX score was significantly different between websites ($F_{2,30} = 3.72$; $p < .05$). Duncan post-hoc testing showed that perceived workload for website 2 was significantly higher than the other two (see Figure 4.4.8.).

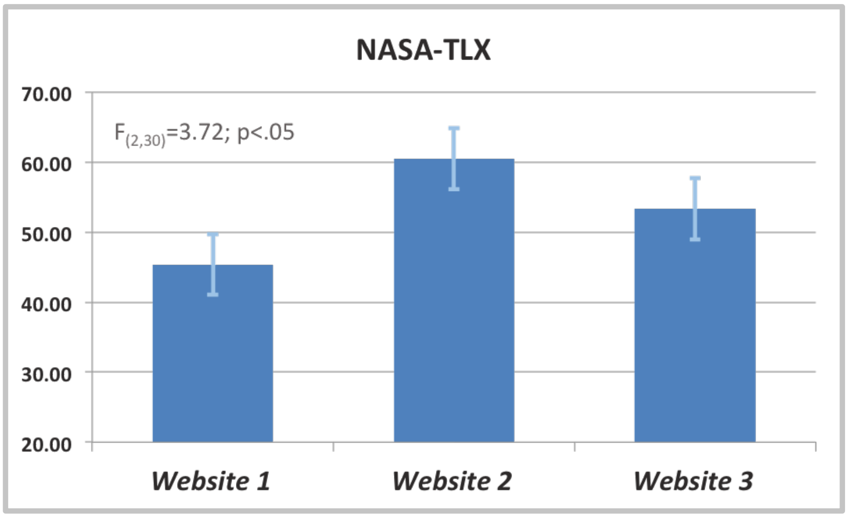


Figure 4.4.8. NASA-TLX scores per website.

*Table 4.4.10. Pairwise comparison among NASA-TLX and website (*p < 0.5).*

NASA-TLX	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.16	.20	

The Nearest Neighbour Index was significantly different between websites ($F_{2,30}$ = 23.9; p < .01). Duncan post-hoc testing showed that the fixation pattern of the medium-complexity websites (websites 2) was significantly more clustered than the other websites (see Figure 4.4.9.).

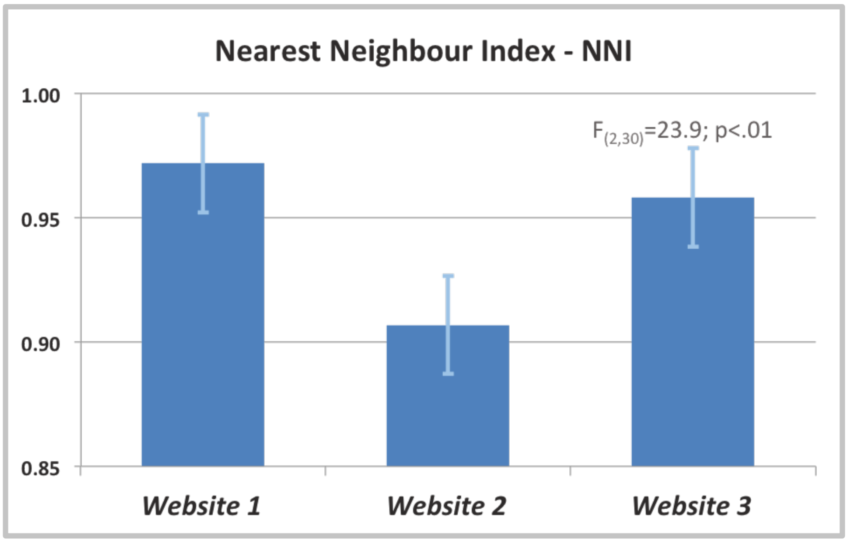


Figure 4.4.9. Nearest Neighbour Index per website.

*Table 4.4.11. Pairwise comparison among NNI and website (*p < 0.5).*

NNI	Website 1	Website 2	Website 3
Website 1			
Website 2	.01*		
Website 3	.22	.01*	

Discussion Study 4

In accordance with the findings of Study 3, the results of this study showed consistency between usability and mental workload measures.

The analysis of performance measures showed that the success rate was higher and completion time shorter for the website 1 than the other two. In this study, website 3 (high complexity) had a higher success rate than website 2. This outcome is supported by the fact that website 2 had several usability problems in its "mobile" version (i.e. poor responsiveness, poor visibility of functions, inconsistent layouts). These problems influenced the performance of the participants, hindering the correct execution of the tasks. A fact that has been reflected in the evaluations of perceived complexity, usability and MWL. Regarding "perceived complexity", the analysis showed that website 2 was the most complex to browse for the subjects involved in this study.

Consistent results can be found in the usability evaluations expressed through the NPS, the Us.E. 2.0, the SOUP and the SUS questionnaires. The analysis showed statistically significantly lower values for website 2 in all the usability questionnaires, while higher values were reported for the websites 1 and 3. A narrow focus on the "Attractiveness" scales of the Us.E. and the SOUP showed no differences between websites, once again reinforcing the diagnostic validity of these tools considering that the three websites had similar aesthetic features.

Regarding perceived MWL, the analysis of NASA-TLX scores highlights that participants perceived a higher MWL while browsing the website 2. The same results emerged through the analysis of subjects' ocular behaviour. In fact, results showed higher NNI values for website 1 and website 3, while lower values were reported for website 2. As previously noted, this in terms of visual exploration strategies a less clustered fixations pattern for website 1 and website 3 and, on the contrary, a more clustered fixations pattern for website 2. Therefore, following the NNI interpretation given by Camilli and colleagues (Camilli, Terenzi & Di Nocera, 2008) participants have experienced a greater mental workload while browsing the website 2.

Overall the results showed that the websites associated with lower levels of mental workload received more positive usability evaluations.

General discussion

Despite the growing interest of the scientific community in issues such as "usability" and "user experience" of web services, the relationship between the complexity of information architecture, perception of usability and mental workload imposed on the user is still not sufficiently investigated. The lack of integration between usability research and mental workload research is mainly due to different fields of application and different user classifications included in the experimental designs.

At first, HCI researchers and designers tried to profile the “typical user” or “average user” with the aim of identifying a series of needs useful to designing modes of interaction with an interface (Johnson, 2007; Norman, 1986). Later, due to the exponential spread of interfaces in daily life, it was necessary to better understand and identify the needs of all the different users who can interact with a given technology in specific contexts. Researchers thus began to distinguish users according to different criteria, such as “system knowledge”, “frequency of use”, “task knowledge”, “motivation in using the system”.

Since the 1970s several user definitions have been proposed (see Carrillo et al. 2017 for review).

In 1981 Schneider (Schneider, 1982) assumed the existence of five types of user based on the user's level of skills and knowledge of the interface. At the lowest level, there are people who use the system without understanding what they are doing, i.e. “parrots”. Continuing on, it is possible to distinguish the “novice”, the “intermediate” and the “expert” users, who have a deeper and deeper knowledge of the system. At the highest level are the "master" users who completely control and manage the system and all its functions.

Nielsen (1994a) proposed an analysis of users based on different dimensions such as “domain knowledge”, “computing experience” and “application experience”. Nielsen classified users into three main categories: “novice”, “expert” and “casual” users. While novice users have no knowledge of the system and they need to learn how it works from scratch, casual users have already used the system before, so they are required to remember rather than learn.

Phil Marsden and Erik Hollnagel (1996) applied the intention to use a system as a classification criterion and they talked about the so-called “accidental user”. The accidental user represents an individual who is “forced” to use technology to achieve a goal but would prefer to achieve it in different ways. From his perspective, the system is perceived more as an obstacle than an aid in achieving the goal (Lewin, 1951).

Turoff (1997) has distinguished different types of users including “novice”, “casual”, “experienced”, “intermediaries”, “frequent” users and “operators”, stressing the importance that a deep knowledge of the user has for the design of interfaces. In Turoff classification, the motivation to use the system plays a fundamental role that will influence the user's future interactions with the interface. While the casual user uses the system sporadically and has no ambition to master it, the “operators” perform a high degree of repetitive work over long periods of time and they have received specific training on the use of the interface, in order to optimize their performance and avoid errors.

More recently Carrillo et al. (2017) have described the “occasional user” as the kind of user who has poor technical knowledge of the interface, and whose purpose is to use it to achieve a certain goal while saving cognitive and temporal resources. This type of user is also not interested in learning how to use the system in-depth, as he may not use it in the future.

In summary, on the one hand, all these definitions place inexperienced or occasional users, on the other hand, operators and expert users. While the first ones have been the focus of User eXperience (UX) studies, the second ones have been more involved in Human Factor (HF) research.

Research involving occasional users focused on the user experience that emerges from the interaction with an interface. The topics of greatest interest in UX research are “User satisfaction”, “Ease of use”, “Consistency”, “Affordance”. The researchers’ efforts are oriented towards the design of intuitive and easy to use interfaces that support the user in achieving their goals. In this field, the concept of “usability” emerged.

On the other hand, Human Factors research focused on the interaction between expert users and interfaces in “high complexity” work environments (e.g: aviation, aerospace, healthcare, etc.). Some of the topics of greatest interest in HF research are “human error”, “situation awareness”, “automation” and the “mental workload” experienced by an operator who interacts with a system. In this field, the concept of mental workload has been used in an attempt to assess the operators’ spare capacity.

Despite the different fields of application, HF and UX share goals (i.e. increasing performance, decreasing errors and cognitive load), experimental techniques and methodologies. Both 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. Today, the challenge for the scientific community is to unify these lines of research and to identify reliable metrics that can provide objective information on these phenomena and their relationship. In this regard, eye movements analysis is a promising measure. Availability of unobtrusive eye-tracking systems allowed researchers to use indices of ocular activity as a measure of the operator mental workload and of the usability of an interface. The most common ocular metrics refer to saccades and fixations, or to the entire path of visual exploration (scanpath). The interpretation of metrics such as the number and average duration of fixations, the amplitude of saccades and the analysis of transitions between certain areas of interest seems to go in the same direction both for usability aspects and for aspects related to the investment and saturation of individual cognitive resources. A practical example of the overlap between usability and mental workload can be found in the fact that during the interaction between an individual and a system, more and longer fixations are associated with poor usability and high mental workload (Callan, 1998; Ehmke & Wilson, 2007; Findlay & Kapoula, 1992; Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Habuchi, Kitajima & Takeuchi, 2008; Kotval & Goldberg, 1998; Moffitt, 1980; Wang et al., 2018). Another example derives from the interpretation given to the analysis of transitions between areas of interest, as transitions between non-contiguous and distant areas of interest are associated with less efficient research strategies, poor perceived usability and increased mental workload (Cowen, Ball, & Delin, 2002; Ehmke & Wilson, 2007; Fitts, Jones & Milton, 1950; Goldberg & Kotval, 1999; Poole et al., 2005).

With that in mind, the present study aims to evaluate the workload imposed on the user by complex interfaces like "information abundant" websites (i.e. websites containing a large quantity and variety of information). It has been hypothesized that a greater complexity of the information architecture could be correlated with the higher mental workload and poor usability evaluations.

The idea that the complexity of information architecture can influence both the cognitive load imposed on users and their perceptions related to the pleasure of the user experience has been already suggested in the literature (e.g. Conklin, 1987). However, there is a lack of experimental research on this issue.

Three Italian government websites with different levels of information complexity have been selected to test the research hypothesis. Results indicate a correspondence between usability measures and mental workload measures. Specifically, the websites associated with lower levels of mental workload (assessed by both objective and subjective measures) received more positive usability evaluations and were associated with a greater success rate in the assigned search tasks.

The analysis of eye movements showed statistically higher NNI values for websites that received more positive usability assessments, while lower values were reported in cases of perceived poor usability. If we consider the visual exploration strategies, this result highlights a less clustered fixations pattern associated with good usability and, on the contrary, a more clustered fixations pattern associated with poor usability. Based on previous research (specifically: Camilli, Terenzi & Di Nocera, 2008), when a task imposes high visual-spatial demand on the user -as in the case of an information search task- fixations clustering (i.e. smaller NNI values) corresponds to a greater mental workload experienced by the user. Therefore, the subjects involved in this study have experienced a greater mental workload while browsing poor usability websites. This result is also supported by the subjective evaluations expressed by the NASA-TLX, the NPS, the SUS and the Us.E. questionnaires and by the analysis of the performance measures. Indeed, in both studies, the success rate was higher, and completion time shorter for websites associated with a lower mental workload and greater usability. Research on MWL (Eggemeier & Wilson, 1991; Sirevaag et al., 1993; Rubio et al., 2004) showed that an increase in cognitive effort is generally associated with a greater number of errors and a lower success rate. Similarly, usability evaluations of a website are generally negative when users take too long to complete the task, make mistakes or fail in its execution (Nielsen & Levy, 1994; Nielsen, 1999; Palmer, 2002). Moreover perceived usability, assessed through the Us.E. 2.0. and the SOUP scales showed the “Handling” scale (which is related to the information architecture and to the cognitive demand imposed by the task) was the most troublesome.

If we analyse the data of the two studies separately, we can notice some differences between the results obtained for the desktop and mobile version of the selected websites.

In fact, despite Study 3 reporting an exact correspondence between the complexity of information architecture, mental workload and usability, this correspondence did not emerge in Study 4. Specifically, in Study 3 the website characterised by a higher complexity of information architecture was also the website where subjects experienced a higher mental workload and for which they reported more unfavourable usability judgments. This result emerged both from the analysis of objective measures (eye movements, performance) and the analysis of subjective measures (self-report questionnaires). In Study 4, however, data analysis did not show any significant difference between the less complex website and the more complex website, either in terms of mental workload or perceived usability. This result was probably influenced by elements related to usability. In fact, despite the selected websites having the same information architecture for the desktop and mobile versions, in the mobile version, website 1 and, even more so, website 2 reported serious usability problems that influenced the performance of the participants. Generally, desktop and mobile versions of a website differ in the way information, menus and images are arranged. Responsive website design facilitates user navigation by adapting website content to the size and technology of the devices. One consequence of this adaptation is that in mobile devices, which have little space to display information compared to desktop devices, "secondary" information is shown at the bottom of the page, after presenting the main content. In the desktop version of websites, on the other hand, all information, primary and secondary, is always accessible to the user. As a result, secondary information can sometimes act as a distractor, increasing the number of options available and thus complicating the users' decision-making processes. Another important difference between desktop and mobile versions of websites can be found in the menu layout. In the desktop version, complex websites have extremely long contextual menus where all categories are always displayed. These long lists are difficult to read and can induce skimming behavior (Fitzsimmons, Weal & Drieghe, 2014) resulting in some useful items to accomplish the task not being read. In the mobile version, however, given the size of the interface, the menu is only partially visible. This peculiarity "forces" the user to read all the items while scrolling through the long list of available options. In this way the user will not lose pieces of information that are important for his goals. We can therefore infer that, given the same complexity of the information architecture, the layout of mobile devices would seem to be more usable for at least two reasons: i) the non-invasiveness of secondary information and ii) the induction of a more careful and accurate reading behavior.

These considerations should obviously be weighed according to the actual usability of the website. In fact, if content is not properly optimized for mobile browsing, even websites with simple information architecture can be complex to navigate. A bad content optimization requires a continuous effort to the users, with zoom and horizontal scroll behaviors (Paternò, Schiavone & Conti, 2017), and may lead they to abandon the website.

Although studies conducted show an association between mental workload and perceived usability, they do not tell us what the actual effect of mental workload is on usability ratings. To obtain this kind of information, it would be useful to set up an experimental setting that would allow us to independently manipulate usability and mental workload. This would allow us to measure changes in perceived usability caused by an increase/decrease in mental workload while keeping usability conditions unchanged. What we can infer from the studies conducted is that usability and mental workload have a lot in common, and in some ways are two overlapping constructs. In fact, talking about usability means paying attention to the efficiency, effectiveness, and satisfaction aspects of an individual. Several studies have found that mental workload plays an important role in determining the efficiency and effectiveness of a task (Lysaght et al., 1989; Xie and Salvendy, 2000; Young and Stanton, 2002). In addition, the individual who experiences a high mental workload will tend to be less satisfied as they are more vulnerable to failure and frustration (Young, Brookhuis, Wickens, & Hancock, 2015).

In the specific case of interacting with digital interfaces, the same elements that influence mental workload, such as a complex information architecture, are associated with poor perceived usability.

"Information architecture" in particular, refers to the structural planning of information within a website. It encompasses three systems: 1) a system for organizing content (organization system), 2) a system that allows users to move from one page to another (navigation system), and 3) a system for labelling the menus and services offered by the site (labelling system) (Garrett, 2010, Rosenfeld & Morville, 2006).

The qualitative and quantitative analysis of the problems encountered during the tests highlighted that Information Architecture (Morville & Rosenfeld, 2006) has a significant impact both in the usability assessments and in determining the mental workload of users. In all the websites examined, the participants have repeatedly experienced feelings of loss ("I don't know how I got here"; "I would like to go back to the previous section but I don't know how”). Moreover, they have taken entirely wrong paths. They have also been confused by the overabundance of options. Finally, they had difficulties in understanding the content of specific labels (“I didn't expect to find this information here”; “I would never have clicked on this link to search for this type of service”). These problems underline the importance of Information Architecture in determining the usability of a website and the related user’ mental workload. To ensure a comfortable user experience, including users in the design and implementation of service and menu labelling systems, and in the content classification and organisation system is a must.

In order to reduce the mental workload imposed on the user, these three systems should be designed following rules and principles that take into account the context of use and the real needs of the user. Regarding the organization system, organizing and classifying contents in a hierarchical way (where some contents are accessible only following a rigid and predetermined path) is useful in websites containing little information and intended for a mostly homogeneous user base, on the contrary, regarding complex sites such as those of PAs, a multifaceted classification is preferable that allows access to a given information through multiple paths (Ruzza et al., 2017).

Similarly, the navigation system should make navigation paths visible and clear, making the user understand at all times where he/she is and what he/she can do on that page/section. Finally, the labelling system should reduce ambiguity and use clear and simple language that respects the users' mental model (lo Storto, C. (2013).

The evaluation of the cognitive processes involved during the interaction between users and digital interfaces has a crucial role both in the design phase and in their usability assessments. HCI research should improve human-machine interaction through the implementation of systems that respect the cognitive processes of the end-users, allowing them to achieve their goals with effectiveness, efficiency and satisfaction.

5. Concluding remarks

The work discussed in this document primarily concerns the possibility of integrating the mental workload assessment into the usability evaluations. In fact, despite the growing interest of the scientific community in issues such as "usability" and "user experience" of web services, the relationship between complexity of information architecture, perception of usability and mental workload imposed on the user is still not sufficiently investigated. A still open research question concerns how the mental workload impacts on usability perceptions and, vice versa, how usability aspects affect the cognitive load experienced by the individual. In this work the use of ocular metrics in UX and HF fields has been analysed. The interpretation of the ocular metrics used in both the usability and mental workload studies shows that perceived usability is somehow influenced by aspects of cognitive processing related, for example, to the processes of classification, categorization and comprehension of stimuli. These processes are associated with the "mental demand" imposed by interaction, a dimension that has a crucial impact on the mental workload (NASA, 1986; Hart & Staveland, 1988). Some recent studies (Fedele et al., 2017; Longo & Dondio, 2015; Kokini et al., 2012) have attempted to answer this question through experimental studies. However, the results are discordant. In fact, while Kokini and collaborators (Kokini et al., 2012) have found that an increase in workload negatively affects perceived usability, Longo and Dondio (Longo & Dondio, 2015) state that there is no relationship between the two constructs and that they should, therefore, be considered separately. Fedele's research group (Fedele et al., 2017) has instead found that positive interaction experiences are associated with less mental workload.

The present work proposes a metric based on the scanpath analysis, the Nearest Neighbour Index, able to return a real time measurement of the mental workload experienced during the use of digital interfaces. The results highlight its validity as an indicator of mental workload under different conditions of task complexity and its association with perceived usability.

Several studies in cognitive science have shown that eye behavior is closely related to users' decision-making, reasoning, and cognitive processing (Di Stasi et al., 2011; Liu & Chuang, 2011; Staub & Rayner, 2007). Eye movement analysis provides important insights into how to design websites with low mental workload.

Consistent with the relevant literature, designers should monitor a few key dimensions:

- *Visual complexity* - web pages should be carefully designed so as not to confuse users. The length of text, the number of images, links and animations should be reduced in order to make it easier for the user (Wang et al., 2014);
- *Language* - the language used should be simple and intuitive. Terminology that does not respect the user's mental model should be avoided. In addition, the text should be structured following specific guidelines so that attention is paid to text length, spacing, and font size (Bernard et al., 2002; Lo Storto, 2013);
- *Visibility and location of links and "call to action"* - the user must be able to recognize links and actions that can be performed within a given page. Similar functional elements must be placed close together in the interface to prevent the user from losing the flow of the process while performing a task (Fitts et al., 1950; Goldberg & Kotval, 1999; Cowen, Ball, & Delin, 2002; Poole et al., 2005; Ehmke & Wilson, 2007).

Finally, some considerations must precede the discussion of this work. The participants at Study 3 and Study 4 were between 46 and 65 years old. Moreover, all the participants were Government employees. Although the cognitive processes investigated are common to all individuals, it could be useful to involve different types of users (e.g. experts vs. naïves, typical vs. occasional users, etc.). Aging-related issues such as decline in sensory, motor, and cognitive abilities play an important role in the ability to use digital interfaces. Differences between the performance of young and adult users in web search tasks have been highlighted in the reference literature. For example, Rogers et al. (2005) investigated this difference in relation to mobile device use. Their study involved participants aged 18-28, and participants aged 51-65. Results showed that older participants experienced more difficulty performing certain behaviors such as pointing and scrolling (Rogers, Fisk, McLaughlin & Pak, 2005). Another study by Al-Showarah et al. (Al-Showarah, Al-Jawad & Sellahewa, 2013) revealed that older users were less efficient in navigating smartphone applications/interfaces than younger users. Another important problem was identified by a recent study by Joseph and Murugesh (Joseph & Murugesh, 2021). The authors found that a mismatch between the visual elements of the interface and the users' mental model can lead to poor task performance and increased cognitive load. For these reasons, the scarce heterogeneity of the experimental sample in terms of age and occupation, may have influenced the experimental results. A second limitation refers to the participants' ability to use the experimental devices: although all the participants declared access to the Internet daily through mobile devices (e.g., smartphones, tablets), during the course of the experiment several problems emerged. For example, some users did not understand how to close a menu or experienced difficulty in finding functions that they typically used when browsing the Internet from a desktop computer. From this limitation clear difficulties emerge in assessing how much the perceptions of usability and MWL have been influenced by elements of the interaction typical of the used device (e.g., touch, zoom) or by the structure of the selected websites.

In conclusion, the results of this study, although not conclusive, underline the need to continue working to improve human-machine interaction, highlighting the importance of also integrating an estimate of the users' mental workload during the design and evaluation of usability of highly complex websites.

The involvement of a more heterogeneous sample, representing different types of users, and the use of evaluation tools tailored to the particular type of the websites evaluated could allow to greater explore in-depth the different dimensions of usability and, therefore, lead to more reliable estimates for the identification of the relationship between information architecture, mental workload and perceptions of usability.

6. References

Ahmad, A., & Khan, M. N. (2017). Developing a website service quality scale: A confirmatory factor analytic approach. Journal of internet Commerce, 16(1), 104-126.

Albert, W., & Tullis, T. (2013). Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes.

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The development and application of models to predict operator workload during system design. In Applications of human performance models to system design (pp. 65-80). Springer, Boston, MA.

Al-Showarah, S., Al-Jawad, N., & Sellahewa, H. (2013, October). Examining eye-tracker scan paths for elderly people using smart phones. In *Sixth York Doctoral Symposium on Computer Science & Electronics* (Vol. 1, p. 7).

Ames, L.L., & George, E.J. (1993). Revision and verification of a seven-point workload estimate scale (No. AFFTC-TIM-93-01). Air Force Test Center Edwards AFB CA.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis.

Andreassi, J. L. (2013). Psychophysiology: Human behavior & physiological response. Psychology Press.

Annett, J. (2002). Subjective rating scales in ergonomics: A reply. Ergonomics, 45(14), 1042-1046.

Artz, M. (2017). NPS—The One Measure You Really Need to Grow?. Controlling & Management Review, 61(1), 32-38.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Psychology of learning and motivation (Vol. 2, pp. 89-195). Academic Press.

Baddeley, A. (1992). Working memory. Science, 255(5044), 556-559.

Baber, C. (2002). Subjective evaluation of usability. Ergonomics, 45(14), 1021-1025.

Backs, R. W., Ryan, A. M., & Wilson, G. F. (1994). Psychophysiological measures of workload during continuous and manual performance. Human Factors, 36, 514-531.

Bahill, A. T., Clark, M. R., & Stark, L. (1975). The main sequence, a tool for studying human eye movements. Mathematical biosciences, 24(3-4), 191-204.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction, 24(6), 574-594.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, 4(3), 114-123.

Barnes, S.J. & Vidgen, R. (2001). An Evaluation of Cyber-Bookshops: The WebQual Method, Journal of Electronic Commerce 6(1), 11-30.

Barnes, S., & Vidgen, R. (2002). An integrative approach to the assessment of e-commerce quality. Journal of Electronic Commerce Research, 3(3).

Barnes, S., & Vidgen, R. (2005). Data Triangulation in action: using comment analysis to refine web quality metrics. 13th European Conference on Information Systems, Regensburg , Germany , May 26–28.

Battiste, V., & Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. In Proceedings of the Human Factors Society Thirty-Second Annual Meeting (pp. 150 –154). Santa Monica, CA: Human Factors Society.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. Handbook of psychophysiology, 2(142-162).

Becker, W., 1989. saccadic eye movements as a control system: metrics. In: Wurtz, R.H., Goldberg, M.E. (Eds.), Reviews of Oculomotor Research. The Neurobiology of saccadic Eye Movements, vol. 3. Elsevier, Hillsdale, pp. 13e67.

Bergstrom, J. R., & Schall, A. (Eds.). (2014). Eye tracking in user experience design. Elsevier.

Bevan, N., & Macleod, M. (1994). Usability measurement in context. Behaviour & information technology, 13(1-2), 132-145.

Bojko, A. (2013). Eye tracking the user experience: A practical guide to research. Rosenfeld Media.

Boles, D. B., & Adair, L. P. (2001). The multiple resources questionnaire (MRQ). In Proceedings of the human factors and ergonomics society annual meeting (Vol. 45, No. 25, pp. 1790-1794). Sage CA: Los Angeles, CA: SAGE Publications.

Boles, D. B., Bursk, J. H., Phillips, J. B., & Perdelwitz, J. R. (2007). Predicting dual-task performance with the Multiple Resources Questionnaire (MRQ). Human factors, 49(1), 32-45.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. International Journal of Human-Computer Interaction, 31(8), 484-495.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: a test of alternative measurement models. Cognitive processing, 10(3), 193-197.

Borsci, S., Kurosu, M., Federici, S., & Mele, M. L. (2013). Computer systems experiences of users with and without disabilities: an evaluation guide for professionals. CRC Press.

Boscarol, M. (2003). Ecologia dei siti web. Come e perché usabilità, accessibilità e fogli di stile stanno cambiando il modo di realizzare i siti internet. Tecniche Nuove.

Boscarol, M. (2010). Come scegliere il tipo di test di usabilità per il nostro progetto. Retrieved from: https://www.usabile.it/512010.htm (Accessed 20 November 2020).

Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.

Bradshaw, J. L. (1968). Load and pupillary changes in continuous processing tasks. British Journal of Psychology, 59(3), 265-271.

Brennan, S. D. (1992). An experimental report on rating scale descriptor sets for the instantaneous self assessment (ISA) recorder. Portsmouth: DRA Maritime Command and Control Division. DRA Technical Memorandum (CAD5), 92017.

Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology,* 42,361-377.

Brookhuis, K. A., & De Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. Accident Analysis & Prevention, 42(3), 898-903.

Buschke, L. (1997). The basics of building a great Website. *Training & Development*, 51(7), 46-49.

Byers, J.C., Bittner, A.C., Hill, S.G., Zaklad, A.L., & Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. In Proceedings of the Human Factors Society.

Byrne, M. D., Anderson, J. R., Douglass, S., & Matessa, M. (1999, May). Eye tracking the visual search of click-down menus. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 402-409). ACM.

Cain, B. (2007). A review of the mental workload literature. Defence Research And Development Toronto (Canada).

Callan, D. J. (1998). Eye movement relationships to excessive performance error in aviation. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 42, No. 15, pp. 1132-1136). Sage CA: Los Angeles, CA: SAGE Publications.

Camilli, M., Terenzi, M., & Di Nocera, F. (2007). Concurrent validity of an ocular measure of mental workload. In D. de Waard, G.R.J. Hockey, P. Nickel, and K.A. Brookhuis (Eds.), *Human Factors Issues in Complex System Performance* (pp. 117-129). Maastricht, the Netherlands: Shaker Publishing.

Camilli, M., Terenzi, M., & 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,* 52(18), 1248-1251.

Card, S. K., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction, Lawrence Erlbaum. Hillside, NJ.

Carrillo, A. L., Martinez, S., Falgueras, J., & Scott-Brown, K. C. (2017). A reflective characterisation of occasional user. Computers in Human Behavior, 70, 74-89.

Carroll, J.M. (1997). Human-computer interaction: Psychology as a science of design. International Journal of Human-Computer Studies, 46(4), 501-522.

Castor, M. C. (2003). Garteur handbook of mental workload measurement. Flight Mechanism Action Group FM-AG13, GARTEUR, Group for Aeronautical Research and Technology in Europe.

Chain Store Age, (1997). Web-based retailers tell disparate tales at NRF. *Chain Store Age*, 45-52.

Chalmers, P. A. (2003). The role of cognitive theory in human–computer interface. Computers in Human Behavior, 19(5), 593–607.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Applied ergonomics*, *74*, 221-232.

Chatterjee, A., Southwood, M. H., & Basilico, D. (1999). Verbs, events and spatial representations. Neuropsychologia, 37(4), 395-402.

Chen, L. S., & Chang, P. C. (2010). Identifying crucial website quality factors of virtual communities. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, pp. 17-19).

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988, May). Development of an instrument measuring user satisfaction of the human-computer interface. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 213-218). ACM.

Clark, P. J., & Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4), 445-453.

Cline D., Hofstetter H.W., Griffin J.R. (1996). Dictionary of Visual Science, 4th edition, Boston, Butterworth-Heinemann, p. 820.

Conklin, J. (1987). Hypertext: an introduction and SurvevJ. computer, 20(9), 17-41.

Corwin, W.H., Sandry-Garza, D.L., Biferno, M.H., Boucek, G.P., Logan, A.L., Jonsson, J.E., & Metalis, S.A. (1989). Assessment of crew workload measurement methods, techniques, and procedures: Process, methods and results. Report WRDC-TR-89-7006. Wright-Patterson Air Force Base, OH: Wright Research and Develop-ment Centre, Air Force Systems Command.

Cowen, L., Ball, L. J., & Delin, J. (2002). An eye movement analysis of web page usability. In People and Computers XVI-Memorable Yet Invisible (pp. 317-335). Springer, London.

Cowley, B. U., Filetti, M., Lukander, K., Torniainen, J., Helenius, A., Ahonen, L., ... & Ravaja, J. N. (2016). The psychophysiology primer: a guide to methods and a broad review with a focus on human-computer interaction. Foundations and Trends in Human-Computer Interaction.

Cummings, M. L., Myers, K., & Scott, S. D. (2006). Modified Cooper Harper evaluation tool for unmanned vehicle displays. In Proceedings of UVS Canada: Conference on Unmanned Vehicle Systems Canada.

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009, July). Eye movement as indicators of mental workload to trigger adaptive automation. In International Conference on Foundations of Augmented Cognition (pp. 219-228). Springer, Berlin, Heidelberg.

De Waard, D. (1996). *The measurement of drivers' mental workload*. Netherlands: Groningen University, Traffic Research Center.

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

Di Nocera, F., Camilli, M., & Terenzi, M. (2006). Using the distribution of eye fixations to assess pilots' mental workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 1, pp. 63-65). Sage CA: Los Angeles, CA: Sage Publications.

Di Nocera, F., Camilli, M., & 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*, 1(3), 271-285.

Di Nocera, F., Ferlazzo, F., & Renzi, P. (1999). Us. E. 1.0: costruzione e validazione di uno strumento in lingua italiana per valutare l'usabilità dei siti Internet. HCITALY, 99.

Di Nocera, F., Ferlazzo, F., & Renzi, P. (2003). L'usabilità a quattro dimensioni. Ricerche di Psicologia, 26(4), 83-104.

Di Stasi, L. L., Álvarez-Valbuena, V., Cañas, J. J., Maldonado, A., Catena, A., Antolí, A., & Candido, A. (2009). Risk behaviour and mental workload: Multimodal assessment techniques applied to motorbike riding simulation. Transportation research part F: traffic psychology and behaviour, 12(5), 361-370.

Di Stasi, L. L., Antolí, A., & Cañas, J. J. (2011). Main sequence: An index for detecting mental workload variation in complex tasks. Applied Ergonomics, 30, 1e7.

Di Stasi, L. L., Antolí, A., Gea, M., & Cañas, J. J. (2011). A neuroergonomic approach to evaluating mental workload in hypermedia interactions. *International Journal of Industrial Ergonomics,* 41(3), 298-304.

Di Stasi, L. L., Marchitto, M., Antolí, A., Baccino, T., & Cañas, J. J. (2010a). Approximation of on-line mental workload index in ATC simulated multitasks. Journal of Air Transport Management, 16(6), 330-333.

Di Stasi, L. L., McCamy, M. B., Martinez-Conde, S., Gayles, E., Hoare, C., Foster, M., ... & Macknik, S. L. (2016). Effects of long and short simulated flights on the saccadic eye movement velocity of aviators. Physiology & Behavior, 153, 91-96.

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., ... & Pannasch, S. (2010b). Saccadic peak velocity sensitivity to variations in mental workload. Aviation, space, and environmental medicine, 81(4), 413-417.

Dillard, M. B., Warm, J. S., Funke, G. J., Funke, M. E., Finomore, Jr., V. S., Matthews, G., Shaw, T. H., & Parasuraman, R. (2014). The Sustained Attention to Response Task (SART) does not promote mindlessness during vigilance performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(8), 1364-1379.

do Amaral, V., Ferreira, L. A., Aquino, P. T., & de Castro, M. C. F. (2013, February). EEG signal classification in usability experiments. In 2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC) (pp. 1-5). IEEE.

Donchin, E., Kramer, A. F., & Wickens, C. (1986). Applications of brain event-related potentials to problems in engineering psychology.

Drachen, A., Nacke, L. E., Yannakakis, G., & Pedersen, A. L. (2010, July). Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (pp. 49-54).

Dubey, M., & Singh, L. (2016). Automatic emotion recognition using facial expression: a review. International Research Journal of Engineering and Technology (IRJET), 3(2), 488-492.

Duchowski, A. (2017). Eye Tracking Methodology: Theory and Practice, third edition. Springer: Cham, Switzerland.

Eggemeier, F.T., Wilson, G.F., Kramer, A.F. & Damos, D.L. (1991).Workload assessment in multi-task environments. In D.L. Damos(Ed.)., *Multiple-task performance*. (pp. 207-216). London: Taylor & Francis.

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1 (pp. 119-128). British Computer Society.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. Environmental psychology and nonverbal behavior, 1(1), 56-75.

Ephrath, A. R., Tole, J. R., Stephens, A. T., & Young, L. R. (1980). Instrument Scan—Is it an Indicator of the Pilot's Workload?. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 24, No. 1, pp. 257-258). Sage CA: Los Angeles, CA: Sage Publications.

Ercolani, A. P., Areni, A., & Leone, L. (2001). Statistica per la psicologia: Fondamenti di psicometrica e statistica descrittiva.-2001.-179 p. Il mulino.

Ericsson, K. A., & Simon, H. A. (1993). 1993: Protocol analysis: verbal reports as data, revised edition. Cambridge, MA: MIT Press.

Esmeria, G. J., & Seva, R. R. (2017, June). Web usability: a literature review. In DLSU Research Congress.

European Commission (2015). "*A Digital Single Market Strategy for Europe*", COM (2015) 192".

Fedele, G., Fedriga, M., Zanuso, S., Mastrangelo, S., & Di Nocera, F. (2017). Can User Experience affect buying intention? A case study on the evaluation of exercise equipment. In: D. de Waard, A. Toffetti, R. Wiczorek, A. Sonderegger, S. Röttger, P. Bouchner, T. Franke, S. Fairclough, M. Noordzij, and K. Brookhuis (Eds.). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference*. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Federici, S., Borsci, S., & Meloni, F. (2009). Le misure dell'usabilità: Studio sulle caratteristiche psicometriche del QUIS e del SUMI nella versione italiana. Giornale di, 3(2), 164.

Findlay, J. M., & Kapoula, Z. (1992). Scrutinization, spatial attention, and the spatial programming of saccadic eye movements. The Quarterly Journal of Experimental Psychology Section A, 45(4), 633-647.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22, 323–327.

Finstad, K. (2013). Response to commentaries on "The Usability Metric for User Experience." Interacting with Computers, 25, 327–330.

Fitts, P. M., Jones, R. E. & Milton, J. L. (1950), Eye Movements of Aircraft Pilots during Instrument-landing Approaches, *Aeronautical Engineering Review* 9(2), 24–29.

Fitzsimmons, G., Weal, M. J., & Drieghe, D. (2014, June). Skim reading: an adaptive strategy for reading on the web. In *Proceedings of the 2014 ACM conference on Web science* (pp. 211-219).

Ganglbauer, E., Schrammel, J., Deutsch, S., & Tscheligi, M. (2009, August). Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility. In *Workshop on user experience evaluation methods in product development*.

Garrett, J. J. (2010). *The elements of user experience: user-centered design for the web and beyond*. Pearson Education, London.

Goldberg, A. (1988). *A history of personal workstations*. ACM.

Goldberg, J. H., & Kotval, X. P. (1998). Eye movement-based evaluation of the computer interface. *Advances in occupational ergonomics and safety*, 529-532.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. International Journal of Industrial Ergonomics, 24(6), 631-645.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In *the Mind's Eye* (pp. 493-516). North-Holland.

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002). Eye tracking in web search tasks: design implications. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 51-58). ACM.

Gopher, D. & Donchin, E. (1986). Workload -an examination of the concept. In K.R. Boff, L. Kaufman & J.P. Thomas (Eds.), *Handbook of perception and human performance. Volume II, cognitive processes and performance.* (pp 41/1-41/49). New York: Wiley.

Graham, H., Cummings, M. L., Donmez, B., & Brzezinski, A. S. (2008). Modified Cooper Harper Scales for Assessing Unmanned Vehicle Displays. MIT Humans and Automation Laboratory.

Groth, A., & Haslwanter, D. (2015). Perceived usability, attractiveness and intuitiveness of responsive mobile tourism websites: a user experience study. In Information and Communication Technologies in Tourism 2015 (pp. 593-606). Springer, Cham.

Grudin, J. (1990, March). The computer reaches out: the historical continuity of interface design. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 261-268).

Habuchi, Y., Kitajima, M., & Takeuchi, H. (2008). Comparison of eye movements in searching for easy-to-find and hard-to-find information in a hierarchically organized information structure. In Proceedings of the 2008 symposium on Eye tracking research & applications (pp. 131-134). ACM.

Hankins, T.C. & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space & Environmental Medicine*, 69, 360-367.

Harbluk, J. L., Noy, Y. I., & Eizenman, M. (2002). The impact of cognitive distraction on driver visual behaviour and vehicle control (No. TP# 13889 E).

Harper, B. D., & Norman, K. L. (1993, February). Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5. In Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference (pp. 224-228).

Harper Jr, R.P., & Cooper, G.E. (1986). Handling qualities and pilot evaluation. Journal of Guidance, Control, and Dynamics, 9(5), 515-529.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139-183.

Hart, S. G., & Wickens, C. D. (1990). Workload assessment and prediction. In Manprint (pp. 257-296). Springer, Dordrecht.

Harris Sr, R. L., Glover, B. J., & Spady Jr, A. A. (1986). Analytical techniques of pilot scanning behavior and their application.

Hendy, K.C., Hamilton, K.M., & Landry, L N. (1993). Measuring subjective workload: when is one scale better than many?. Human Factors, 35(4), 579-601.

Hering, H., & Coatleven, G. (1996). ERGO (Version 2) For instantaneous self assessment of workload in a real-time ATC simulation environment. EEC Note, 10, 96.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611), 1190-1192.

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., ... & Verplank, W. (1992). ACM SIGCHI curricula for human-computer interaction. ACM.

Hill, S.G., Byers, J.C., Zaklad, A.L., Christ, R.E., & Bittner, A.C. (1988). Workload assessment of a mobile air defences system. In Proceedings of the Human Factors Society Thirty-Second Annual Meeting (pp. 1068 –1072). Santa Monica, CA: Human Factors Society.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad, A.L. and Christ, R.E. 1992, Comparison of four subjective workload rating scales, Human Factors, 34, 429-439.

Hoffman, H. G. (1998, March). Physically touching virtual objects using tactile augmentation enhances the realism of virtual environments. In Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998 (pp. 59-63). IEEE.

Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye Tracking. A comprehensive guide to methods and measures. Oxford University Press, New York.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. International journal of human-computer studies, 64(2), 79-102.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004, April). Task-evoked pupillary response to mental workload in human-computer interaction. In CHI'04 extended abstracts on Human factors in computing systems (pp. 1477-1480).

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005, April). Towards an index of opportunity: understanding changes in mental workload during task execution. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 311-320).

ISO 9241-11 (1998). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)–Part II Guidance on Usability.

ISO 9241-210 (2009). Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems.

Ishida, T., Ikeda, M. (1989). Temporal properties of information extraction in reading studied by a text- mask replacement technique. Journal of the Optical Society A: Optics and Image Sciences 6:1624–32.

Isreal, J. B., Wickens, C. D., Chesney, G. K., & Donchin, E. (1980). The event-related brain potential as an index of display-monitoring workload.Human Factors, 22, 211-224.

Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. Communications of the ACM, 26(10), 785-793.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573-605).

Johnson, M. (2007). Unscrambling the average user of Habbo hotel. Human Technology: An Interdisciplinary Journal on Humans in ICT Environments.

Joseph, A. W., & Murugesh, R. (2021). Eye-Tracking Evaluation of Age-Related Differences in User Behaviour on Mobile Applications. *Journal of Scientific Research*, 65(1).

Juris, M., & Velden, M. (1977). The pupillary response to mental overload. *Physiological Psychology*, 5(4), 421-424.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological review, 87(4), 329.

Kahneman, D. (1973). Attention and effort (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual deficit during a mental task. Science, 157(3785), 218-219.

Käthner, I., Wriessnegger, S. C., Müller-Putz, G. R., Kübler, A., & Halder, S. (2014). Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain–computer interface. *Biological psychology*, *102*, 118-129.

Katoh, Z. (1997). Saccade amplitude as a discriminator of flight types. Aviation, Space, and Environmental Medicine, 68(3), 205-208.

Kimura, M., Uwano, H., Ohira, M., & Matsumoto, K. I. (2009, July). Toward constructing an electroencephalogram measurement method for usability evaluation. In International Conference on Human-Computer Interaction (pp. 95-104). Springer, Berlin, Heidelberg.

Kirakowski, J., & Cierlik, B. (1998, October). Measuring the usability of web sites. In Proceedings of the Human Factors and Ergonomics Society annual meeting (Vol. 42, No. 4, pp. 424-428). Sage CA: Los Angeles, CA: SAGE Publications.

Kirakowski, J., Claridge, N., & Whitehand, R. (1998, June). Human centered measures of success in web site design. In Proceedings of the Fourth Conference on Human Factors & the Web.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. British journal of educational technology, 24(3), 210-212.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). Nonverbal communication in human interaction. Cengage Learning.

Knop, E., Knop, N., Zhivov, A., Kraak, R., Korb, D. R., Blackie, C., ... & Guthoff, R. (2011). The lid wiper and muco□cutaneous junction anatomy of the human eyelid margins: an in vivo confocal and histological study. Journal of anatomy, 218(4), 449-461.

Kokini, C. M., Lee, S., Koubek, R. J., & Moon, S. K. (2012). Considering context: The role of mental workload and operator control in users' perceptions of usability. *International Journal of Human-Computer Interaction*, 28(9), 543-559.

Kotval, X. P., & Goldberg, J. H. (1998). Eye movements and interface component grouping: An evaluation method. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 42, No. 5, pp. 486-490). Sage CA: Los Angeles, CA: SAGE Publications.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. Multiple-task performance, 279-328.

Krebs, M. J., Wingert, J. W., & Cunningham, T. (1977). Exploration of an oculometer-based model of pilot workload.

Kruizinga, A., Mulder, B., & de Waard, D. (2006). Eye scan patterns in a simulated ambulance dispatcher's task. Developments in human factors in transportation, design, and evaluation, 305-317.

Kun, A. L., Palinko, O., Medenica, Z., & Heeman, P. A. (2013, August). On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues. In INTERSPEECH (pp. 3766-3770).

Kurosu, M., & Kashimura, K. (1995, May). Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In Conference companion on Human factors in computing systems (pp. 292-293).

Kutas M, McCarthy G, Donchin E. (1997). Augmenting mental chronometry: P300 as a measure of stimulus evaluation time. Science. ;197:792–795.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology, 30(3), 261-273.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.

Lean, Y., & Shan, F. (2012). Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *22*(3), 177-187.

Lens, A., Nemeth, S. C., & Ledford, J. K. (2008). *Ocular anatomy and physiology*. Slack Incorporated.

Leplat, J., (1978). Accident analysis and work analysis. *Journal of Occupational Accidents* 1, 331–340.

Lewin, K. (1951). Field theory in social science: selected theoretical papers (Edited by Dorwin Cartwright.).

Lewis, C. (1982). Using the" thinking-aloud" method in cognitive interface design. Yorktown Heights, NY: IBM TJ Watson Research Center.

Lewis, J. R. (2014). Usability: Lessons learned ... and yet to be learned. International Journal of Human-Computer Interaction, 30, 663–684.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), Human centered design (Vol. 5619, pp. 94–103). Berlin, Germany: Springer.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: when there's no time for the SUS. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2099-2102).

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. International Journal of Human-Computer Interaction, 31(8), 496-505.

Liu, H. C., & Chuang, H. H. (2011). An examination of cognitive processing of multimedia information based on viewers' eye movements. *Interactive Learning Environments*, 19(5), 503-517.

Longo, L., & Dondio, P. (2015, December). On the relationship between perception of usability and subjective mental workload of web interfaces. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 345-352). IEEE.

lo Storto, C. (2013). Evaluating ecommerce websites cognitive efficiency: An integrative framework based on data envelopment analysis. *Applied ergonomics*, 44(6), 1004-1014.

Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. Ergonomics, 44(3), 229-243.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., & Linton, P. M. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies (No. TR-2075-3). ANALYTICS INC WILLOW GROVE PA.

Maier, A., Baltsen, N., Christoffersen, H., & Störrle, H. (2014). Towards diagram understanding: A pilot study measuring cognitive workload through eye-tracking. In International Conference on Human Behaviour in Design 2014 (pp. HBiD2014-114).

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing* (pp. 39-65). Springer, London.

Mandal, F. B. (2014). Nonverbal communication in humans. Journal of human behavior in the social environment, 24(4), 417-421.

Mandel, T. (1997). Elements of user interface design. New York, NY: John Wiley & Sons.

Manning, C., Mills, S., Fox, C., & Pfleiderer, E. (2001). Investigating the validity of performance and objective workload evaluation research (POWER) (No. DOT/FAA/AM-01/10). Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Marsden, P., & Hollnagel, E. (1996). Human interaction with technology: The accidental user. Acta Psychologica, 91(3), 345-358.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on (pp. 7-7). IEEE.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854-2861.

Martinez-Conde S, Macknik SL & Hubel DH (2004). The role of fixational eye movements in visual perception. Nature Reviews Neuroscience; 5, 229-240.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica*, 75(1), 75-89.

McCarthy, G., & Donchin, E. (1981). A metric for thought: a comparison of P300 latency and reaction time. Science, 211(4477), 77-80.

McCracken, J., & Aldrich, T. B. (1984). Analysis of selected LHX mission functions. Implications for Operator Workload and System Automation Goals, Ft. Rucker, AL.

Miller, G. (1956). Human memory and the storage of information. IRE Transactions on Information Theory, 2(3), 129-137.

Moffitt, K. (1980). Evaluation of the fixation duration in visual search. Attention, Perception, & Psychophysics, 27(4), 370-372.

Moray, N. (1979). Models and measures of mental workload. In *Mental Workload* (pp. 13-21). Springer, Boston, MA.

Morris, T. L., & Miller, J. C. (1996). Electrooculographic and performance indices of fatigue during simulated flight. Biological psychology, 42(3), 343-360.

Morville, P., & Rosenfeld, L. (2006). *Information architecture for the World Wide Web: Designing large-scale web sites*. O'Reilly Media, Inc. Sebastopol, CA.

Myers, B. A. (1984). The user interface for Sapphire. IEEE Computer Graphics and Applications, 4(12), 13-23.

Nakayama, M., Takahashi, K., & Shimizu, Y. (2002). The act of task difficulty and eye-movement frequency for the'Oculo-motor indices'. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (pp. 37-42). ACM.

N. A. S. A., Human Performance Research Group. (1986). *NASA Task Load Index (TLX) v. 1.0: Paper and Pencil Package*. Moffett Field, CA: NASA Ames Research Center.

Nataupsky, M., & Abbott, T.S. (1987). Comparison of workload measures on computer-generated primary flight displays. In Proceedings of the Human Factors Society Thirty-First Annual Meeting (pp. 548 – 552). Santa Monica, CA: Human Factors Society.

Nielsen, J. (1994a). *Usability engineering*. Elsevier.

Nielsen, J. (1994b). *Estimating the number of subjects needed for a thinking aloud test*. International journal of human-computer studies, 41(3), 385-397.

Nielsen, J. (1994c). Usability inspection methods. In *Conference companion on Human factors in computing systems* (pp. 413-414). ACM.

Nielsen, J. (1999). *Designing web usability: The practice of simplicity.* New Riders Publishing.

Nielsen, J. (2001). Success rate: The simplest usability metric. Retrieved from https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/ (Accessed 20 November 2020).

Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. Communications of the ACM, 37(4), 66-76.

Nielsen, J., & Loranger, H. (2006). *Prioritizing web usability*. Berkeley, CA: New Riders Publishing.

Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 249-256).

Nielsen, J., & Pernice, K. (2010). *Eyetracking web usability*. New Riders.

Norman, D. A. (1986). Cognitive engineering. User centered system design, 31, 61.

North, R.A., & Riley, V.A. (1989). W/INDEX: A predictive model of operator workload. In Applications of human performance models to system design (pp. 81-89). Springer, Boston, MA.

Noyes, J. M., & Bruneau, D. P. (2007). A self-analysis of the NASA-TLX workload measure. Ergonomics, 50(4), 514-519.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Human Factors, 33(1), 17-33.

O’Donnel, R. D., & Eggemeier, F. T. (1986). Cognitive processes and performance. Handbook of Perception and Human Performance; Kaufman, BK, Wiley, TJ, Eds, 41-42.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010, March). Estimating cognitive load using remote eye tracking in a driving simulator. In Proceedings of the 2010 symposium on eye-tracking research & applications (pp. 141-144). ACM.

Palmer, J. W. (2002). Web site usability, design, and performance metrics. Information systems research, 13(2), 151-167.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004, March). The determinants of web page viewing behavior: an eye-tracking study. In Proceedings of the 2004 symposium on Eye tracking research & applications (pp. 147-154).

Papillo, J.F. & Shapiro, D., (1990). The Cardiovascular System. In: L.G. Tassinary, Eds. *Principles of Psychophysiology: Physical, Social, and Inferential Elements* (pp. 456-512). Cambridge: Cambridge University Press.

Parameswaran, M., & Whinston, A. B. (2007). Social computing: An overview. *Communications of the Association for Information Systems*, 19(1), 37.

Parasuraman, R. (1990). Event-related brain potentials and human factors research. Oxford University Press.

Pashler, H. E., & Sutherland, S. (1998). The psychology of attention (Vol. 15). Cambridge, MA: MIT press.

Paternò, F., Schiavone, A. G., & Conti, A. (2017, September). Customizable automatic detection of usability bad smells in mobile-access web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-11)

Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L., & Morrisroe, G. (2005). The Integrated Workload Scale (IWS): a new self-report tool to assess railway signaller workload. Applied Ergonomics, 36(6), 681-693.

Polillo, R. (2010). Facile da usare-Una moderna introduzione all'ingegneria della usabilità (pp. 1-413). Apogeo.

Poole, A., Ball, L. J., & Phillips, P. (2005). In search of salience: A response-time and eye-movement analysis of bookmark recognition. In People and computers XVIII—Design for life (pp. 363-378). Springer, London.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of human computer interaction*, 1, 211-219.

Preece, J., Rogers, Y., & Sharp, H. (2015). Interaction design: beyond human-computer interaction. John Wiley & Sons.

Ragot, R. (1984). Perceptual and motor space representation: An event□related potential study. Psychophysiology, 21(2), 159-170.

Rayner, K. (1998). "Eye movements in reading and information processing: 20 years of research". Psychological Bulletin 134(3):372–422.

Rayner, K. & Pollatsek, A. (1989) The psychology of reading. Prentice Hall.

Recarte, M.A., & Nunes, L.M. (2000). Effects of verbal and spatial imagery tasks on eye fixations while driving. Journal of Experimental Psychology: Applied, 6, 31-43.

Recarte, M. A., & Nunes, L. (2002). Mental load and loss of control over speed in real driving.: Towards a theory of attentional speed control. Transportation Research Part F: Traffic Psychology and Behaviour, 5(2), 111-122.

Recarte, M.A., & Nunes, L.M. (2003). Mental workload and driving: Effects on visual search, discrimination and decision making. Journal of Experimental Psychology: Applied, 9, 119-137.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Advances in psychology (Vol. 52, pp. 185-218). North-Holland.

Reichheld, F. F. (2003). The one number you need to grow. Harvard business review, 81(12), 46-55.

Reichheld, F. F. (2006). Questions about NPS–and some answers. accessed March, 1, 2007.

Rogers, Y. (2004). New theoretical approaches for human-computer interaction. Annual review of information science and technology, 38(1), 87-143.

Rogers, W. A., Fisk, A. D., McLaughlin, A. C., & Pak, R. (2005). Touch a screen or turn a knob: Choosing the best device for the job. *Human factors*, *47*(2), 271-288

Roscoe, A.H. (1984). Assessing pilot workload in flight. Flight test techniques. In Proceedings of NATO Advisory Group for Aerospace Research and Development (AGARD).

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA□TLX, and workload profile methods. *Applied Psychology,* 53(1), 61-86.

Russell, J.A., 1995. Facial expressions of emotion: what lies beyond minimal universality? Psychol. Bull. 118 (3), 379–391.

Ruzza, M., Tiozzo, B., Mantovani, C., D'Este, F., & Ravarotto, L. (2017). Designing the information architecture of a complex website: A strategy based on news content and faceted classification. *International Journal of Information Management*, 37(3), 166-176.

Sauro, J. (2011a). Measuring usability with the system usability scale (SUS). Retrieved from https://measuringu.com/sus/ (Accessed 20 November 2020).

Sauro, J. (2011b). What is a good task-completion rate? Retrieved from https://measuringu.com/task-completion/ (Accessed 20 November 2020).

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. Journal of usability studies, 10(2).

Sauro, J. (2010). Does Better Usability Increase Customer Loyalty? The Net Promoter Score and the System Usability Scale (SUS). Retrieved from http://www.measuringu.com/usability-loyalty.php (Accessed on 20 November 2020).

Sauro, J., & Lewis, J. R. (2012). Quantifying the user experience: Practical statistics for user research. Burlington, MA: Morgan Kaufmann.

Sawin, D.A., & Scerbo, M.W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. Human Factors, 37, 752 –765.

Schneider, M. L. (1982). Models for the design of static, software systems. Gathering Information for Problem Formulation, 107.

Scholtz, J. (2004). *Usability evaluation*. National Institute of Standards and Technology, 1.

Sharp, H., Rogers, Y., & Preece, J. (2007). Interaction design. Beyond human–computer interaction (2nd ed.). Chichester, UK: John Wiley & Sons.

Shively, R., Battiste, V., Matsumoto, J., Pepiton, D., Bortolussi, M., & Hart, S.G. (1987). In flight evaluation of pilot workload measures for rotorcraft research. In Proceedings of the Fourth Symposium on Aviation Psychology (pp. 637– 643). Columbus, OH: Department of Aviation, Ohio State University.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the user interface: strategies for effective human-computer interaction*. Pearson.

Slabey, R. (1990), "QFD: a basic primer. Excerpts from the implementation manual for the three day QFD workshop", Transactions from the Second Symposium on Quality Function Deployment, Novi, MI, 18-19 June

Smith, D. C., Irby, C., Kimball, R., & Harslem, E. (1982, June). The Star user interface: An overview. In Proceedings of the June 7-10, 1982, national computer conference (pp. 515-528).

Staiano, J., Menéndez, M., Battocchi, A., De Angeli, A., & Sebe, N. (2012, June). UX_Mate: from facial expressions to UX evaluation. In Proceedings of the Designing Interactive Systems Conference (pp. 741-750).

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. *The Oxford handbook of psycholinguistics*, 327, 342.

Stein, E.S. (1985). Air traffic controller workload: An examination of workload probe. (Report No. DOT/FAA/CT-TN84/24). Atlantic City, NJ: Federal Aviation Administration Technical Centre.

Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink. Psychophysiology, 21(1), 22–33.

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of fatigue. Human Factors, 36, 285–297.

Takahashi, Y., Rayman, J. B., & Dynlacht, B. D. (2000). Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression. *Genes & development*, 14(7), 804-816.

Tatum IV, W. O. (Ed.). (2014). Handbook of EEG interpretation. Demos Medical Publishing.

Terzis, V., Moridis, C. N., & Economides, A. A. (2010, August). Measuring instant emotions during a self-assessment test: the use of FaceReader. In Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research (pp. 1-4).

Tole, J. R., Stephens, A. T., Harris, R. L., & Ephrath, A. R. (1982). Visual scanning behavior and mental workload in aircraft pilots. Aviation, Space, and Environmental Medicine.

Tole, J. R., Stephens, A. T., Vivaudou, M., Ephrath, A. R., & Young, L. R. (1983). Visual scanning behavior and pilot workload (NASA Contractor Report No. 3717). Hampton, VA: NASA LangleyResearch Center.

Took, R. (1990). *Putting design into practice: Formal specification and the user interface*. In Formal methods in human-computer interaction (pp. 63-96). Cambridge University Press.

Trimmel, M., Meixner-Pendleton, M., & Haring, S. (2003). Stress response caused by system response time when searching for information on the Internet. Human Factors, 45(4), 615-622.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. Ergonomics, 39(3), 358-381.

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. Handbook of human factors and ergonomics, 3, 243-268.

Turner, C. W., Lewis, J. R., & Nielsen, J. (2006). Determining usability test sample size. *International encyclopedia of ergonomics and human factors*, 3(2), 3084-3088.

Turoff, M. (1997). The design and evaluation of interactive systems. Section 1.4.3: "Userroles and types". http://web.njit.edu/~turoff/coursenotes/IS732/book/tablecon.htm. (Accessed 22 October 2020).

Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. Mind & Language, 25(1), 3-29.

Van Camp, M., De Boeck, M., Verwulgen, S., & De Bruyne, G. (2018, July). EEG technology for UX evaluation: A multisensory perspective. In International Conference on Applied Human Factors and Ergonomics (pp. 337-343). Springer, Cham.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behaviour & information technology, 22(5), 339-351.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. Human factors, 43(1), 111-121.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. Biological psychology, 42(3), 323-342.

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. Ergonomics, 41(5), 656–669.

Vidulich, M. A., & Tsang, P. S. (1986). Techniques of subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods. Ergonomics, 29(11), 1385-1398.

Vidullch, M. A., Ward, G. F., & Schueren, J. (1991). Using the subjective workload dominance (SWORD) technique for projective workload assessment. Human Factors, 33(6), 677-691.

Virzi, R. A. (1992). *Refining the test phase of usability evaluation: How many subjects is enough?*. Human factors, 34(4), 457-468.

Wainwright, W.A. (1987). Flight test evaluation of crew workload. In A.H. Roscoe (Ed.), The practical assessment of pilot workload, AGARDograph No. 282 (pp.60-68). Neuilly sur Seine, France: AGARD.

Wang, J., Antonenko, P., Celepkolu, M., Jimenez, Y., Fieldman, E., & Fieldman, A. (2018). Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data. International Journal of Human–Computer Interaction, 1-12.

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different WEB page designs. International Journal of Human-Computer Studies, 59(1-2), 199-212.

Ward, R. D., Marsden, P. H., Cahill, B., & Johnson, C. (2002, April). Physiological responses to well-designed and poorly designed interfaces. In Proceedings of CHI 2002 workshop on physiological computing.

Wattal, S., Schuff, D., Mandviwalla, M., & Williams, C. B. (2010). Web 2.0 and politics: the 2008 US presidential election and an e-politics research agenda. MIS quarterly, 669-688.

Wharton, C. (1994). *The cognitive walkthrough method: A practitioner's guide*. Usability inspection methods.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), Varieties of attention (pp. 63-102). San Diego, CA: Academic Press.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. Human factors, 37(3), 473-494.

Wickens, C.D., & Hollands, J.G. (2000). *Engineering Psychology and Human Performance (3rd Edition)*. Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D. (2002). Multiple resources and performance prediction. Theoretical issues in ergonomics science, 3(2), 159-177.

Wickens, C. D. (2008). Multiple resources and mental workload. Human factors, 50(3), 449-455.

Wiebe, E. N., Roberts, E., & Behrend, T. S. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. Computers in Human Behavior, 26(3), 474-481.

Wierwille, W. W., & Casali, J. G. (1983, October). A validated rating scale for global mental workload measurement applications. In Proceedings of the human factors society annual meeting (Vol. 27, No. 2, pp. 129-133). Sage CA: Los Angeles.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors,* 35(2), 263-281.

Wilson, G. F., & Eggemeier, F. T. (2006). Mental workload measurement. International encyclopedia of ergonomics and human factors, 1.

Wilson, G. F., Grandt, M., Svensson, E., Schlegel, R. E., Veltman, H., Van Orden, K. F., ... & Fraser, W. (2004). Operator functional state assessment. [Hampton: NASA langley research center].

Wilson, G. M. (2001, March). Psychophysiological indicators of the impact of media quality on users. In CHI'01 Extended Abstracts on Human Factors in Computing Systems (pp. 95-96).

Wilson, G. M., & Sasse, M. A. (2000). Investigating the impact of audio degradations on users: subjective vs objective assessment methods. CHISIG: the Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia.

Winton, W. M., Putnam, L. E., & Krauss, R. M. (1984). Facial and autonomic manifestations of the dimensional structure of emotion. Journal of Experimental Social Psychology, 20(3), 195-216.

Witkowski, T. (2012). A review of research findings on neuro-linguistic programming. Scientific Review of Mental Health Practice, 9(1).

Wolverton, G.S., & Zola, D. (1983). The temporal characteristics of visual information extraction during reading. In K. Rayner (Ed.), Eye movements in reading: Perceptual and language processes. New York: Academic.

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress*, 14(1), 74-99.

Yannakakis, G. N., Hallam, J., & Lund, H. H. (2008). Entertainment capture through heart rate activity in physical interactive playgrounds. User Modeling and User-Adapted Interaction, 18(1-2), 207-243.

Yarbus, A. (1967). Eye movements and vision, New York, Plenum Press.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, 58(1), 1-17.

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. Human factors, 44(3), 365-375.

Young, M. S., & Stanton, N. A. (2004). Taking the load off: investigations of how adaptive cruise control affects mental workload. Ergonomics, 47(9), 1014-1035.

Zaman, B., & Shrimpton-Smith, T. (2006, October). The FaceReader: Measuring instant fun of use. In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles (pp. 457-460).