2022

An Investigation of the Relationship Between Subjective Mental Workload and Objective Indicators of User Activity

Greg Byrne
Technological University Dublin, Ireland

An Investigation of the Relationship Between Subjective Mental Workload and Objective Indicators of User Activity

Greg Byrne

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computer Science (Advanced Software Development)

16.06.2022

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Stream), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Greg Byrne

Date: 16.06.2022

Abstract

Whilst the concept of physical workload is intuitively understood and readily applicable in system design, the same cannot be said of mental workload (MWL), despite its importance in our increasingly technological society. Despite its origin in the mid 20^{th} century, the very concept of "mental workload" is still a topic of debate in the literature, although it can be loosely defined as "the amount of mental work necessary for a person to complete a task" (Miller, 1956; Longo, 2014). Several methods have been utilized to measure of MWL, including physiological methods such as neuro-imagery, performance-based metrics, and subjective measures *via* questionnaires, such as the NASA-TLX method (NASA, 2022).

In this work, the relationship between subjective measures of MWL and objective indicators of activity is examined. Herein, a series of web-based tasks have been developed with mouse-activity monitoring implemented in JavaScript in order to study this relationship. The experimental results indicate that user mouse activity does not correlate with subjective indicators of MWL.

Keywords: Mental Workload, Human-Computer Interfaces, Ergonomics, Usability, Web Design

Acknowledgments

I would like to express my thanks to Dr. Bujar Raufi for his support throughout this project, without which this work would not have been possible. Similarly, I would like to thank Prof. Luca Longo for his guidance in this project.

Furthermore I would like to thank my family for their encouragement and support during both this project, and throughout my postgraduate studies as a whole.

Contents

List of Figures

List of Tables

List of Acronyms

HCI	Human-computer Interaction
MWL	Mental Workload
NASA-TLX	National Aeronautics and Space Administration Task Load Index
WP	Workload Profile
SWAT	Subjective Workload Assessment Technique
AI	Artificial Intelligence
ML	Machine Learning
MRT	Multiple Resource Theory

Chapter 1

Introduction

1.1 Background

As the world grows increasingly digital, accelerated by the Covid-19 pandemic, the need for well-designed and accessible computer systems is at an all-time high (Vargo, Zhu, Benwell, & Yan, 2021; Lottridge, 2020; Kucirkova, Evertsen-Stanghelle, Studsrød, Jensen, & Størksen, 2020). Past research has generated a range of procedures for assessing usability of computer systems (Kaur, Kaur, & Kaur, 2016; Freire, Arezes, & Campos, 2012; Thomas, Onyimbo, & Logeswaran, 2016). This research, however, tends to omit an key metric relating to the context of the user and the perceived difficulty of the task being carried out - this phenomena is known as the mental workload (MWL), or alternatively cognitive workload. Although a precise definition of this complex psychological phenomenon has not yet been developed, there exist a range of subjective self-reporting tools for its assessment (Hart, 2006; F. T. Eggemeier, Shingledecker, & Crabtree, 1985; Cain, 2007; Tsang & Velazquez, 1996; C. D. Wickens, Helton, Hollands, & Banbury, 1992).

In recent years an increasingly sophisticated range of technologies have been applied to investigate the usability of web based systems (Souza, Seruffo, De Mello, Souza, & Vellasco, 2019; Aviz, Souza, Ribeiro, De Mello, & Seruffo, 2019), but relatively few attempts have been made to explore MWL with such techniques. Longo *et al.* have carried out promising research in this domain, finding that - although not correlating

to usability - MWL and could be "employed to improve the prediction of human performance, thus enhancing the description of user experience". In one study, Grimes and Valacich found that a user's mouse activity could be used as a proxy for MWL, with interesting implications for e-learning and user experience (Grimes & Valacich, 2015).

1.2 Research Project/Problem

This work is concerned with the subject of MWL, its subjective assessment metrics and their potential relationship with tangible indicators of user activity, such as mouse activity. Due to the increasing importance of ergonomics in the field of HCI, much research has been previously carried out into predictors of MWL and user activity (Zöllner et al., 2011; Mock et al., 2016; Atterer, Wnuk, & Schmidt, 2006a; Arroyo, Selker, & Wei, 2006; Xie & Salvendy, 2000a).

Methods for measuring MWL are divided into three broad categories, subjective methods (questionnaires), physiological methods (*e.g.* Brain neuro-imagery), and performance-based metrics (Cain, 2007). Whilst a wide variety of techniques have been utilized to investigate the relationship between user activity (mouse clicking, movement, focus *etc*) and user experience; much remains to be investigated regarding a user's subjective assessment of MWL and its relationship to objective indicators of activity.

With this in mind, this study aims to investigate the relationship between tangible indicators of user activity and subjective MWL scores. For this purpose, a well established multi-dimensional measure of MWL will be employed: NASA's Task Load Index (NASA-TLX). The aim of this investigation is to develop a suite of tools (using JavaScript) to gather metrics pertaining to user activity (mouse movement, clicks, focus, scrolling), and investigate their relationship with a user's subjective measure of MWL. With this in mind, this study aims to explore the following research question:

"Is there a relationship between a user's subjective MWL when performing web-based tasks and objective indicators of tangible activity in the web browser?"

In order to investigate this research question, the following hypotheses are introduced:

Null Hypothesis (H0) Metrics of user activity, obtained using JavaScript embedded in web tasks, exhibit no correlation with indicators of MWL.

Alternative Hypothesis (H1) Metrics of user activity, obtained using JavaScript embedded in web tasks, correlate highly with indicators of MWL.

1.3 Research Objectives

The objective of this research is to answer the research question posed above; that is, to determine whether a correlation exists between objective indicators of user activity when performing web-based tasks, and subjective measures of MWL. This project considers several mouse interaction metrics: mouse position, mouse clicks, mouse hovering, mouse velocity.

For this purpose, an application will be developed using JavaScript, where users can perform tasks, during which mouse metrics are tracked including position, clicks, element hover-time *etc* (Mozilla, 2022c). This data will be gathered, parsed and stored transparently to the user. By storing mouse position and time, composite metrics can be calculated such as mouse velocity and mouse distance travelled. mouse-clicks, *etc.*. After completing each task, users will be asked to assess MWL according to the NASA Task Load Index (NASA-TLX) and Workload Profile (WP).

1.4 Research Methodologies

Initially, this project consisted of a period of secondary research to gain an understanding of the literature and state-of-the-art in the areas of HCI, ergonomics, MWL and their applications in web-design. This provided a grounding in the methodologies and heuristics involved in this area, in particular techniques and considerations for effective MWL assessment.

For the primary research, a web application was developed, incorporating two well

established MWL evaluation questionnaires, with fully automated data capture and storage for user behavioural indicators. This was deployed to a publicly available domain, and distributed to participants. The correlation between user behavioural metrics and self-reported MWL assessments was analysed using statistical methods, to investigate the extent of correlation between the two quantities. Based on the outcome of this analysis, the null or alternate hypothesis will be accepted.

1.5 Scope and Limitations

The scope of this project relates to the aforementioned research objectives. This project is concerned with indicators of user activity, MWL, and the nature of any relationship that exists between these quantities. Much work exists in the literature involving the relationship between physiological indicators and MWL, but this work focuses specifically on user mouse activity in the context of web-based tasks (Delliaux, Delaforge, Deharo, & Chaumet, 2019; Backs & Seljos, 1994; Causse et al., 2022; Gevins & Smith, 2003).

This experiment faces limitations in terms of the broadness and variety of web-based tasks available (this work features two custom-built tasks), past research demonstrates that there are numerous use cases where behavioural indicators can be tracked for web-based tasks (Romero, 2017; McFarland, 2016). Furthermore, as an online experiment, the environment in which the tasks take place is less controlled than a lab-based alternative. One cannot ensure, for instance, that a user is paying attention during the relatively short time allocated to each task. Similarly, users will perform the experiment on different devices, with different screen sizes, resolutions *etc.*

Participants for the experiment were selected from university groups, friends, colleagues *etc.* which naturally limits the group in terms of demographics and geographical region. This was unavoidable in the time0frame of this research, but ideally a participant pool should be drawn from as diverse and representative a group as possible.

1.6 Document Outline

Literature Review This chapter documents the the secondary research component of this project. A broad body of literature was reviewed, with an emphasis on the areas of HCI, ergonomics, usability, design and MWL. The chapter first introduces the concept of MWL, before diving into the rich body of research that exists in the area covering an array of measurement techniques and applications. Gaps in the research are explored, introducing the goals of the current work.

Experiment Design This chapter outlines the experimental design used in this work, in the context of the experiment's objectives. This include the reasoning behind the choice MWL techniques utilized, the software design and implementation, with a focus on the literature and current best-practices in the field. Finally, the data collected in the experiment is discussed, as well as the calculation of the MWL results utilized.

Results and Evaluation This chapter details the results obtained from the experiment, and compares them with similar work in the literature. The results and analysed and used to make a decision regarding the null or alternative hypotheses, thereby satisfying the research objective.

Conclusion Finally, the results of this work are summarized and discussed in the broader context of this field. The successes and challenges of the experiment are reflected upon, and used a the basis to suggest further areas of research.

Chapter 2

Review of existing literature

This chapter discusses the relevant literature concerning the field of Mental Workload (MWL), and its relation to human-computer interaction (HCI) and usability. This chapter is divided into three sub-chapters. The first sub-chapter gives a brief overview of HCI and usability to provide context and introduce the reader to key concepts relating to human interaction with computer systems.

Next, the phenomena of MWL is introduced and rich psychological history of the field is explored. Context is provided in relation to the ongoing debate regarding the definition of the concept of MWL. The guiding principles and theories of MWL are discussed, and an overview of the methodologies commonly employed in this field is provided. This provides a grounding in the fundamental concepts and challenges relating to this field, illustrating the need for further research.

In the final sub-chapter, "Related Work", contemporary research is explored to gain insight into state of the art and gaps in the research. The aim of this section is to provide insight into the need for further research in this area as it pertains to this study. Given that MWL lies at the interface of fields including HCI, usability and psychology - this section outlines relevant research which closely relates to this study, rather than providing an exhaustive review of the various and constantly developing array of research in this area.

2.1 Human-Computer Interaction

Human-computer interaction (HCI) is the study of usability in relation to the usage of computers by humans. The aim of studying HCI is to increase the ease of use and value derived from a persons' interaction with a computer system (Bansal & Khan, 2018). The study of HCI is thought to have begun in 1959 with Shakel's paper "The ergonomics of a computer", followed by Licklider with his seminal research on "Man - Computer Symbiosis" which, in hindsight, is prophetic when considering todays "era of enhanced digital connectivity" (Shackel, 1969; Licklider, 1960; Pantic, Nijholt, Pentland, & Huanag, 2008). HCI encompasses both mechanical interaction - such as the ergonomic design of mice and keyboards, as well as interaction through software. When a person interacts with a computer their experience is affected by a vast range of criteria - the software interface, the colour choices, the input device, the keyboard layout, the interaction response times *etc.* These criteria all play a role in the relationship between a human and computer - and have been studied extensively in the literature (Shneiderman, 1988; Cowley et al., 2016; Javaid, 2013).

Karray *et al.* discuss how the value of a computer system "is visible only when it becomes possible to be efficiently utilised by the user" and describe a computer systems' efficacy as a "balance between functionality and usability of a system" (Karray, Alemzadeh, Abou Saleh, & Nours Arab, 2008). As more sophisticated computer systems are developed, HCI has become a growing concern - with much recent work investigating the usability and explainability of machine-learning and artificial intelligence (ML/AI) systems (Grudin, 2009; Winograd, 2006; Li, Kumar, Lasecki, & Hilliges, 2020; Bhatt et al., 2020). Many studies have been carried out in recent years to increase the transparency, interpretability and explainability of such systems to better facilitate the derivation of knowledge from systems that have been formerly categorized as "black boxes" (Roscher, Bohn, Duarte, & Garcke, 2020; Pawar, O'Shea, Rea, & O'Reilly, 2020).

2.2 Human Mental Workload

2.2.1 Usability and MWL

The increasing prevalence of the internet and the necessity for digital literacy necessitates that systems can be evaluated and designed to ensure an adequate level of usability is achieved. Many heuristics have thus been developed for assessing usability of digital systems. Standard ISO 9241-11 defines usability as the "extent to which a product can be used by specified users to achieved specified goals with effectiveness, efficiency and satisfaction" (ISO-9241-11, 2018). These terms are defined as follows:

- **User**: Person who interacts with the product
- **Goal**: Intended outcome
- **Effectiveness**: Accuracy and completeness with which users achieve specified goals
- **Efficiency**: Resources expended in relation to the accuracy and completeness with which users achieve goals
- **Satisfaction**: Freedom from discomfort and positive attitudes
- **Context of use**: Accuracy and completeness with which users achieve specified goals

Tan *et al.* recognized the two most important usability evaluation techniques - heuristic analysis and user testing (Tan, Liu, & Bishu, 2009). This is an evolving field, however, and as the use cases for web-applications proliferate due to increased digitization - new heuristics are being developed (Quiñones & Rusu, 2017). Nielsen's 10 heuristics are widely recognized in the field of HCI and usability (Nielsen, 1994):

Table 2.1: Nielsen's 10 Usability Heuristics (Nielsen, 1994)

H1	Visibility of system status
H2	Match between system and the real world
H3	User control and freedom
H4	Consistency and standards
H5	Error prevention
H6	Recognition rather than recall
H7	Flexibility and efficiency of use
H8	Aesthetic and minimalist design
H9	Help users recognize, diagnose, and recover from errors
H10	Help and documentation

These heuristics have been applied extensively to enhance the usability of computer products across a wide spectrum of use cases (Li et al., 2020; Paz, Paz, Pow-Sang, & Collantes, 2014; Alsumait & Al-Osaimi, 2009).

The concept of physical workload has been effectively used for decades in the fields of ergonomics and physiology. Many methods have been developed to quantify physical workload - relating the amount of physical work done with the energy cost (*e.g.* oxygen consumption associated with a particular exercise) (Gawron, 2019). However, as society becomes increasingly technological - an increasing amount of work is done requiring little physical exertion. Instead, tasks involve mentally strenuous tasks such as information processing, multitasking, decision making and monitoring *etc.* This leaves a vacuum in the literature where a large amount of work is being done without adequate metrics to evaluate the workload associated with these tasks. This has led to much work being carried out, originally in the psychological literature with for applications in aviation - but more recently with a focus on HCI and usability (Reid & Nygren, 1988; Hart & Staveland, 1988; Gopher & Donchin, 1986; Young, Brookhuis, Wickens, & Hancock, 2015).

As observed by Longo in 2015 "there has been a tendency to overlook aspects of the context and characteristics of the users during the usability assessment process".

One such characteristic is a user's mental exertion when performing a given task. Human mental workload (MWL) can be intuitively defined as "the amount of mental work necessary for a person to complete a task over a given period of time" (Longo, 2014). Although the concept has been in circulation for the last four decades (Hart & Staveland, 1988), no precise definition yet exists for the phenomenon. The concept of MWL has seen application in a variety of fields including aviation, transport, HCI, educational psychology *etc* (Hart & Staveland, 1988; Longo & Dondio, 2016; Biondi, Cacanindin, Douglas, & Cort, 2021; Kim & Ji, 2013; Xie & Salvendy, 2000b; Paxion, Galy, & Berthelon, 2014; Mehler, Reimer, Coughlin, & Dusek, 2009). The multidisciplinary nature of MWL contributes to the difficulty in deriving a single, concrete definition.

2.2.2 Defining Mental Workload

Whilst MWL can be intuitively defined as the the "amount of mental work necessary for a person to complete a task"(Longo, 2014), there still exists much debate surrounding the precise definition of the term. Hart and Staveland suggested that MWL "is not an inherent property, but rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills,behaviours and perceptions of the operator" (Hart & Staveland, 1988). MWL has been defined in the psychological literature by Huey *et al.* as a mental construct, or "intervening variable (not reducible to empirical terms and not directly observable) rather than a hypothetical construct" (Council, 1993; Gopher & Donchin, 1986; MacCorquodale & Meehl, 1948; Longo, 2014). Here, MWL is described in terms of the cognitive demand a task places on a human operator (Cain, 2007). Eggemeier and Wilson defined it as "the portion of operator information processing capacity... required to meet system demands" (F. Eggemeier, Wilson, Kramer, & Damos, 1993). Arguing that "performance is not all that matters in the design of a good system", Wickens argues that a consideration of the mental strain a task places on its operator leads to better designed systems (C. D. Wickens et al., 1992). These definitions are all underpinned by a common theme; the phenomena of the mental demand a task

places on a user.

Xie *et al.* describe how, due to the subjective, psychological nature of the metric, "nobody seems to know what MWL is", and instead make do with an "intuitively 'right'" definition. Similarly to Gopher et al, they postulate that "Mental workload cannot be detected directly, but through the measurement of some other variables that are thought to correlate highly with it, such as subjective rating, performance and some physiological data"(Xie & Salvendy, 2000b). The multifaceted, "emergent" nature of cognitive workload has resulted in Gevins describing its measurement as "perhaps the most basic issue" in the field (Gevins & Smith, 2003).

Cognitive workload is a multifaceted domain, composed of three broad areas – the amount of work and number of things to do, time and the subjective psychological experiences of the human operator (Lysaght et al., 1989). Gopher and Dochin note that no absolute measure can indicate MWL, (Gopher & Donchin, 1986); Curry, *et al.* instead applied a relative measure to describe MWL, which captures the ephemeral, "non empirical" aspect of this quantity (Curry, Jex, Levison, & Stassen, 1979). They defined MWL as follows: "the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort ... workload is never greater than unity".

2.2.3 Supporting Theories

The concept of MWL is underpinned by two core concepts; *limited processing capacity* and *performance* (Longo, 2014). In 1975 Kahneman envisioned that, due to limited cognitive processing capacity, mental resources could be viewed as a "single undifferentiated capacity", or a limited pool of resources available for humans to perform tasks (Egeth & Kahneman, 1975). Due to limited nature of mental resources - concurrent tasks competed for mental resources, and performance suffered as a result.

Wickens' *multiple resource theory* (MRT) was proposed some years later, building on this foundation. Wickens acknowledges Kahnemans' influential work on attention, and proposes a multiple resource model, illustrated with the 4D cube model, with each dimension representing one of the "dichotomies of information processing"

(C. D. Wickens et al., 1992). Here, each dimension corresponds to a different pool of resources. The four dimensions are as follows:

- **Stages of Processing**: Perceptual, central, response. This denotes that cognitive activities (*e.g.* working memory) utilizes different resources than the selection of actions (response). This is supported by the example of an air-traffic controller instructing a pilot of updates to traffic (response) whilst simultaneously maintaining an overview of the current state of air-traffic (cognitive).
- **Codes of Processing**: Verbal, spatial. Consider a driver struggling to focus on driving whilst dialing a phone number (competing for visual resources), versus using a voice command to dial a number. This illustrates a distinction between verbal and spatial resources.
- **Modalities Dimension**: This indicates that auditory perception and visual perception employ different pools of resources.
- **Visual Channels**: Object recognition, peripheral vision, *etc.* This dimension distinguishes between focal and ambient vision. Longo uses the example of walking through a hallway (guided by peripheral vision) whilst reading a book (focal vision) (Longo, 2014).

Wickens' 4D model is based both on physiological and human-centric rationale - each dimension corresponds to a region of the brain, as indicated by FF-MRI studies. Furthermore, each dimension should "correlate with relatively straightforward decisions that a design could make... to support multitask activities" (C. Wickens, 2008).

In 1988, Reid and Nygren developed an assessment technique for mental workload based on a three dimensional model comprised of time load, mental-effort load, and psychological-stress load (Reid & Nygren, 1988). Each criterion can be sub-divided into three categories of low to high load - with 27 possible permutations of load levels.

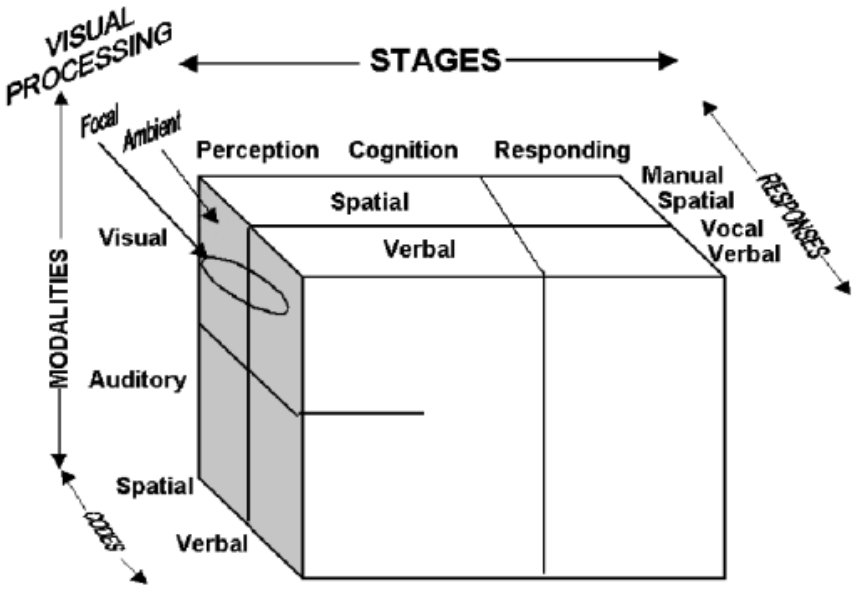

Figure 2.1: Wickens 4-D multiple resource model (C. Wickens, 2008)

- **Time Load**

  1. Often have spare time. Interruptions or overlapping activities occur infrequently.
  2. Occasionally have spare time. Interruptions occur frequently.
  3. Minimal spare time. Interruptions occur almost constantly.

- **Mental Effort Load**

  1. Very little conscious mental effort required. Activity is almost automatic.
  2. Moderate concentration is required.
  3. Extensive mental effort and concentration is necessary. Very complex activity.

- **Psychological Stress Load**

  1. Little confusion, risk, frustration, anxiety or stress present.
  2. Moderate stress for the aforementioned reasons Significant compensation is required to maintain performance.

3. Intense stress due to aforementioned reasons. Extreme determination and self-control required.

This is illustrated as follows:

Figure 2.2: 3-D workload construct (Reid & Nygren, 1988)

2.2.4 Mental Workload Methodologies

Classes of Measurement

The multitude of definitions and theories about the precise definition of the phenomena is reflected in its numerous measurement techniques (F. Eggemeier et al., 1993; Hart & Staveland, 1988; Muñoz-de Escalona, Cañas, Leva, & Longo, 2020; Hendy, Hamilton, & Landry, 1993; Reid & Nygren, 1988). As early as 1978, Wierwille and Williges identified 28 specific techniques to measure workload (Williges & Wierwille, 1979). In general, measurement techniques are organised into three broad categories (F. Eggemeier et al., 1993; Young & Stanton, 2002; Tsang & Velazquez, 1996; Cain, 2007; Longo, 2018):

- **Subjective Measures**: These are psychological measures which includes self-reported measures and subjective rating scales.
- **Performance-based Measures**: These measures assume that the cognitive workload of an operator is acquires importance only if it has an influence on task performance. This technique is therefore considered most valuable in system design (Longo & Dondio, 2016).

- **Physiological Measures**: Those derived from the physiology of the human subject (Mehler et al., 2009)

Subjective Measures

These measurement techniques require operators to judge and report their own experience of the workload associated with a particular task. The most widely accepted subjective measures are as outlined as follows.

- **NASA Task Load Index (NASA-TLX)**: NASA-TLX is a multi-dimensional scale consisting of six sub-scales that represent different aspects of workload: Mental, Physical and Temporal Demands, Frustration, Effort and Performance. This scale was developed by Hart and Staveland in 1988 for application in aeronautics (Hart & Staveland, 1988). When carrying out NASA-TLX, these attributes are typically gathered after the execution of a task, and the weighted average is computed considering to yield a metric indicative of MWL.

- **Subjective Workload Assessment Technique (SWAT)**: Here, the dimensions of time load, mental effort and psychological stress are considered. Each variable is described by three discrete values and users are required to sort the 27 different combinations by MWL from lowest to the highest. This is the most common method employed according to the literature (Cain, 2007; Reid & Nygren, 1988).

- **Workload Profile (WP)**: This self-assessment was developed by Tsang and Valazques (Tsang & Velazquez, 1996) and is based on the multiple-resource theory of Wickens (C. D. Wickens et al., 1992; C. Wickens, 2008). WP considers the dimensions of solving and deciding, time and space, verbal, auditory, speech and response selection; Visual attention; Manual activity. WP is often measured using the dual-task technique, where subjects carry out two tasks simultaneously. If subjects have a constant size pool of cognitive resources to draw upon, the amount of effort employed in the primary task will be inversely proportional to the performance on the secondary task (Dennis, Bruza, & McArthur, 2002).

Despite the difficulties of obtaining reliable data from subjective self-reported questionnaires, the NASA-TLX, SWAT, and WP have been demonstrated as reliable means of measuring MWL, and have been evaluated in a multitude of comparative studies (Hart, 2006; Kim & Ji, 2013; Rubio, Díaz, Martín, & Puente, 2004; Byers, Bittner, & Hill, 1989).

Performance-Based Measures

Performance-based measures derive an index of workload from some aspect of operator behaviour or activity. These can be subdivided into two groups: *primary task measures* and *secondary task measures*. Primary task measures specify the adequacy of operator performance on the principal task (e.g. driving errors made while a person drives a car). Secondary task measures provide a proxy to gauge workload of a primary task - inferring the workload *via* an operators capacity for handling additional tasks (e.g. carry out a task while talking on the phone). Tsang and Vidulich have criticised performance based methodologies for overlooking difficulties with the task itself - and attributing performance only to the operator's traits. This can be overly simplistic in the case of a user interacting with a poorly designed interface, for example (Tsang & Vidulich, 2006).

Physiological Measures

It is well-established that physiological indicators of arousal are sensitive to mental events such as emotion and stress. Physiological measures aim to infer the level of mental workload from some metric obtained from the operator (e.g. pulse, pupillary reflex *etc*). Physiological measures can be broadly categorized into three groups:

- **Brain-related Measures**: Many studies have demonstrated a correlation between neuroimaging data and MWL. fMRI and EEG techniques are particularly favoured for their portability and unobtrusiveness (Gevins & Smith, 2003; Ayaz et al., 2012; Causse, Chua, Peysakhovich, Del Campo, & Matton, 2017).

- **Eye-related Measures**: Pulat estimates that 80% of information processed by

the brain is through visual channels (Pulat, 1997). Many studies investigate the relationship between eye and gaze activity with MWL (Schall, 2014; Serra et al., 2019; Aviz et al., 2019). In one such study, Holland demonstrates a relationship between blink interval and MWL in auditory input tasks (Holland & Tarlow, 1972).

- **Heart-related Measures**: Results from heart-rate, blood pressure and electrocardiogram (ECG) analysis have been utilized to to investigate MWL (Vincent, Craik, & Furedy, 1996; Backs & Seljos, 1994). The relationship between MWL and cardiovascular activity has important implications in society, as it high levels of MWL are related to increased cardiovascular risk (Delliaux et al., 2019).

According to Ward and Marsden, "psycho-physiological testing is perhaps not as robust as HCI usability testers might like it to be", as the invasive nature of the measurement may have a non-trivial effect on the user performing the task (Ward & Marsden, 2003). According to Longo, however, the problem is improved significantly with the development of more sophistical and miniaturized tools and sensors currently available to researchers (Longo, 2014). The precise metrics acquired vary widely in this domain, from interaction data obtained from traditional peripherals to more sophisticated technologies such as touch screens, eye-tracking and brain scanning technologies (Maslov & Nikou, 2020; Ayaz et al., 2012; Pimenta, Carneiro, Novais, & Neves, 2013; Mock et al., 2016).

Due to the range of assessment categories, certain measures may be more appropriate than others for a given experiment. Thus, Eggmeier and O'Donnell propose the following criteria be considered when selecting measurement techniques (T. O'Donnell & Eggemeier, 1986):

- **Sensitivity**: Capability of a technique to discriminate significant variations in the workload imposed by a task.

- **Diagnosticity**: Capability of a technique to discriminate the amount of workload imposed on different operator capacities (e.g. perceptual vs. motor resources).

- **Intrusiveness**: The tendency for a technique to cause disruptions in primary task performance - becoming a significant contributor of MWL by in and of itself.
- **Implementation Requirements**: The ease of implementing a particular technique.
- **Operator Acceptance**: Degree of willingness of the operators to follow instructions and utilize a particular technique.

2.2.5 Applications of Mental Workload

The significance of a consideration for cognitive workload is evident across a range of domains, from early work related to aviation systems (Reid & Nygren, 1988) to the applications in driver performance, e-learning, nursing (Mehler et al., 2009; Schewe & Vollrath, 2020; Yamagishi et al., 2007), augmented reality (Jeffri & Awang Rambli, 2021) *etc.*. In one such study, Pimenta *et al.* demonstrated that mental workload could be quantified by keyboard/mouse tracking and utilized to monitor mental fatigue in e-learning platforms (Pimenta et al., 2015). In another recent study, Pourteimour *et al.* investigated the impact of MWL on the performance of nurses providing care for Covid-19 patients (Pourteimour, Yaghmaei, & Babamohamadi, 2021).

2.3 Related Work

2.3.1 MWL, Usability and HCI

The relationship between usability and performance was disucssed in a recent review by Saket *et al.* (Saket, Endert, & Stasko, 2016). This is supported by Lehmann who discussed the necessity for multiple metrics when considering usability, including cognitive engagement (Lehmann, Lalmas, Yom-Tov, & Dupret, 2012). By understanding the impact of cognitive workload on digital task performance, computer systems can be better designed to optimize task performance. Despite interest in the relationship between cognitive workload and usability (Tracy, 2007), the two quantities were not

found to be correlated when investigated further. Luca *et al.* found that usability and cognitive workload, although not directly correlated, "can be jointly employed to improve the prediction of human performance, thus enhancing the description of user experience" (Longo & Dondio, 2016; Longo, 2018). It has been demonstrated that both cognitive underload and overload have negative consequences on performance (Lysaght et al., 1989; Young & Stanton, 2002; Biondi et al., 2021). This has significant implications in UX design, as situations of underload or overload may result in websites losing users, negatively affecting the website itself (Longo, 2014).

2.3.2 User Activity Analysis

Much research has been carried out to gain insight into internet user activity in the literature. A review by Woods *et al.* in 2015 highlighted the significance of widespread internet access within the context of usability research - as researchers are no longer limited to the "constraints of the Western, Educated, Industrialised, Rich and Democratic (*WEIRD*)" - opening up the possibility of cross-cultural psychological research. Woods discusses some downsides of technological platforms for user activity research, such as a lack of control over experimental conditions (screen size, resolution *etc.*), but maintains that the advantages of having access to a large, diverse pool of subjects in an inexpensive manner is hugely advantageous for such research (Knoeferle, Woods, Käppler, & Spence, 2015; Woods, Velasco, Levitan, Wan, & Spence, 2015).

Early work in user activity research involved the collecting and static analysis of log files (Rodriguez, 2002; Atterer, Wnuk, & Schmidt, 2006b). This yielded useful insight into website navigation, but analysis was slow, labour-some and asynchronous (not real-time) (Roy, Pattnaik, & Mall, 2014). Later, as webpages became more sophisticated and JavaScript grew in popularity, more sophisticated, real-time tools were developed. This typically involved the use of frameworks whereby the researcher selected the types of user events to measure - and the framework would then insert data-gathering JavaScript into HTTP file of the target Web site. The JavaScript code would then handle the measurement and logging of activity such as clicks, mouse movement *etc.* Examples of this technology include *Webvip* and *WebCAT* (Rodriguez,

2002; Scholtz, Laskowski, & Downey, 1998).

An increasingly diverse and sophisticated range of techniques are being employed in recent years to gain insight into usability and ergonomics, including the application of eye-tracking software and machine learning. With recent advances in artificial intelligence (AI), investigations have been carried out into the potential of AI as a means of evaluating user experience (Bakaev, Khvorostov, Heil, & Gaedke, 2017; Yang, Wei, He, Yan, & Liu, 2021; Sahi, 2018; Amelio, Draganov, Janković, & Tanikić, 2019). One such study, for example, applied fuzzy logic and clustering techniques to develop a tool for assessing UX based on mouse movement (Souza et al., 2019). In another recent study, Aviz *et al.* employed eye tracking as a means of detecting "hot spots" or areas of interest, which may have significance for UX purposes – gaining insight into features which draw a user's attention (Aviz et al., 2019).

2.3.3 User Activity and MWL

User activity research has significant implications in the field of MWL - and such techniques have been studied extensively in the literature. Research by Chen *et al.* demonstrated significant differences in handwriting corresponding to three distinct levels of MWL, and upon further research demonstrated a relationship between emotional state and stress levels with mouse and keyboard usage - corroborated by Liu *et al.* in 2003 (Chen et al., 2012; Liu, Wong, & Hui, 2003). A study by Pimenta *et al.* consisted of asking a group of students to perform tasks at the beginning and end of the day (when subjects were expected to be more tired). Multiple indicators of user interaction were collected and machine-learning (ML) was applied to compare the interaction data from the first and second assessments - demonstrating that fatigue is related to performance.

2.4 Current Research and State of the Art

As predicted by Longo in 2014 (Longo, 2014), advances in technology are facilitating physiological assessment of MWL. A study by Serra *et al* in 2019 utilized eye-tracking technology to investigate the relationship between workload and website complexity, and found that lower levels of MWL correlate with more positive evaluations of usability (Serra et al., 2019).

A study by Grimes and Valacich explored the potential of mouse movement as an indicator of cognitive workload, yielding promising results. This has particularly relevant implications for e-learning systems during the Covid-19 pandemic – by "observing when students exhibit changes in mouse behaviour, it may be possible to identify when they are having trouble understanding a concept - similar to seeing a confused look on a student's face..." (Grimes & Valacich, 2015).

2.4.1 Gaps in the Research

Whilst a wide variety of techniques have been utilized to investigate the relationship between user activity (mouse clicking, movement, focus *etc*) and user experience; much remains to be investigated regarding a user's subjective assessment of MWL and its relationship to objective indicators of activity.

With this in mind, this study aims to investigate the relationship between tangible indicators of user activity and subjective MWL scores. Well established multidimensional measures of MWL will be employed: NASA's Task Load Index (NASA-TLX), and the Workload Profile (WP) method. To achieve this, a suite of tools will be developed, using JavaScript, to gather metrics pertaining to user activity (mouse movement, clicks, focus, scrolling), and investigate their relationship with a user's subjective measure of MWL. This study is concerned primarily with the following research question:

> *"Is there a relationship between a user's subjective MWL when performing web-based tasks and objective indicators of tangible activity in the web browser?"*

Chapter 3

Experiment design and methodology

This chapter introduces the research hypothesis this study aims to investigate, and the design of the web interface which was implemented for to this end. Furthermore, it outlines the basis for which the interface was developed; including the data gathered, choice of tasks to induce cognitive workload, user surveys, *etc.* The primary objective of the experiment is to determine whether user behavioural data, gathered using JavaScript embedded in a web interface, is correlated with MWL, as determined by subjective user surveys.

Firstly, this chapter outlines the research hypothesis, and gives a high-level overview of the experiment by which this hypothesis will be investigated, briefly outlining some considerations and justifications for the methodologies used.

The subsequent section discusses the myriad types of data this experiment is concerned with from a psychological perspective, and provides some justification for these choices. This includes both the objective behavioural data as ascertained by the web-interface, and psychological data obtained from subjective user surveys.

Next, the design of the software component of the experiment is outlined. The automated gathering of user metrics is discussed here, as well as the data-model architecture which facilitates gathering of such metrics.

Lastly, this chapter summarizes some of the advantages and limitations the study,

and proposes some alternatives which may provide additional insight.

3.1 Research Hypothesis

By developing a framework which gathers user behavioural metrics, using JavaScript embedded in the browser, the following hypotheses will be investigated:

Null Hypothesis (H0) Metrics of user activity, obtained using JavaScript embedded in web tasks, exhibit no correlation with indicators of MWL.

Alternative Hypothesis (H1) Metrics of user activity, obtained using JavaScript embedded in web tasks, correlate highly with indicators of MWL.

3.2 Experiment Overview

In this work, a web-based application was developed and distributed to the participants. After implementation, the web-application was hosted on a publicly available domain, and left running over 2 months. This allowed adequate time to maximise the size of the pool of volunteers. The application features two tasks (or 'puzzles'), which the user is prompted to complete. Each task consists of multiple difficulty levels, (henceforth referred to as 'trials'), where each trial is proceeded by a survey. Whilst a user completes a given trial, their mouse activity is collected for analysis.

A more detailed breakdown of the experiment's core elements is given below:

- **Users**: Users are central to any investigation into MWL. Here, participants have been drawn from several sources - including, but not limited to, peers in the university, friends, family, colleagues *etc.* The application assigns an identifier to a particular user session based on browser metadata, which allows for identification of a single, unique visitor to the application and associates the behavioural data and surveys with that user, whilst retaining their anonymity. An optional form is displayed which allows a user to provide their age/gender.

- **Tasks**: In this study, two task interfaces (or 'tasks') have been implemented. In the first, a grid of coloured blocks is shown to the user, and an animation is played which demonstrates a particular pattern, which the user then attempts to repeat. Previous studies indicate that high levels of MWL can be induced by placing high levels of demand on a participants' working memory. (Mohammadian, Parsaei, Mokarami, & Kazemi, 2022; Norman, 2013).

  In the second task, users are presented with a grid of numbers, and an arithmetic challenge (*e.g.* 4×22). The user must select the correct answer from the grid, after which a new arithmetic problem and grid of numbers appears.

  In both tasks, one minute is allocated per difficulty level. The task refreshes with a new randomized version of the same task after each answer is submitted, until the time is elapsed, at which point the users are brought to a survey form. This ensures users are engaged for the entirety of the minute, and time-pressure is emphasized, with by a large stopwatch interface displayed at the top of the screen which displays the time remaining. Time-pressure is a key component of MWL, and was the first factor proposed for the SWAT measurement of MWL (Reid & Nygren, 1988).

- **Trials**: Each of the two tasks designed in this experiment consists of multiple difficulty levels, or trials. The 'repeat the pattern' game described above has been increased in difficulty by simply varying the grid size (*e.g.* 9 blocks *vs* 4 blocks), whist the arithmetic challenge features larger digits at the higher difficulty level (single digit numbers only *vs* double digits). The premise that increased task difficulties results in higher levels of cognitive workload is taken as a ground truth, and is supported in the literature (Allison & Polich, 2008; Fan, Zhao, Zhang, Luo, & Zhang, 2020).

- **Data Collection**: When a user begins a trial, a hidden HTML element is present in the background of the page. Using basic JavaScript, event-handlers are assigned to this HTML element which monitor and record mouse movement, clicks, hovers *etc* (Mozilla, 2022c). This data is associated with a particular trial,

which in turn is linked to a particular (anonymised) user. This data is stored in an SQL database for later analysis.

- **Surveys**: After each trial was completed (one minute elapsed), the user was immediately brought to a survey page, and asked to complete questions relating to the previous trial. By ensuring that the user is brought to the survey page after each trial, it is ensured that the user can rate the survey as accurately as possible from their recent experience. Questions from both the NASA-TLX and WP methodologies were used here.

A high-level overview of the application flow is illustrated below:

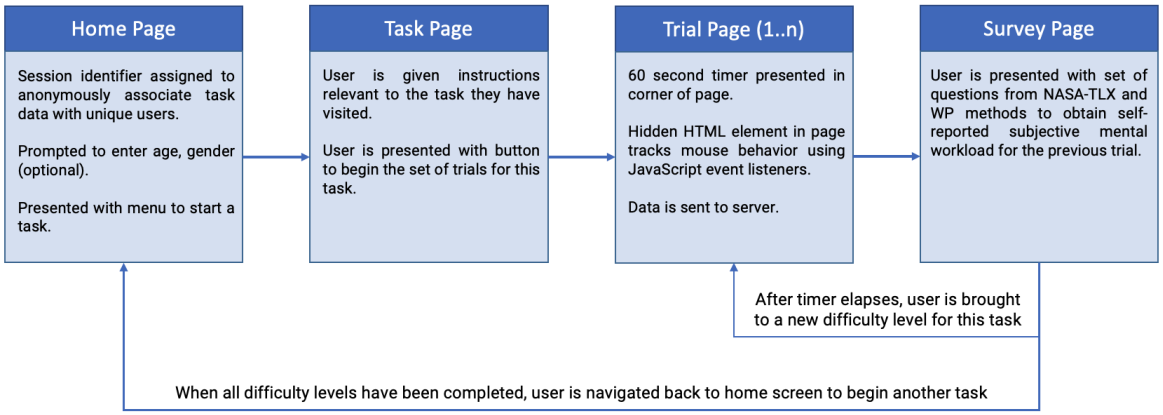

Figure 3.1: Overview of user-flow for the application used in the experiment. Here, '1..n' represents the relationship between tasks and trials, as a task may have 3 trials (*e.g.* easy, medium and hard difficulties).

3.3 Metrics

3.3.1 Mental Workload

Cain argues that it is "appropriate that mental workload be measured by subjective means, as it is a psychological construct" (Cain, 2007). Jex states that "In the absence of any single objective measure of the diffuse metacontroller's activity, the fundamental measure, against which all objective measures must be calibrated, is the

individual's subjective workload evaluation in each task" (Jex, 1988). In both cases, the researchers argue that, whilst "an operator is often an unreliable and invalid measuring instrument" (Gopher & Donchin, 1986), subjective measures of workload are the still considered appropriate methods to assess a loosely defined psychological construct such as mental workload. An advantage of survey-based assessment techniques are their non-intrusive nature, particularly important here when any distractions may impart additional cognitive load on a participant, skewing the results. Eggmeier argues that intrusion produces "significant problems in interpreting the data that result from the use of an assessment technique... The results of a procedure associated with significant degradation in primary task performance cannot accurately represent the degree of load required..." (R. D. O'Donnell & Eggemeier, 1986).

In the current work, two validated means of assessment are utilized; the NASA-TLX (Hart & Staveland, 1988) and WP (Tsang & Velazquez, 1996) processes. These are multi-dimensional and adopt an absolute scale design, and have been compared in several studies (Vidulich, 1988; Longo, Rusconi, Noce, & Barrett, 2012), concluding that both are reliable, with WP showing slightly greater sensitivity (Rubio et al., 2004). Both methodologies have low implementation requirements and high subject acceptability (Longo et al., 2012). As questionnaires were administered following trial performance, interference of these methodologies with performance is negligible.

Mental Workload

For both questionnaires, the user provides their answers in a 1-100 point scale. For the NASA-TLX questionnaire, the user was asked to rate the statements from "very low" (1) to "very high" (100) (NASA, 2022):

To allow for the simplest user-experience and higher participation, the typical sub-scale weighing is omitted in this experiment. This is known as the Raw TLX modification (Hart, 2006). Whilst the questionnaire applied in its unmodified form, the *physical demand* component of the questionnaire was omitted from calculations, as it plays an insignificant role in the web-based tasks studied herein.

MWL, according to the raw modification of the NASA-TLX survey is calculated

as follows:

$$\frac{NTQ_{MD} + NTQ_{TD} + NTQ_E + NTQ_P + NTQ_F}{5} = \frac{\sum_{x1} NTQ_x}{5} \tag{3.1}$$

$$NTQ_x : [0 \dots 100] \in N \tag{3.2}$$

$$X_1 : MD|TD|F|P|F$$

where MD = mental demand, E = effort, F = frustration, TD = temporal demand, P = performance.

For the workload profile questionnaire results, the standard formula was utilized (Longo et al., 2012; Longo & Dondio, 2016), with a slight modification, omitting the speech, verbal and auditory dimensions as they play no role in the tasks herein.

$$\frac{WP_{SD} + WP_{TS} + WP_{RS} + WP_{VA} + WP_{MA}}{100} \tag{3.3}$$

$$= WP_{HMW} = \frac{\sum_{i=1}^{5}(WP_i)}{100} \tag{3.4}$$

$$WP_{HMW} : [0 \dots 5] \in R$$

$$WP_{HMW} : [0 \dots 100] \in R$$

$$i : SD|TS|RS|VA|MA$$

where SD = solving and deciding, TS = task and space, RS = response selection, VA = visual attention, MA = manual activity.

The questionnaires are outlined in tables 3.1 and 3.2, overleaf, along with their relevant mappings for storage in the database.

Table 3.1: NASA-TLX question-database mapping

Dimension	Question	Database label
Mental	How mentally demanding was this task?	mental
Temporal	How hurried or rushed was the pace of the task?	temporal
Performance	How successful were you in accomplishing what you were asked to do?	performance
Effort	How hard did you have to work to accomplish your level of performance?	effort
Frustration	How insecure, discouraged, irritated, stressed and annoyed were you?	frustration

Table 3.2: WP question-database mapping

Dimension	Question	Database label
Perceptual	How much attention was required for activities like remembering, problem-solving decision-making, perceiving (detecting, recognizing and identifying objects)?	wp_perceptual
Response	How much attention was required for selecting the proper response channel (manual - keyboard/mouse, or speech - voice) and its execution?	wp_response
Spatial	How much attention was required for spatial processing (spatially pay attention around you)?	wp_spatial
Visual	How much attention was required for executing the task based on the information visually received (eyes)?	wp_visual
Manual	How much attention was required for manually respond to the task (eg. keyboard/mouse usage)?	wp_manual

Performance Indicators

Task performance is correlated with mental workload, with both cognitive underload and overload resulting in lesser task performance (Wilson & Sharples, 2015). As such performance metrics are commonly employed as indicators of MWL (Tsang & Vidulich, 2006; Hart & Staveland, 1988), and will be measured in this work as a user submits their answer for each trial. This calculation is given in equation 3.5, below:

$$correct\_answer\_frequency = \frac{N_C}{N_{Tot}} \tag{3.5}$$

where N_C = number of correct answers for a trial, N_{Tot} = total number of anwers for a trial.

For a particular trial, *e.g. block pattern task and easy difficulty*, the user may complete 10 rounds (or *TrialAttempts*) in the 60 second limit. As users' submitted answers are recorded, performance can be calculated by taking the ratio of correct attempts to total attempts.Furthermore, the number of *TrialAttempts* completed within the 60 second limit gives an indication of users' response time. The rate of *TrialAttempts* per minute as follows can be calculated as follows:

$$trial\_attempt\_frequency = \frac{N}{60} \tag{3.6}$$

where N = number of answers submitted for a trial.

Correlation Analysis

To determine the extent to which tangible indicators of user behaviour correlate with subjective indicators of MWL, a correlation analysis will be performed. Here, both following correlation coefficients will be applied (Miinitab, 2022):

- Pearson correlation, r, measures the extent of a linear relationship between two variables.
- Kendall Rank Correlation, τ, is an alternative to Pearson's correlation. It is non-parametric, and is favoured for small sample sizes with categorical data.

Using these analysis methods, the correlation between the following was investigated:

- All user interaction indicators and the MWL results as calculated *via* NASA-TLX.
- All user interaction indicators and the MWL results as calculated *via* WP.
- Specific indicators of user interaction the MWL results as calculated *via* NASA-TLX.
- Specific indicators of user interaction the MWL results as calculated *via* WP.
- Specific indicators of user interaction for a specific task the MWL results as calculated *via* NASA-TLX.
- Specific indicators of user interaction for a specific task the MWL results as calculated *via* WP.

3.3.2 User Interaction

The utility of mouse-tracking data as a behavioural indicator has been demonstrated extensively in the fields of ergonomics and HCI. Research indicates that unconscious mouse activity during the completion of a task is related to the level of cognitive load experienced by the subject (Cha & Min, 2022). In general, during periods of higher cognitive load, less mouse activity is observed (Grimes & Valacich, 2015; Rheem, Verma, & Becker, 2018; Garavan, Ross, Murphy, Roche, & Stein, 2002). In 2012, Papesh and Goldinger utilized *via* a simple digital interface, whereby users were asked to distinguish words they memorized from newly presented words by clicking "old" or "new" buttons. They demonstrated that, as participants' confidence in their decisions increased (measured by the 7 point Likert scale), their mouse movements were faster and their answers required shorter response times (Papesh & Goldinger, 2012).

Mouse-tracking data offers a range of practical advantages when compared with other techniques. It was first introduced as a cost-effective alternative for EEG and

eye-tracking methods (Freeman & Ambady, 2010), and is less intrusive than both methods. Furthermore, it provides a rich set of information, such as response time, mouse trajectory and velocity all at high temporal resolution (Rheem et al., 2018).

In this work, the following mouse data is collected:

- **Mouse Click**: By tracking clicks, metrics such as response time can be calculated (Mozilla, 2022a). The coordinates of the clicks as well as their time is recorded, which will allow for insight into the intent of the user.
- **Mouse Position**: Tracking mouse position allows for the calculation of a range of derived metrics, such as average mouse velocity, mouse distance travelled, *etc* (Mozilla, 2022d). This metric is considered especially useful for web developers (Atterer et al., 2006b), as it indicates where a user's attention is focused. With modern high-resolution screens, a mouse traverses a large number of pixels in a small time-frame, which can be captured with a timestamp, giving high-resolution data for user behaviour.
- **Element Hover**: Hovering over an element may indicate uncertainty about a decision, and it is anticipated that this metric will prove useful for the arithmetic game where a user must select an answer from a range of numbers (Mozilla, 2022e).

3.4 Tasks

3.4.1 Guiding Principles

When considering the design of tasks appropriate for inducing mental workload, it is necessary to consider the factors which contribute to MWL. Galy *et al.* define three primary criteria to consider when evaluating a tasks' MWL demands; task difficulty, time pressure, and mental arousal (Galy, Cariou, & Mélan, 2012).

Task Difficulty

In 1994, Backs and Seljos demonstrated the effect of task difficulty on mental workload as measured by the NASA-TLX method, which agreed with physiological measures (Backs & Seljos, 1994). This is corroborated by Ayres' research in 2006, who showed that error-rates correlate strongly with task difficulty as determined by subjective measures (Ayres, 2006). This research indicates that, for the purposes of task selection and design in MWL research, granular control of task difficulty is desirable. Allison and Polich applied this principle, and demonstrated a correlation in physiological measures of MWL and difficulty settings in a video game (Allison & Polich, 2008).

Time Pressure

Whilst not directly impacting task difficulty, time pressure imparts a psychological pressure on subjects, which "activates an emotional component and would thus have an indirect effect on cognitive load" (Monod & Kapitaniak, 2003). The effect of time-pressure on a subjects perceived level of mental workload is well documented in the literature (Yi, Qiu, Fan, Zhang, & Ming, 2022; Inzana, Driskell, Salas, & Johnston, 1996).

Mental Arousal

A rich body of research exists on the relationship between physiological health and mental workload (Zacks, 2004; Zare et al., 2016; Barnes, 2015). It is unsurprising that a subjects' performance in a given task directly depends on their functional state. Monk and Leng refer to this as the "time of day effect" on performance, which parallels the circadian rhythm (Monk & Leng, 1982). Research indicates that memory-based tasks done in the evening time result in higher performance. Smit *et al.* demonstrated a correlation between subject alertness (measured *via* EEG), task performance and self-reported MWL (Smit, Eling, Hopman, & Coenen, 2005).

Taking the above into consideration the following design principles were applied

for the tasks:

- **Task Difficulty**: The tasks must allow for reliable variation of difficulty levels. At higher difficulty levels, greater cognitive load can be observed in the users' behaviour.
- **Time Pressure**: An element of time pressure should be emphasized in the tasks. By applying a time limit to the tasks, the user allocates greater cognitive resources rather than working slower at the same level of cognitive load for a given task.
- **Mental Arousal**: As mental arousal affects cognitive workload and task performance, it is preferable to allow the user to complete the task at a time convenient for them, rather than at a given time when they may be otherwise fatigued (*e.g.* early in the morning or after work).

Two tasks have been developed for this experiment to best fit the above design criteria. *Block Pattern* and *Math Challenge*. They are discussed in detail in section 3.4.3.

3.5 Software Design

For the software component of this study, several key principles were considered:

- **Availability**: By designing the application as a publicly available web application, test is accessible and convenient for a potentially large group of participants. Whilst a less controlled environment than a laboratory-based setting, web-based experimentation offers several benefits. Steenbergen and Bocanegra argue that "doing a research in a lab is inherently constraining in terms of the participants' characteristics and the contextual variables that can be systematically investigated", and online methods offer "a relativey inexpensive method to test hundreds of participants in a couple of hours... also facilitate(s) transparency and quicker replication by fellow researchers". (van Steenbergen & Bocanegra,

2016). This is already the case in many big-data studies, where vast quantities of social media data are analysed (Griffiths, 2015).

- **Scalability**: The web application should be available a potentially large group of participants, independent of geographical differences. Furthermore, the application should be extensible, allowing for facile fine-tuning of tasks, development of new tasks, *etc.* This opens up a variety of interesting avenues for future research, such as comparisons between users, differences in behaviour between tasks and the effect on observed MWL, *etc.*

- **Automation**: The application should be designed in order to minimize manual intervention by the experiment organizers. To this end, the NASA-TLX and WP surveys are integrated into the application as web-forms, raw mouse behavioural data is collected and parsed into an appropriate format before storing in a database for subsequent analysis *etc.* This ensures that the experiment can cater to a potentially large sample group, and ensures that future development of the application (*e.g.* adding a third task) requires no additional overhead in terms of software development. This is known as modularity in software development (Sullivan, Griswold, Cai, & Hallen, 2001).

With this in mind, the a multi-layer design was utilized for the application, based on a standard client-server architecture (Özsu, 2016). The application consists of four layers; the HCI layer (client), the server layer, a series of data preprocessig utilities, and the persistence layer.

- **Client**: The client is a standard front-end web application, built using the *ReactJS* framework (Meta, 2022). React allows for rapid prototyping and development of responsive web applications, based on a modular, component-based architecture. As of 2021, *ReactJS* ranked first in terms of popularity amongst web developers, surpassing *jQuery* in usage (Statista, 2021). The client-side application will render a user interface, including tasks, and send the task performance data, as well as user behavioural data to the server to processing and subsequent storage.

- **Server**: The server will follow a representational state transfer (REST) design pattern, which maps HTTP requests and their methods (get, post, put, delete) to user navigation and the create, read, update, delete (CRUD) operations of domain objects in the application (Jazayeri, 2007). The REST design pattern was developed by Roy Fielding during his PhD research, and has since become a cornerstone of modern server-side web development (Fielding & Taylor, 2000; Richards, 2006).

- **Data Preprocessing**: Raw interaction data, captured *via* the event API (Mozilla, 2022b), is sent to the server, and will be parsed and stored in a format amenable to storage in a relational database. This includes the nested association of tasks with unique user sessions, each task having multiple trials (difficulty levels for a given task), each trial having multiple user attempts. By offloading the complex relationships between the data to the sever, an intuitive user interface is presented where a user simply completes a set of tasks and the complexity is transparently handled at the server layer. "Transparency" is a common design pattern in software engineering whereby complexity is hidden from the user, in our case, the end-user of the web-application. Leite *et al.* provide a review of the concept (Leite & Cappelli, 2010).

- **Data Persistence**: Keeping the relational nature of the data in mind, as previous outlined (users having multiple tasks, with multiple trials *etc*), a relational SQL database was utilized in this work.

The four-layered design of the application is outlined below:

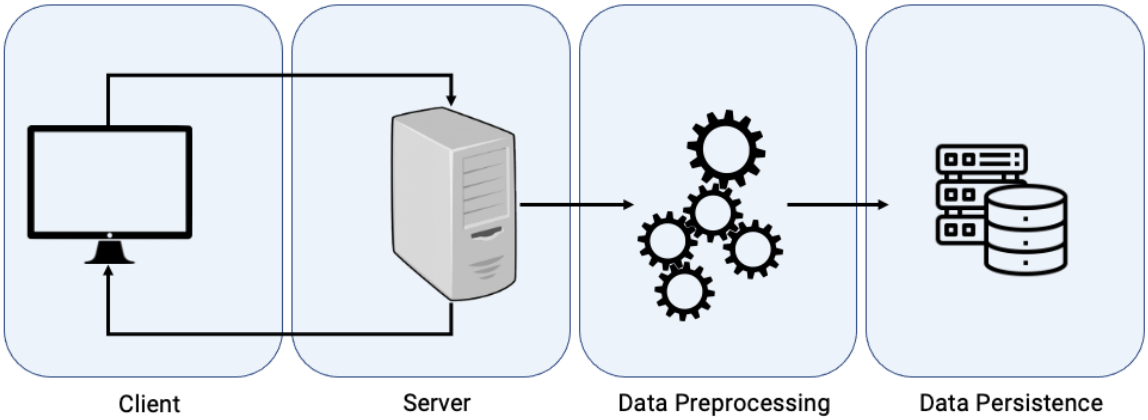

Figure 3.2: High-level overview of the application design.

3.5.1 Data Model Architecture

The relational nature of the data naturally lends itself to a hierarchical data model. This provides opportunity for interesting insight into the results, comparing performance and perceived MWL on at the task level or at the group level. For example, with this architecture, it is possible to study mouse behaviour for a particular task, or for a particular user, by isolating subsets of the data-model graph outlined below.

Using this data model with a relational database allows for powerful querying capabilities. One potential use case, for example, might involve investigating the mouse click activity for a particular user. The following SQL command would retrieve this data:

```
SELECT * FROM mouse_clicks
JOIN trials t on t.id = mouse_clicks.trial_id
JOIN tasks t2 on t2.id = t.task_id
JOIN users u on u.id = t2.user_id
WHERE user_id = 3;
```

Listing 3.1: SQL command to fetch mouse clicks for a particular user.

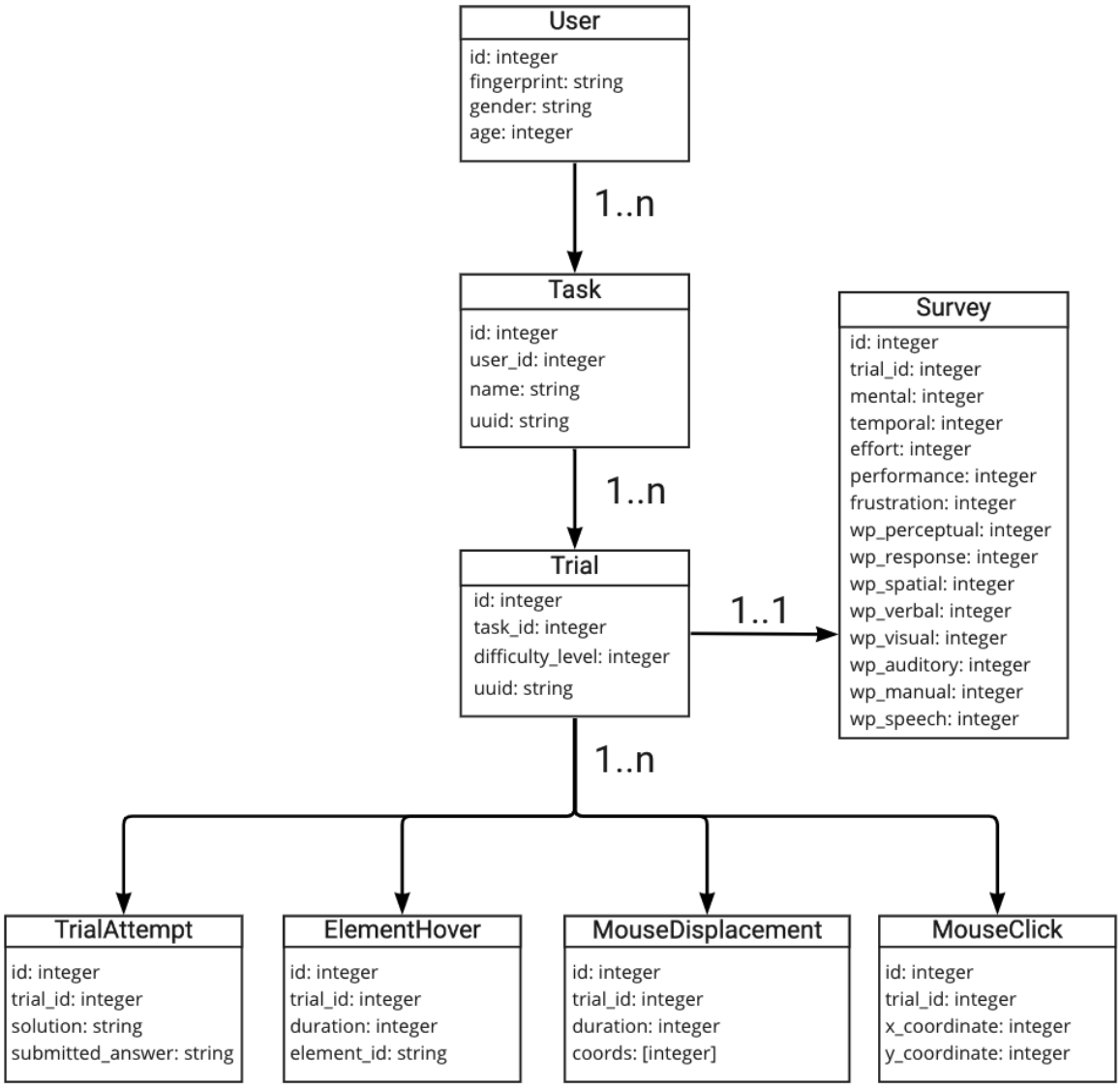

Figure 3.3: Data model architecture for the web application. Note the relationships between the domain objects; *i.e.* a trial is associated with many mouse clicks (1..n), but a single survey (1..1).

3.5.2 Interface Design

The following user interface (UI) flow design was implemented for this experiment:

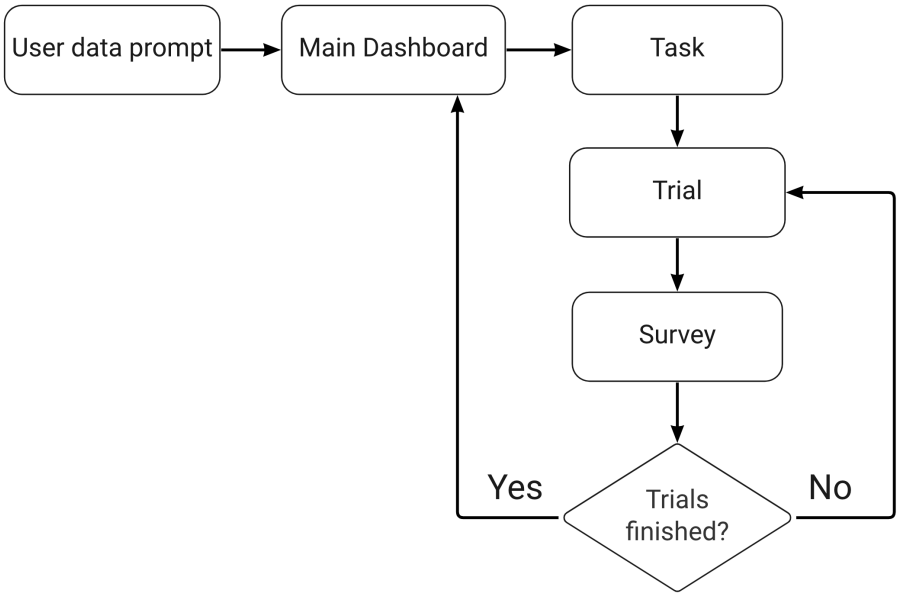

Figure 3.4: UI flow diagram for application.

After proceeding to task selection (stage 3 in the above diagram), the user may start the trials. A set of difficulty levels is defined for each task on the server, and these parameters are used to generate random trials for a given task at a specific difficulty level. When the user chooses to start a task, the client sends a HTTP request to the server. The server then responds with the set of trials (easy, medium, hard difficulties in random order) as JSON, which the client consumes before rendering a trial.

The user interface was built using ReactJS, and is illustrated overleaf.

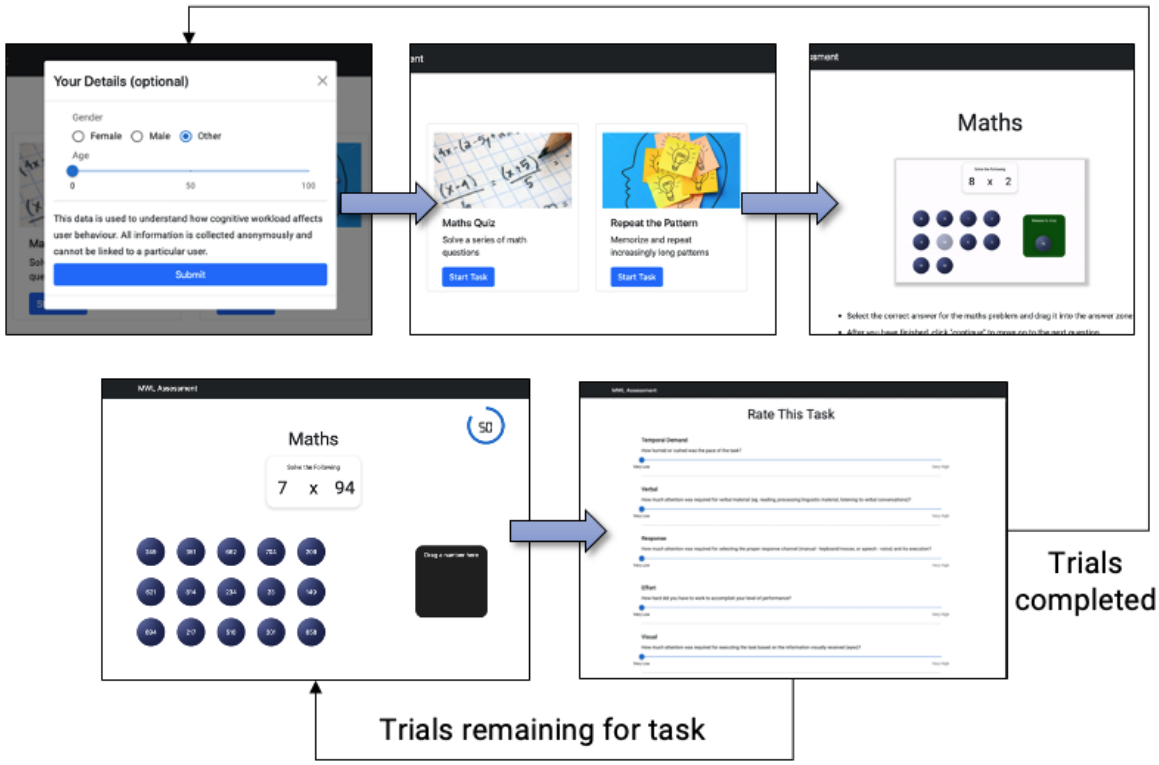

Figure 3.5: UI developed for MWL experiment. From top left: 1. Optional prompt to capture user metadata (age/gender). 2. Main application dashboard where users select a task. 3. Introductory screen for a task, with instructions and an animation of the task being at hand. 4. Trial in progress (task at a given difficulty level). 5. Survey screen.

3.5.3 Tasks

Block Pattern

For this task, the user is presented with a grid of boxes. An animation appears whereby the boxes temporarily glow in a particular pattern. The user must then recreate the pattern displayed by clicking on the appropriate boxes. This task primarily involves a users' working memory. Research on MWL suggests that the amount of information held in working is strongly related to cognitive load (Rheem et al., 2018). In this task, the user must rely entirely on visual observation, working memory, and mouse control to repeat the pattern observed. In one study, Sweller demonstrated that the amount of information held in working memory has a direct impact on cognitive load

(Sweller, 1988). In another, Garavan *et al.* found a relationship between goal-driven motor activity is related to MWL (Garavan et al., 2002). Thus, it is expected that the fine motor control required for this task, in combination with its load on working memory, will have a significant and impact on MWL which will be observable in the mouse behavioural data recorded.

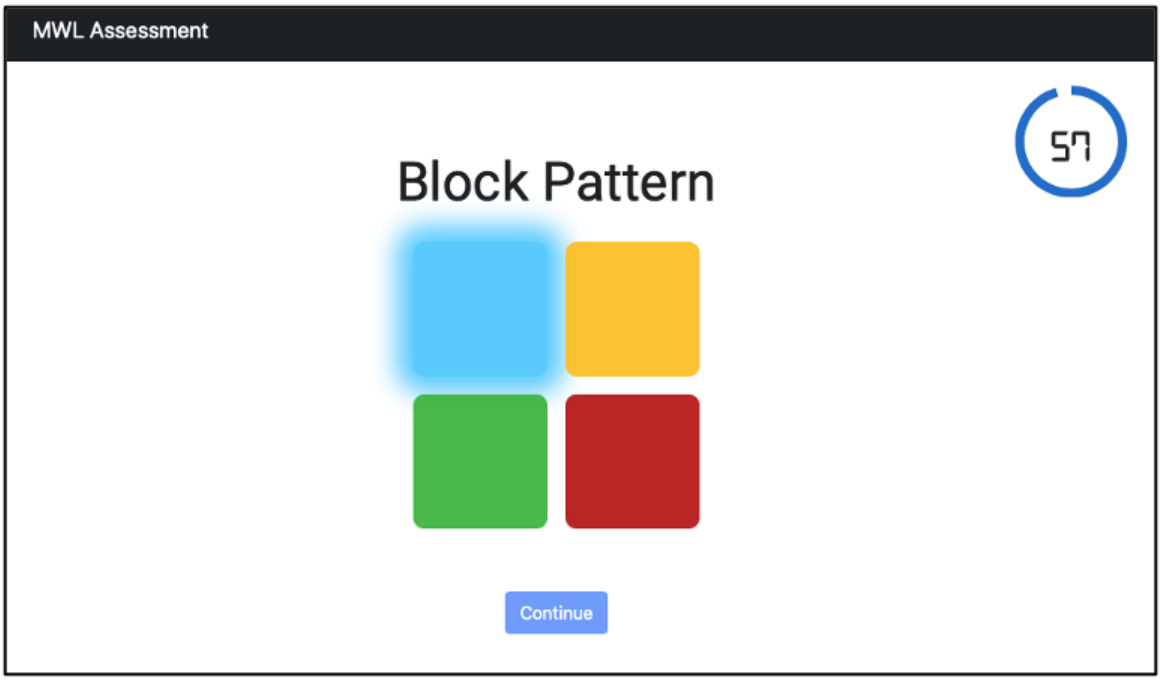

Figure 3.6: Block pattern task developed for this experiment.

Math Challenge

In *Math Challenge*, users will be presented with an arithmetic problem, and must click the correct answer from an group of possible choices. The challenge here is twofold; solving the arithmetic problem, and finding the correct number in a grid of numbers. This goal here is to induce stress on visual resources, working memory and numeracy skills.

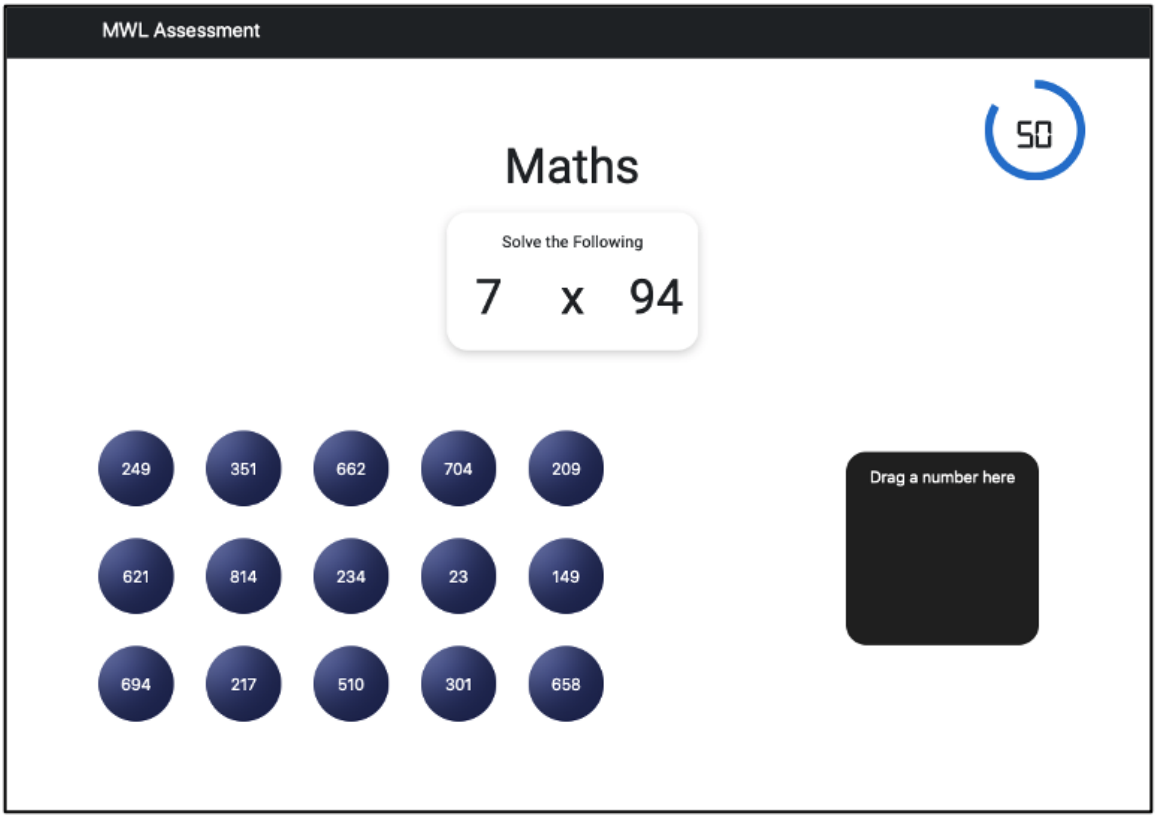

Figure 3.7: Arithmetic challenge task developed for this experiment.

In both tasks, difficulty can easily be attenuated by changing task parameters; pattern length for *Block Pattern*, or number size for *Math Challenge*. To elicit the desired time-pressure response, each task (at a given difficulty level) has a time limit of 60 seconds, which will be displayed as a timer which counts down at the top of the screen. When the time has elapsed, the user will be prompted to complete a survey, before beginning the next task. The parameter(s) for controlling the difficulty of the two tasks are outlined in the tables below.

Table 3.3: Parameter for difficulty adjustment in arithmetic task

Parameter	Easy	Hard
Number size	0..9	0..100

Table 3.4: Parameters for difficulty adjustment in block pattern task

Parameter	Easy	Medium	Hard
Pattern length	4	5	6
Number of blocks	4	4	6

3.6 Strengths and Limitations of the Design

This chapter outlines the design of the experiment from a psychological and ergonomics perspective, and also from a software design perspective. Keeping in mind the goal of the experiment; to investigate the correlation between concrete indicators of user activity and subjective indicators of MWL, this experiment possesses both strengths and limitations.

3.6.1 Strengths

- Being web-based, the experiment is inexpensive to perform, and can scale to a potentially large sample group without the practical issues a lab-based study would face.
- The methods for measuring MWL are well established in the literature, and show high sensitivity, reliability and subject acceptance. These methods are non-intrusive, and, by carrying out the questionnaires after each trial, there is a negligible effect on subject performance.
- The user is seamlessly navigated the user through the experiment with no involvement necessary on the part of the experiment operator, including questionnaires.
- Using JavaScript and HTML event-listeners, data is collected in at a millisecond level of granularity.
- Data is collected and processed without affecting the user experience.
- The design of the application allows for facile implementation of new tasks.

- The client consumes trial data from the server. By simply updating *via* configuration files on the server, difficulty levels can be easily introduced or modified without affecting the user experience.

- By designing multiple tasks, the likelihood of specific interaction indicators being under-represented in the dataset is reduced.

3.6.2 Limitations

- Being web-based, the experiment is conducted in an inherently less controlled environment. One cannot ensure the user wasn't distracted during the course of an experiment. Furthermore, there is a possibility that users can submit random values to skip the lengthy questionnaire.

- By focusing only on mouse activity, several common indicators of user activity have been omitted, including scrolling, keystrokes *etc.*

- A greater number of tasks would be desirable; the tasks designed herein are simple in nature and test only specific areas of cognition, in particular working memory.

Chapter 4

Results, Evaluation and Discussion

In this chapter the results of the experiment are presented, analysed and discussed, in the context of the objectives of this investigation. Initially, an exploratory overview of the data is presented, to provide better insight into the results set. The data obtained for behavioural indicators and MWL questionnaires is then presented. Finally, a correlation analysis is undertaken, thereby fulfilling the core objective of this research; to determine if a correlation between subjective MWL metrics and objective indicators of user activity exists. Finally finally, the results of the investigation are summarised and critically analysed, forming the basis of suggestions for future research.

4.1 Evaluation

In total, participation for the experiment totals 33 unique users. As previously discussed, a major disadvantage of conducting an experiment online is a lack of control over the environment in which the experiment takes place. As a consequence of this, a large portion of users abandoned the experiment before trial data could be recorded. The resulting dataset totals 15 unique users, completing a sum total of 65 trials. From the 15 unique users, 13 provided optional personal information (gender, age). This is summarised below:

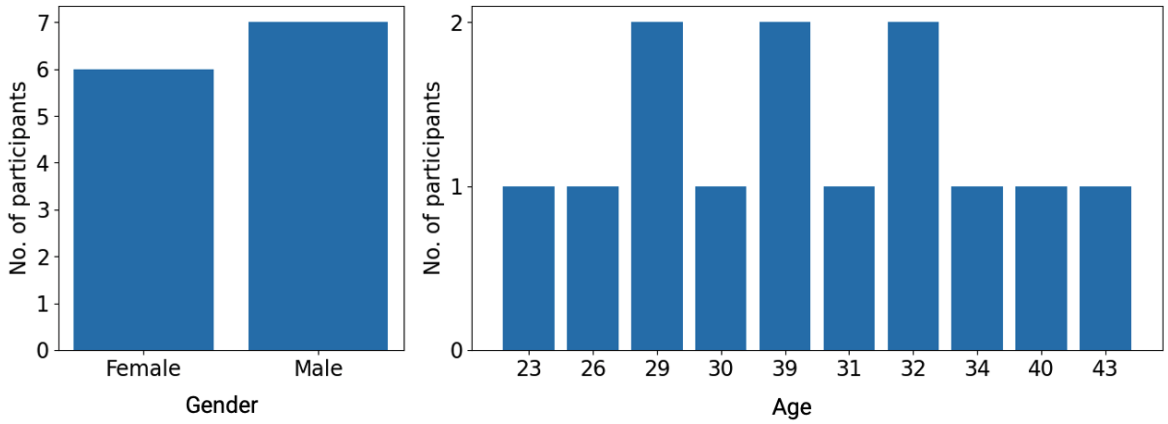

Figure 4.1: Gender (left) and age (right) of participants who voluntarily submitted this information.

Of the 65 trials, 53 have completed both NASA-TLX and WP surveys. The trial data completion-rate is summarised in Table 4.1, below. The dataset can be viewed in detail in Appendix Fig. A1.

Table 4.1: Results metadata

Column	Non-null count	Type
trial_id	65	integer
user_id	65	integer
task_name	65	integer
difficulty	65	integer
nasa_mwl	53	float
wp_mwl	53	float
tot_mouse_clicks	65	integer
total_mouse_distance	54	float
average_mouse_velocity	54	float
element_hover_time	65	float
correct_answer_frequency	57	float
attempts_frequency	65	float

The results of the mental workload measurements are illustrated in the box plot dia-

grams below. For the NASA-TLX methodology, MWL varies in an intuitive manner, increasing with difficulty. Workload profile however does not show a clear trend with respect to the difficulty levels of the trial. This may in be due to the small sample size. The NASA-TLX results, however, demonstrate that design of the tasks, with multiple difficulty levels for the trials, imposes the desired cognitive load on users.

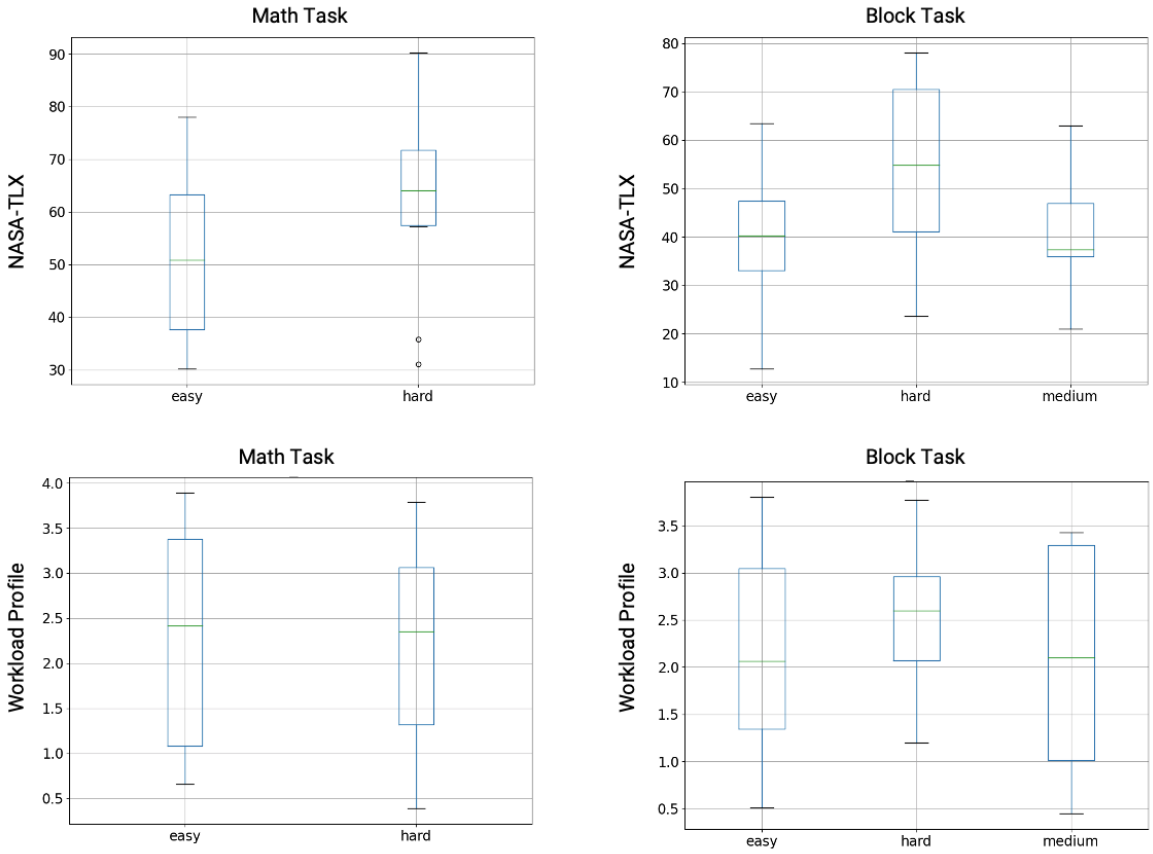

Figure 4.2: Box plots of the MWL results obtained from experimentation.

To gain an initial insight into any potential correlations that exist within the dataset, pairwise scatter plot matrices were generated. This will give a visual overview of the data and an early insight into possibly correlation between variables (Pandas, 2022b).

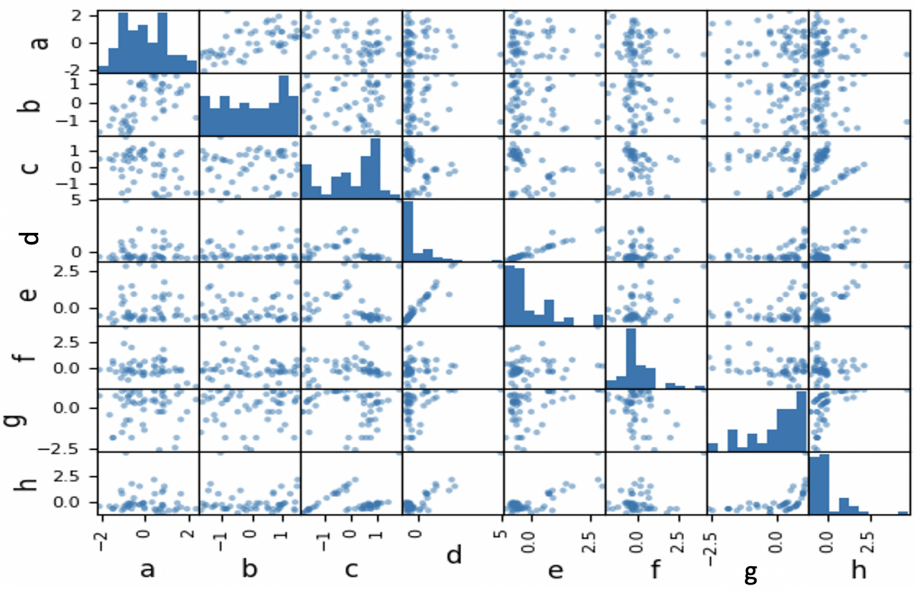

Figure 4.3: Pairwise scatter matrices of numerical data obtained across all trials. A - MWL (NASA-TLX), B-MWL (WP), C - Total mouse clicks, D - Total mouse distance, E - Average mouse velocity (px/s), F - Element hover time (s), G - Correct answer frequency (%), H - Trial attempt frequency (attempts/min)

From an initial overview of the scatter matrix of all trials (where each column of numerical data is plotted against each other in the table), some clear correlation can be observed. These will be outlined below using their grid coordinates:

- **(a, b)**: It is unsurprising that the two measures of mental workload employed show some correlation. This verifies that, for the most part, the participants were consistent in their ratings of the tasks. A user that rated a task high according to NASA-TLX appears to have done likewise for WP. This however, does not yield insight into the primary objective, relating MWL to mouse activity.

- **(a, g)**: There appears to be a loose correlation between NASA-TLX rating and correct answer frequency. The sample size, however, is too small to draw conclusive observations, and outliers are significant.

- **(c, d),(c, h)**: Total mouse clicks (c) shows some linearity with respect to total mouse distance (d). This can be intuitively understood as a mouse-move event

typically precedes a click event. Mouse clicks (c) similarly appears to display correlation with trial attempts frequency (h), which can be understood as the user progressing through trials is likely to click more frequently. Whilst not an indicator of correlation between MWL and mouse activity, these intuitive relationships demonstrate the effective design and implementation of the experiment. This is illustrated in the pairwise scatter matrices obtained from both trial types, below.

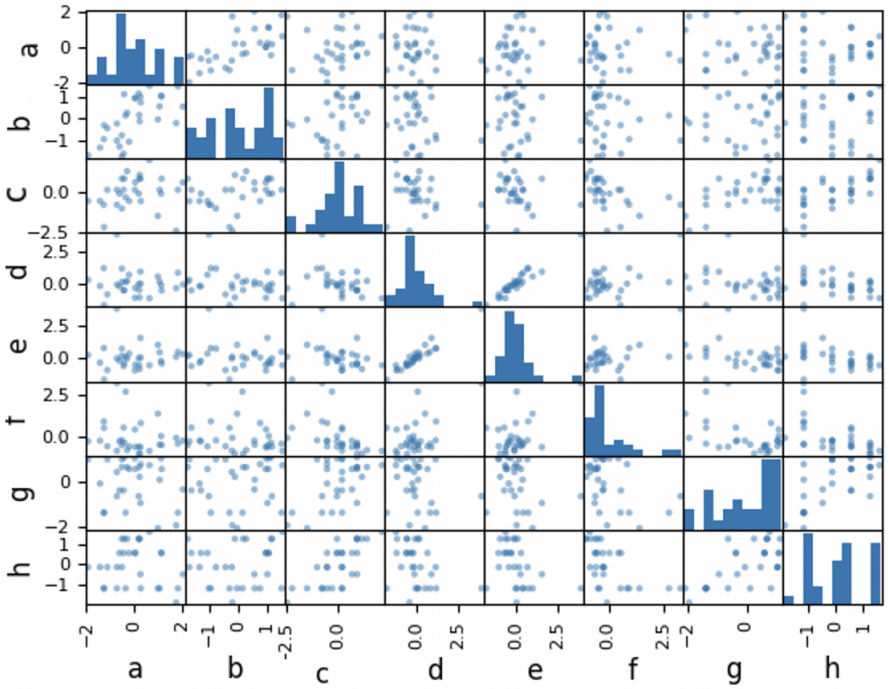

Figure 4.4: Pairwise scatter matrices of numerical data obtained from block pattern trial. A - MWL (NASA-TLX), B - MWL (WP), C - Total mouse clicks, E - Total mouse distance, E - Average mouse velocity (px/s), F - Element hover time (s), G - Correct answer frequency (%), H - Trial attempt frequency (attempts/min)).

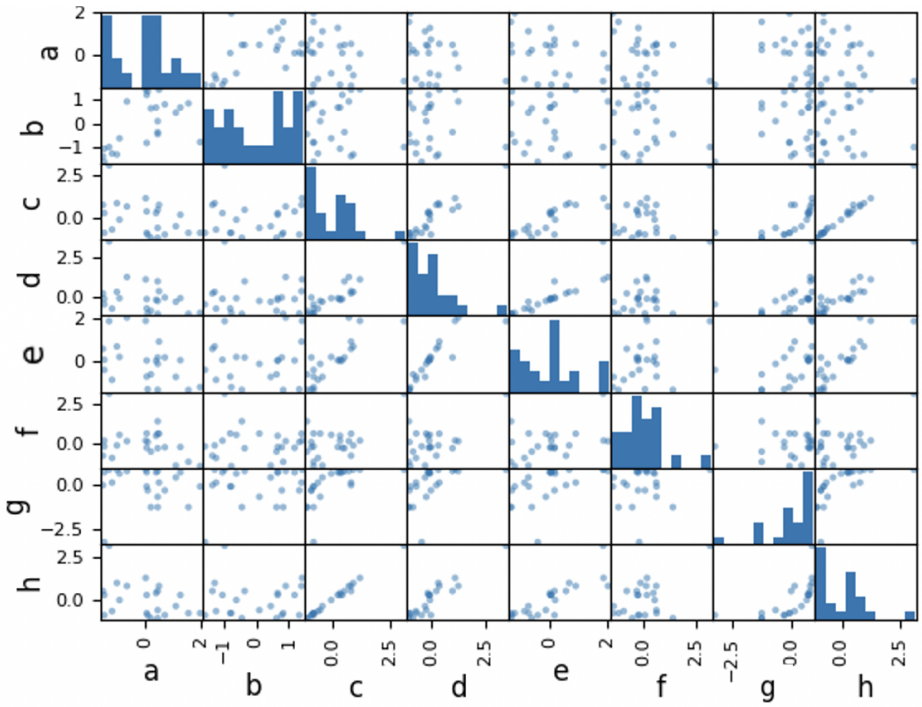

Figure 4.5: Pairwise scatter matrices of numerical data obtained from maths trial. A - MWL (NASA-TLX), B - MWL (WP), C - Total mouse clicks, E - Total mouse distance, E - Average mouse velocity (px/s), F - Element hover time (s), G - Correct answer frequency (%), H - Trial attempt frequency (attempts/min)).

A full breakdown of the correlations between different measurements is given in table 4.2, below.

Table 4.2: Pearson correlation analysis of trials data. Columns: A = MWL value calculated from NASA-TLX, B = MWL value calculated from WP, C = number of mouse clicks, D = total mouse distance moved (px), E = average mouse velocity (px/s), F = total mouse hover time (s), G= percentage of successful trial answers, H = trial answer submissions per minute.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	NA	(0.57, 0.00)	(-0.38, 0.01)	(-0.01, 0.92)	(0.02, 0.89)	(-0.24, 0.08)	(0.02, 0.88)	(-0.09, 0.50)
B	(0.57, 0.00)	NA	(0.05, 0.71)	(-0.05, 0.73)	(0.04, 0.80)	(-0.13, 0.37)	(0.08, 0.56)	(0.00, 0.98)
C	(-0.38, 0.01)	(0.05, 0.71)	NA	(-0.03, 0.82)	(-0.16, 0.26)	(0.18, 0.21)	(-0.10, 0.47)	(0.22, 0.11)
D	(-0.01, 0.92)	(-0.05, 0.73)	(-0.03, 0.82)	NA	(0.91, 0.00)	(0.44, 0.00)	(0.41, 0.00)	(0.90, 0.00)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
E	(0.02, 0.89)	(0.04, 0.80)	(-0.16, 0.26)	(0.91, 0.00)	NA	(0.33, 0.01)	(0.41, 0.00)	(0.81, 0.00)
F	(-0.24, 0.08)	(-0.13, 0.37)	(0.18, 0.21)	(0.44, 0.00)	(0.33, 0.01)	NA	(-0.15, 0.29)	(0.35, 0.01)
G	(0.02, 0.88)	(0.08, 0.56)	(-0.10, 0.47)	(0.41, 0.00)	(0.41, 0.00)	(-0.15, 0.29)	NA	(0.51, 0.00)
H	(-0.09, 0.50)	(0.00, 0.98)	(0.22, 0.11)	(0.90, 0.00)	(0.81, 0.00)	(0.35, 0.01)	(0.51, 0.00)	NA

(Pearson r, Two-tailed p value)

A correlation analysis was then carried out using the Kendall's tau (τ), which measures the correspondence between two rankings. In our case, the difficulty level of a trial and mouse behaviour. Values close to 1 indicate strong agreement, and values close to -1 indicate strong disagreement

Table 4.4: Kendall's τ correlation analysis of a trials' difficulty level and data recorded from experimentation across all trials.*NASA* = MWL value calculated from NASA-TLX, *WP* = MWL value calculated from WP, *Clicks* = number of mouse clicks, *tot. mouse* = total mouse distance moved (px), *Mouse_vel* = average mouse velocity (px/s), *Tot. hover* = total mouse hover time (s), *cor. freq* = percentage of successful trial answers, *answer/min* = trial answer submissions per minute.

Metric	τ
NASA	(0.017,0.875)
WP	(-0.0136,0.901)
clicks	(0.048,0.665)
tot. mouse	(-0.296,0.006)
mouse_vel	(-0.310,0.004)
tot. hover	(-0.037,0.734)
cor. freq	(-0.282,0.011)
answer/min	(-0.330,0.003)

(Kendall's τ, p value)

4.2 Discussion

The work presented demonstrates the design, implementation and application of a framework for assessing mouse activity and mental workload required during web-based tasks, and subsequent data analysis was carried out on the results obtained. From the experimental results obtained, correlation was found between several of the behavioural metrics themselves (*e.g.* mouse velocity and trial attempts in Table 4.2), which reflects positively on the design of the system.

With regards to behavioural metrics and MWL, however, the same can not be said. Weak correlation was found between several indicators of mouse activity and MWL, but a statistically significant correlation could not be demonstrated. The very limited size of the dataset, as well as incomplete data, resulted in a participant group of only 15 users, out of 33 total (55% reduction in size after removing incomplete data), with 58 trials completed. As noted by Reips, a major disadvantage of web-based experimentation is the lack of a controlled environment (Reips, 2000). Whilst this was anticipated as a possible issue during the design review (section 3.5.2), it was hoped that the accessibility of a web-based experiment would be compensatory, even if a large quantity of data was invalid.

It is hoped that this preliminary work and lessons learned herein are utilized to conduct further research, with an emphasis on increasing the participant population. This work nonetheless demonstrates a proof-of-concept for an automated MWL analysis system with interesting potential for future work.

4.3 Hypotheses

Whilst further research with a larger dataset is required for conclusive evidence, the results obtained in this work support the null hypothesis (H0):

Null hypothesis (H0): Metrics of user activity, obtained using JavaScript embedded in web tasks, exhibit no correlation with indicators of MWL.

4.4 Summary

The dataset obtained in this work was first subject to a preliminary round of "cleaning", whereby incomplete data was removed. Next, scatter plots were utilized to gain an initial insight into potential correlations in the data. Correlation analysis was then conducted between MWL results, and mouse behavioural metrics obtained *via* the web-application. Both Pearson and Kendall correlation methods were applied, and no significant correlation could be demonstrated between mouse metrics and MWL results obtained. Whilst the results obtained are in agreement with the null hypothesis, it is anticipated that the size of the dataset was a limiting factor here. Further investigation is required in this area, and recommendations have been made for such work.

Chapter 5

Conclusion

This chapter provides an overview of the project and its findings. This works' contribution to the body of knowledge for the field is considered, and suggestions and recommendations for further research are outlined.

5.1 Research Overview and Problem Definition

The aim of this research is to determine whether measurable indicators of user activity correlate with subjective, self-reported ratings of mental workload within the context of human-computer interaction. This is an area of growing significance in the 21^{th} century, as the world grows increasingly digital. This project outlines the development of a web-application which monitors and records user mouse activity in real-time, and its application in investigating the relationship between said data and subjective measures of mental workload.

5.2 Design, Experimentation, Evaluation and Results

This project entails the design and development of a web-based framework which provides users with a series of challenges, at a range of customizable difficulty levels, and records mouse activity throughout. Participation in the online experiment totalled

33 unique users, although incomplete surveys and abandonment resulted in a limited final dataset for analysis. Preliminary findings indicate some correlation between specific indicators of mouse activity, but no correlation could be found between these indicators and mental workload ratings.

5.3 Contributions and Impact

The the objective of this work is to determine whether a correlation can be found between objective indicators of user activity and subjective measures of mental workload. Due to limitations in the size of the dataset analysed, further research is warranted. This results obtained in this project reveal no correlation between the aforementioned quantities. In summary, these preliminary findings support the null hypothesis, *H0*:
Null Hypothesis (H0): Metrics of user activity, obtained using JavaScript embedded in web tasks, do not correlate with indicators of MWL.

5.4 Future Work and recommendations

A number of considerations and recommendations for future research have come to light during this work. The primary concern for subsequent research in this area is the necessity of an extended trials period. Due to insufficient time and participation, the participant pool was too small to draw conclusive results. On one hand, conducting the experiments online allowed for easier access to a potentially large group of participants, but similarly, the experimental environment is inherently less controlled, and thus abandonment rates and invalid data submission was high. In order to have a data set of substantial size, accounting for these issues, a participant pool numbering several hundred to a thousand applications is recommended - as it must be anticipated that a sizeable portion of the data will be discarded after parsing and cleaning the results set.

Furthermore, consideration should be paid toward the makeup of the sample population. In this study, participant age ranged from 20-40, a narrow bracket. Most

participants were also involved in computer science or software development (most participants were students or colleagues). This group is thus not representative of the population as a whole, and this should be mitigated in further research.
In relation to the design of the experiment itself, several recommendations can be made:

- **Tasks**: Designing and implementing a more diverse range of tasks will result in two main benefits. Firstly, more variation in task design would allow for different 'games' which require (and track) different interaction metrics, such as keyboard usage, scrolling, *etc.* Secondly, a broader task selection may lead to higher participation rates, as users are more likely to find a task they enjoy. One avenue which

- **Alternative measures of MWL**: This study focused exclusively on subjective measures of MWL and the correlation with objective indicators of mouse activity. A natural question that arises here concerns the possibility of mouse activity being more closely related with physiological measures of MWL. It is easy to understand why eye movement, for example, may be more closely related to mouse activity - as a user must see where they are aiming their cursor prior to a click event. Kapowski *et al.* have conducted research into the relationship between eye movement and mouse dynamics, with interesting implications for MWL research (Kasprowski & Harezlak, 2018).

- **Data Mining**: Due to limitations with the size of the dataset in this experiment, knowledge extraction techniques were not applied. For future work, machine learning could be applied on a larger dataset to find interesting relationships in the data and create a classifier to potentially rate a web UI as "easy", "medium" or "hard" to use, based on mouse activity. Machine learning has been applied previously for application in web usability (Sahi, 2018).

References

Allison, B. Z., & Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological Psychology*, *77*(3), 277–283. doi: 10.1016/j.biopsycho.2007.10.014

Alsumait, A., & Al-Osaimi, A. (2009). Usability heuristics evaluation for child e-learning applications. In *Proceedings of the 11th international conference on information integration and web-based applications amp; services* (p. 425–430). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1806338.1806417` doi: 10.1145/1806338.1806417

Amelio, A., Draganov, I. R., Janković, R., & Tanikić, D. (2019). Analysis of usability for the dice CAPTCHA. *Information (Switzerland)*, *10*(7), 1–18. doi: 10.3390/INFO10070221

Arroyo, E., Selker, T., & Wei, W. (2006). Usability tool for analysis of web designs using mouse tracks. In *Chi '06 extended abstracts on human factors in computing systems* (p. 484–489). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1125451.1125557` doi: 10.1145/1125451.1125557

Atterer, R., Wnuk, M., & Schmidt, A. (2006a). Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction. In *Proceedings of the 15th international conference on world wide web* (p. 203–212). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1135777.1135811` doi: 10.1145/1135777.1135811

Atterer, R., Wnuk, M., & Schmidt, A. (2006b). Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction. *Proceedings of the 15th International Conference on World Wide Web*, 203–212. doi: 10.1145/1135777.1135811

Aviz, I. L., Souza, K. E., Ribeiro, E., De Mello, H., & Seruffo, M. C. d. R. (2019). Comparative study of user experience evaluation techniques based on mouse and gaze tracking. *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019*, 53–56. doi: 10.1145/3323503.3360623

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*(1), 36–47. Retrieved from `http://dx.doi.org/10.1016/j.neuroimage.2011.06.023` doi: 10.1016/j.neuroimage.2011.06.023

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400. doi: https://doi.org/10.1016/j.learninstruc.2006.09.001

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. *International Journal of Psychophysiology*, *16*(1), 57–68. doi: 10.1016/0167-8760(94)90042-6

Bakaev, M., Khvorostov, V., Heil, S., & Gaedke, M. (2017). Evaluation of user-subjective web interface similarity with Kansei engineering-based ANN. *Proceedings - 2017 IEEE 25th International Requirements Engineering Conference Workshops, REW 2017*, 125–131. doi: 10.1109/REW.2017.13

Bansal, H., & Khan, R. (2018). A Review Paper on Human Computer Interaction. *International Journal of Advanced Research in Computer Science and Software Engineering*, *8*(4), 53. doi: 10.23956/ijarcsse.v8i4.630

Barnes, J. N. (2015). Exercise, cognitive function, and aging. *Advances in physiology education*, *39*(2), 55–62. doi: 10.1152/advan.00101.2014

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (p. 648–657). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3351095.3375624 doi: 10.1145/3351095.3375624

Biondi, F. N., Cacanindin, A., Douglas, C., & Cort, J. (2021). Overloaded and at Work: Investigating the Effect of Cognitive Workload on Assembly Task Performance. *Human Factors*, *63*(5), 813–820. doi: 10.1177/0018720820929928

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989). Traditional and raw task load index (tlx) correlations: Are paired comparisons necessary?.

Cain, B. (2007). A Review of the Mental Workload Literature. *Defence research and development Toronto (Canada)*(1998), 4–1–4–34. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports*, *7*(1), 5222. Retrieved from https://doi.org/10.1038/s41598-017-05378-x doi: 10.1038/s41598-017-05378-x

Causse, M., Lepron, E., Mandrick, K., Peysakhovich, V., Berry, I., Callan, D., & Rémy, F. (2022, feb). Facing successfully high mental workload and stressors: An fMRI study. *Human brain mapping*, *43*(3), 1011–1031. doi: 10.1002/hbm.25703

Cha, G. E., & Min, B. C. (2022). *Correlation between Unconscious Mouse Actions and Human CognitiveWorkload* (Vol. 1) (No. 1). Association for Computing Machinery. doi: 10.1145/3491101.3519658

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M. A., Taib, R., ... Wang, Y. (2012). Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems*, *2*(4). doi: 10.1145/2395123.2395127

Council, N. R. (1993). *Workload Transition: Implications for Individual and Team Performance* (B. M. Huey & C. D. Wickens, Eds.). Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/catalog/2045/workload-transition-implications-for-individual-and-team-performance doi: 10.17226/2045

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., ... Jacucci, G. (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*, *9*(3-4), 151–308. doi: 10.1561/1100000065

Curry, R., Jex, H., Levison, W., & Stassen, H. (1979). Final Report of Control Engineering Group. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 235–252). Boston, MA: Springer US. doi: 10.1007/978-1-4757-0884-4_13

Delliaux, S., Delaforge, A., Deharo, J.-C., & Chaumet, G. (2019). Mental Workload Alters Heart Rate Variability, Lowering Non-linear Dynamics. *Frontiers in Physiology*, *10*. Retrieved from https://www.frontiersin.org/articles/10.3389/fphys.2019.00565 doi: 10.3389/fphys.2019.00565

Dennis, S., Bruza, P., & McArthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, *53*(2), 120–133. doi: 10.1002/asi.10015

Egeth, H., & Kahneman, D. (1975). *Attention and Effort* (Vol. 88) (No. 2). doi: 10.2307/1421603

Eggemeier, F., Wilson, G., Kramer, A., & Damos, D. (1993, 07). Workload assessment in multi-task environments. In (p. 207-216). doi: 10.1201/9781003069447-12

Eggemeier, F. T., Shingledecker, C. A., & Crabtree, M. S. (1985). Workload Measurement in System Design and Evaluation. *Proceedings of the Human Factors So-*

ciety Annual Meeting, *29*(3), 215–219. Retrieved from https://doi.org/10.1177/154193128502900302 doi: 10.1177/154193128502900302

Fan, X., Zhao, C., Zhang, X., Luo, H., & Zhang, W. (2020). Assessment of mental workload based on multi-physiological signals. *Technology and Health Care*, *28*(S1), S67–S80. doi: 10.3233/THC-209008

Fielding, R. T., & Taylor, R. N. (2000). *Architectural styles and the design of network-based software architectures* (Unpublished doctoral dissertation). (AAI9980887)

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241. doi: 10.3758/BRM.42.1.226

Freire, L. L., Arezes, P. M., & Campos, J. C. (2012). A literature review about usability evaluation methods for e-learning platforms. *Work*, *41*, 1038–1044. doi: 10.3233/WOR-2012-0281-1038

Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology*, *83*(3), 269–275. Retrieved from http://dx.doi.org/10.1016/j.ijpsycho.2011.09.023 doi: 10.1016/j.ijpsycho.2011.09.023

Garavan, H., Ross, T. J., Murphy, K., Roche, R. A. P., & Stein, E. A. (2002, dec). Dissociable executive functions in the dynamic control of behavior: inhibition, error detection, and correction. *NeuroImage*, *17*(4), 1820–1829. doi: 10.1006/nimg.2002.1326

Gawron, V. (2019). *Human performance, workload, and situational awareness measures handbook, third edition - 2-volume set.* doi: 10.1201/9780429019562

Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 113–131. Retrieved from https://doi.org/10.1080/14639220210159717 doi: 10.1080/14639220210159717

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In *Handbook of perception and human performance, vol. 2: Cognitive processes and performance.* (pp. 1–49). Oxford, England: John Wiley Sons.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. doi: https://doi.org/10.1016/j.cognition.2014.11.026

Grimes, G. M., & Valacich, J. S. (2015). Mind over mouse: The effect of cognitive load on mouse movement behavior. *2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015*(December 2015).

Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *AI Magazine*, *30*(4), 48–57. doi: 10.1609/aimag.v30i4.2271

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society*, 904–908. doi: 10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Toward Development Of a Cohesive Model Of Workload. *The Journal Of San Jose State University*, *52*(Human Mental Workload), 381.

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. Retrieved from `https://doi.org/10.1177/001872089303500401` doi: 10.1177/001872089303500401

Holland, M. K., & Tarlow, G. (1972). Blinking and mental load. *Psychological reports*, *31*(1), 119–127. doi: 10.2466/pr0.1972.31.1.119

Inzana, C. M., Driskell, J. E., Salas, E., & Johnston, J. H. (1996). Effects of preparatory information on enhancing performance under stress. *Journal of Applied Psychology*, *81*(4), 429–435. doi: 10.1037/0021-9010.81.4.429

ISO-9241-11. (2018). *Iso - iso 9241-11:2018 - ergonomics of human-system interaction — part 11: Usability: Definitions and concepts.* `https://www.iso.org/standard/63500.html`. ((Accessed on 02/27/2022))

Javaid, M. A. (2013). Review and Analysis of Human Computer Interaction (HCI) Principles. *SSRN Electronic Journal*(1998). doi: 10.2139/ssrn.2333608

Jazayeri, M. (2007). Some trends in Web application development. *FoSE 2007: Future of Software Engineering*, 199–213. doi: 10.1109/FOSE.2007.26

Jeffri, N. F. S., & Awang Rambli, D. R. (2021). A review of augmented reality systems and their effects on mental workload and task performance. *Heliyon*, *7*(3), e06277. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2405844021003820` doi: https://doi.org/10.1016/j.heliyon.2021.e06277

Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. In *Human mental workload.* (pp. 5–39). Oxford, England: North-Holland. doi: 10.1016/S0166-4115(08)62381-X

Karray, F., Alemzadeh, M., Abou Saleh, J., & Nours Arab, M. (2008). Human-Computer Interaction: Overview on State of the Art. *International Journal on Smart Sensing and Intelligent Systems*, *1*(1), 137–159. doi: 10.21307/ijssis-2017-283

Kasprowski, P., & Harezlak, K. (2018). Fusion of eye movement and mouse dynamics for reliable behavioral biometrics. *Pattern Analysis and Applications*, *21*(1), 91–103. Retrieved from `https://doi.org/10.1007/s10044-016-0568-5` doi: 10.1007/s10044-016-0568-5

Kaur, S., Kaur, K., & Kaur, P. (2016). Analysis of website usability evaluation methods. In *2016 3rd international conference on computing for sustainable global development (indiacom)* (p. 1043-1046).

Kim, J. Y., & Ji, Y. G. (2013). A Comparison of Subjective Mental Workload Measures in Driving Contexts. *Journal of the Ergonomics Society of Korea*, *32*(2), 167–177. doi: 10.5143/jesk.2013.32.2.167

Knoeferle, K., Woods, A., Käppler, F., & Spence, C. (2015). That sounds sweet: Using cross-modal correspondences to communicate gustatory attributes. *Psychology and Marketing*, *32*, 107-120. doi: 10.1002/mar.20766

Kucirkova, N., Evertsen-Stanghelle, C., Studsrød, I., Jensen, I. B., & Størksen, I. (2020). Lessons for child–computer interaction studies following the research challenges during the Covid-19 pandemic. *International Journal of Child-Computer Interaction*, *26*, 100203. Retrieved from `https://doi.org/10.1016/j.ijcci.2020.100203` doi: 10.1016/j.ijcci.2020.100203

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of User Engagement. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User modeling, adaptation, and personalization* (pp. 164–175). Berlin, Heidelberg: Springer Berlin Heidelberg.

Leite, J. C. S. d. P., & Cappelli, C. (2010). Software Transparency. *Business Information Systems Engineering*, *2*(3), 127–139. Retrieved from `https://doi.org/10.1007/s12599-010-0102-z` doi: 10.1007/s12599-010-0102-z

Li, Y., Kumar, R., Lasecki, W. S., & Hilliges, O. (2020). Artificial intelligence for hci: A modern approach. In *Extended abstracts of the 2020 chi conference on human factors in computing systems* (p. 1–8). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3334480.3375147` doi: 10.1145/3334480.3375147

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4-11. doi: 10.1109/THFE2.1960.4503259

Liu, J., Wong, C. K., & Hui, K. K. (2003). An Adaptive User Interface Based on Personalized Learning. *IEEE Intelligent Systems*, *18*(2), 52–57. doi: 10.1109/MIS.2003.1193657

Longo, L. (2014). *Formalising human mental workload as a defeasible computational concept* (Unpublished doctoral dissertation).

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, *13*(8), 1–36. Retrieved from `https://doi.org/10.1371/journal.pone.0199661` doi: 10.1371/journal.pone.0199661

Longo, L., & Dondio, P. (2016). On the relationship between perception of usability and subjective mental workload of web interfaces. *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, *1*(December), 345–352. doi: 10.1109/WI-IAT.2015.157

Longo, L., Rusconi, F., Noce, L., & Barrett, S. (2012). The importance of Human Mental Workload in Web design. *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies*, 403–409. doi: 10.5220/0003960204030409

Lottridge, D. (2020). Insights from videochat research in the context of covid-19. *Interactions*, *27*(4), 9–10. Retrieved from `https://doi.org/10.1145/3406104` doi: 10.1145/3406104

Lysaght, R. J., Hill, S. G., Dick, a. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., ... Wherry, R. J. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies. *United States Army Research Institute for the Behavioral Sciences, Technical Report*, *851*, 903–986.

MacCorquodale, K., & Meehl, P. E. (1948). *On a distinction between hypothetical constructs and intervening variables.* (Vol. 55). US: American Psychological Association. doi: 10.1037/h0056029

Maslov, I., & Nikou, S. (2020). Usability and UX of Learning Management Systems: An Eye- Tracking Approach. *Proceedings - 2020 IEEE International Conference on Engineering, Technology and Innovation, ICE/ITMC 2020*. doi: 10.1109/ICE/ITMC49519.2020.9198333

matplotlib. (2022). *Matplotlib.* Retrieved 2022-07-25, from https://matplotlib.org/

McFarland, J. (2016). Mental workload measurement for competitive video games. *University of Lousville*, 102. Retrieved from http://ir.library.louisville.edu/etd

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record*(2138), 6–12. doi: 10.3141/2138-02

Meta. (2022). *ReactJS.* Retrieved 2022-07-12, from https://reactjs.org/

Miinitab. (2022). *Pearson and Spearman Correlation Methods.* Retrieved 2022-07-23, from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

Miller, G. A. (1956). *The magical number seven, plus or minus two: Some limits on our capacity for processing information.* (Vol. 63) (No. 2). US: American Psychological Association. doi: 10.1037/h0043158

Mock, P., Gerjets, P., Tibus, M., Trautwein, U., Möller, K., & Rosenstiel, W. (2016). Using touchscreen interaction data to predict cognitive workload. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 349–356. doi: 10.1145/2993148.2993202

Mohammadian, M., Parsaei, H., Mokarami, H., & Kazemi, R. (2022). Cognitive demands and mental workload: A filed study of the mining control room operators. *Heliyon*, *8*(2), e08860. Retrieved from https://doi.org/10.1016/j.heliyon.2022.e08860 doi: 10.1016/j.heliyon.2022.e08860

Monk, T. H., & Leng, V. C. (1982). Time of day effects in simple repetitive tasks: Some possible mechanisms. *Acta Psychologica*, *51*(3), 207–221. doi: 10.1016/0001-6918(82)90035-X

Monod, H., & Kapitaniak, B. (2003). *Ergonomics.* Retrieved 2022-07-03, from https://www.unitheque.com/ergonomie/abreges/elsevier-masson/Livre/2466

Mozilla. (2022a). *Click Event.* Retrieved 2022-07-03, from https://developer.mozilla.org/en-US/docs/Web/API/Element/click_event

Mozilla. (2022b). *Event Reference.* Retrieved 2022-07-15, from https://developer.mozilla.org/en-US/docs/Web/Events

Mozilla. (2022c). *MouseEvent.* Retrieved 2022-07-03, from https://developer.mozilla.org/en-US/docs/Web/API/MouseEvent

Mozilla. (2022d). *Mousemove Event.* Retrieved 2022-07-03, from https://developer.mozilla.org/en-US/docs/Web/API/Element/mousemove_event

Mozilla. (2022e). *Mouseover Event.* Retrieved 2022-07-03, from https://developer.mozilla.org/en-US/docs/Web/API/Element/mouseover_event

Muñoz-de Escalona, E., Cañas, J. J., Leva, C., & Longo, L. (2020). Task Demand Transition Peak Point Effects on Mental Workload Measures Divergence. *Communications in Computer and Information Science*, *1318*, 207–226. doi: 10.1007/978-3-030-62302-9_13

NASA. (2022). *NASA Task Load Index.* Retrieved 2022-07-22, from https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 152–158). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/191666.191729 doi: 10.1145/191666.191729

Norman, G. (2013). Working memory and mental workload. *Advances in Health Sciences Education*, *18*(2), 163–165. doi: 10.1007/s10459-013-9451-y

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of perception and human performance, vol. 2: Cognitive processes and performance.* (pp. 1–49). Oxford, England: John Wiley Sons.

O'Donnell, T., & Eggemeier, R. D. (1986). *Workload assessment methodology. Handbook of Perception and Human Performance* (Vol. 2).

Özsu, M. T. (2016). Client-Server Architecture. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 1–3). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4899-7993-3_664-2 doi: 10.1007/978-1-4899-7993-3_664-2

Pandas. (2022a). *Pandas.* Retrieved 2022-07-25, from https://pandas.pydata.org/

Pandas. (2022b). *Pandas Scatter Matrix.* Retrieved 2022-07-25, from https://pandas.pydata.org/docs/reference/api/pandas.plotting.scatter_matrix.html

Pantic, M., Nijholt, A., Pentland, A., & Huanag, T. S. (2008). Human-Centred Intelligent Human-Computer Interaction (HCI2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, *1*(2), 168–187. doi: 10.1504/ijaacs.2008.019799

Papesh, M. H., & Goldinger, S. D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychonomic Bulletin and Review*, *19*(5), 906–913. doi: 10.3758/s13423-012-0281-3

Pawar, U., O'Shea, D., Rea, S., & O'Reilly, R. (2020). Incorporating explainable artificial intelligence (XAI) to aid the understanding of machine learning in the healthcare domain. *CEUR Workshop Proceedings*, *2771*, 169–180.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, *5*(DEC), 1–11. doi: 10.3389/fpsyg.2014.01344

Paz, F., Paz, F. A., Pow-Sang, J. A., & Collantes, L. (2014). Usability heuristics for transactional web sites. In *2014 11th international conference on information technology: New generations* (p. 627-628). doi: 10.1109/ITNG.2014.81

Pimenta, A., Carneiro, D., Novais, P., & Neves, J. (2013). Monitoring mental fatigue through the analysis of keyboard and mouse interaction patterns. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8073 LNAI*(September), 222–231. doi: 10.1007/978-3-642-40846-5_23

Pimenta, A., Gonçalves, S., Carneiro, D., Fde-Riverola, F., Neves, J., & Novais, P. (2015). Mental workload management as a tool in e-learning scenarios. *PECCS 2015 - 5th International Conference on Pervasive and Embedded Computing and Communication Systems, Proceedings*, 25–32. doi: 10.5220/0005237700250032

Pourteimour, S., Yaghmaei, S., & Babamohamadi, H. (2021). The relationship between mental workload and job performance among Iranian nurses providing care to COVID-19 patients: A cross-sectional study. *Journal of Nursing Management*, *29*(6), 1723–1732. doi: 10.1111/jonm.13305

Pulat, B. M. (1997). *Fundamentals of industrial ergonomics*. Waveland Press.

Quiñones, D., & Rusu, C. (2017). How to develop usability heuristics: A systematic literature review. *Computer Standards and Interfaces*, *53*, 89–122. Retrieved from `http://dx.doi.org/10.1016/j.csi.2017.03.009` doi: 10.1016/j.csi.2017.03.009

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. In *Human mental workload.* (pp. 185–218). Oxford, England: North-Holland. doi: 10.1016/S0166-4115(08)62387-0

Reips, U.-D. (2000). Chapter 4 - The Web Experiment Method: Advantages, Disadvantages, and Solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89–117). San Diego: Academic Press. doi: https://doi.org/10.1016/B978-012099980-4/50005-8

Rheem, H., Verma, V., & Becker, D. V. (2018). Use of mouse-tracking method to measure cognitive load. *Proceedings of the Human Factors and Ergonomics Society*, *3*, 1982–1986. doi: 10.1177/1541931218621449

Richards, R. (2006). Representational State Transfer (REST). In *Pro php xml and web services* (pp. 633–672). Berkeley, CA: Apress. Retrieved from https://doi.org/10.1007/978-1-4302-0139-7_17 doi: 10.1007/978-1-4302-0139-7_17

Rodriguez, M. G. (2002). Automatic data-gathering agents for remote navigability testing. *IEEE Software*, *19*(6), 78–85. doi: 10.1109/MS.2002.1049396

Romero, J. F. (2017). An Investigation of the Correlation Between Mental Workload and Web User ' s Interaction An investigation of the correlation between Mental Workload and Web User ' s Interaction Joaquim Filipe Braga Simões Romero. Retrieved from https://arrow.dit.ie/scschcomdis

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. doi: 10.1109/ACCESS.2020.2976199

Roy, S., Pattnaik, P. K., & Mall, R. (2014). A quantitative approach to evaluate usability of academic websites based on human perception. *Egyptian Informatics Journal*, *15*(3), 159–167. Retrieved from http://dx.doi.org/10.1016/j.eij.2014.08.002 doi: 10.1016/j.eij.2014.08.002

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. doi: 10.1111/j.1464-0597.2004.00161.x

Sahi, G. (2018). Performance evaluation of artificial neural network for usability assessment of e-commerce websites. In *2018 3rd international conference for convergence in technology (i2ct)* (p. 1-6). doi: 10.1109/I2CT.2018.8529613

Saket, B., Endert, A., & Stasko, J. (2016). Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the sixth*

workshop on beyond time and errors on novel evaluation methods for visualization (p. 133–142). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2993901.2993903 doi: 10.1145/2993901.2993903

Schall, A. (2014). 6 - Information Architecture and Web Navigation. In J. Romano Bergstrom & A. J. Schall (Eds.), *Eye tracking in user experience design* (pp. 139–162). Boston: Morgan Kaufmann. doi: https://doi.org/10.1016/B978-0-12-408138-3.00006-6

Schewe, F., & Vollrath, M. (2020). Ecological interface design effectively reduces cognitive workload – The example of HMIs for speed control. *Transportation Research Part F: Traffic Psychology and Behaviour*, *72*, 155–170. Retrieved from https://doi.org/10.1016/j.trf.2020.05.009 doi: 10.1016/j.trf.2020.05.009

Scholtz, J., Laskowski, S., & Downey, L. (1998). Developing usability tools and techniques for designing and testing web sites. *... on Human Factors the Web*, *I*, 1–10. Retrieved from http://zing.ncsl.nist.gov/hfweb/att4/proceedings/proceedings.en.html%5Cnhttp://zing.ncsl.nist.gov/hfweb/att4/proceedings/scholtz/

SciPy. (2022). *SciPy*. Retrieved 2022-07-25, from https://docs.scipy.org/doc/scipy/index.html

Serra, G., Falco, F. D., Maggi, P., Forsi, R., Cocco, A., Gaudino, G., ... Nocera, F. D. (2019). The role of mental workload in determining the relation between website complexity and usability : an eye-tracking study. , *4959*(8).

Shackel, B. (1969). Man-Computer Interaction—The Contribution of the Human Sciences. *Ergonomics*, *12*(4), 485–499. doi: 10.1080/00140136908931075

Shneiderman, B. (1988). We can design better user interfaces: A review of human-computer interaction styles. *Ergonomics*, *31*(5), 699–710. doi: 10.1080/00140138808966713

Smit, A. S., Eling, P. A., Hopman, M. T., & Coenen, A. M. (2005). Mental and physical effort affect vigilance differently. *International Journal of Psychophysiology*, *57*(3), 211–217. doi: 10.1016/j.ijpsycho.2005.02.001

Souza, K. E., Seruffo, M. C., De Mello, H. D., Souza, D. D. S., & Vellasco, M. M. (2019). User Experience Evaluation Using Mouse Tracking and Artificial Intelligence. *IEEE Access*, *7*, 96506–96515. doi: 10.1109/ACCESS.2019.2927860

Statista. (2021). *Web Framework Popularity.* Retrieved 2022-07-12, from `https://www.statista.com/statistics/1124699/worldwide-developer-survey-most-used-frameworks-web/`

Sullivan, K. J., Griswold, W. G., Cai, Y., & Hallen, B. (2001). The structure and value of modularity in software design. *Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering*, 99–108. doi: 10.1145/503209.503224

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi: 10.1016/0364-0213(88)90023-7

Tan, W.-s., Liu, D., & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics*, *39*(4), 621–627. doi: https://doi.org/10.1016/j.ergon.2008.02.012

Thomas, M. O., Onyimbo, B. A., & Logeswaran, R. (2016). Usability evaluation criteria for internet of things. *International Journal of Information Technology and Computer Science*, *8*, 10-18.

Tracy, J. P. (2007). *Measuring Cognitive Load to Test the Usability of Websites.* University of Memphis. Retrieved from `https://books.google.ie/books?id=IhPBtgAACAAJ`

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. doi: 10.1080/00140139608964470

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. In *Handbook of human factors and ergonomics, 3rd ed.* (pp. 243–268). Tsang, Pamela S.: Department of Psychology, Wright State University, 3640 Colonel Glenn Highway, Dayton, OH, US, pamela.tsang@wright.edu: John Wiley Sons, Inc. doi: 10.1002/0470048204.ch9

van Steenbergen, H., & Bocanegra, B. R. (2016). Promises and pitfalls of Web-based experimentation in the advance of replicable psychological science: A reply to Plant (2015). *Behavior Research Methods*, *48*(4), 1713–1717. doi: 10.3758/s13428-015-0677-x

Vargo, D., Zhu, L., Benwell, B., & Yan, Z. (2021). Digital technology use during COVID-19 pandemic: A rapid review. *Human Behavior and Emerging Technologies*, *3*(1), 13–24. doi: 10.1002/hbe2.242

Vidulich, M. A. (1988). The Cognitive Psychology of Subjective Mental Workload. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 219–229). North-Holland. doi: 10.1016/S0166-4115(08)62388-2

Vincent, A., Craik, F. I. M., & Furedy, J. J. (1996). Relations among memory performance, mental workload and cardiovascular responses. *International Journal of Psychophysiology*, *23*(3), 181–198. doi: https://doi.org/10.1016/S0167-8760(96)00058-X

Ward, R. D., & Marsden, P. M. (2003). Physiological responses to different WEB page designs. *International Journal of Human Computer Studies*, *59*(1-2), 199–212. doi: 10.1016/S1071-5819(03)00019-3

Wickens, C. (2008, 07). Multiple resources and mental workload. *Human factors*, *50*, 449-55. doi: 10.1518/001872008X288394

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (1992). *Engineering psychology and human performance*. Routledge.

Williges, R. C., & Wierwille, W. W. (1979). Behavioral Measures of Aircrew Mental Workload. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *21*(5), 549–574. doi: 10.1177/001872087902100503

Wilson, J. R., & Sharples, S. (2015). *Evaluation of human work*. CRC press.

Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human-computer interaction. *Artificial Intelligence*, *170*(18), 1256–1258. doi: 10.1016/j.artint.2006.10.011

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *2015*(7). doi: 10.7717/peerj.1058

Xie, B., & Salvendy, G. (2000a, 09). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics*, *4*, 213-242. doi: 10.1207/S15327566IJCE0403_3

Xie, B., & Salvendy, G. (2000b). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work & Stress*, *14*(1), 74–99. Retrieved from `https://doi.org/10.1080/026783700417249` doi: 10.1080/026783700417249

Yamagishi, M., Kobayashi, T., Kobayashi, T., Nagami, M., Shimazu, A., & Kageyama, T. (2007). Effect of web-based assertion training for stress management of Japanese nurses. *Journal of Nursing Management*, *15*(6), 603–607. doi: 10.1111/j.1365-2834.2007.00739.x

Yang, H., Wei, H., He, X., Yan, Y., & Liu, X. (2021). User Experience Evaluation of Cross-Channel Consumption: Based on Grounded Theory and Neural Network. *Wireless Communications and Mobile Computing*, *2021*. doi: 10.1155/2021/1133414

Yi, W., Qiu, S., Fan, X., Zhang, L., & Ming, D. (2022). Evaluation of Mental Workload Associated with Time Pressure in Rapid Serial Visual Presentation Tasks.

IEEE Transactions on Cognitive and Developmental Systems, *14*(2), 608–616. doi: 10.1109/TCDS.2021.3061564

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. doi: 10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194. doi: 10.1080/14639220210123789

Zacks, R. T. (2004). Psychology and Aging: Editorial. *Psychology and Aging*, *19*(1), 3. doi: 10.1037/0882-7974.19.1.3

Zare, S., Hasheminezhad, N., Dehesh, T., Hasanvand, D., Ahmadi, S., & Hemmatjo, R. (2016). The relationship between mental workload and prevalence of musculoskeletal disorders among welders of Tehran heavy metal structures company in 2016. *Journal of Biology and Today's World*, *5*(12), 218–223. doi: 10.15412/J.JBTW.01051203

Zöllner, M., Huber, S., Jetter, H.-C., Reiterer, H., Campos, P., Graham, N., ... Winckler, M. (2011). Human-Computer Interaction – INTERACT 2011. , *6949*, 584–587. Retrieved from `http://www.springerlink.com/content/wp8642lr28800747/` doi: 10.1007/978-3-642-23768-3

Appendix A

Additional Content

A.1 Dataset

Table A.1: Trial experimental results overview. Columns: *task* = task name, *difficulty* = trial difficulty_level, *NASA* = MWL value calculated from NASA-TLX, *WP* = MWL value calculated from WP, *clicks* = number of mouse clicks, *tot. mouse* = total mouse distance moved (px), *mouse_vel* = average mouse velocity (px/s), *tot. hover* = total mouse hover time (s), *cor. freq* = percentage of successful trial answers, *answer/min* = trial answer submissions per minute.

task	difficulty	NASA	WP	clicks	tot. mouse	mouse_vel	tot. hover	cor. freq	answer/min
maths	easy	31	1.02	17	30229	530.69	24.1	0.94	0.28
maths	hard	35.8	0.79	7	15853	272.43	11.51	0.86	0.12
maths	easy	45	3.3	21	61999	1074.04	23.74	0.95	0.37
maths	hard	64	2.35	8	27008	477.66	29.26	0.75	0.13
maths	easy	34	1.1	46	113964	1027.14	58.39	0.96	0.77
block	medium	36	1.45	28	14268	243.46	20.87	0.88	0.13
block	easy	41	1.15	3	12665	217.81	18.47	0.78	0.15
block	hard	32.6	1.2	29	23200	455.26	43.43	0.5	0.10
maths	easy	64.8	3.87	22	41251	764.33	29.17	0.86	0.37
maths	hard	59.8	2.97	2	12652	342.94	16.13	0.67	0.05
block	medium	37.6	3.4	36	8624	166.53	19.95	0.9	0.17
block	hard	62	2.82	33	10844	224.32	19.17	0.67	0.10
block	easy	39.6	2.52	38	11057	191.15	17.04	0.8	0.17

task	difficulty	NASA	WP	clicks	tot. mouse	mouse_vel	tot. hover	cor. freq	answer/min
maths	hard	57.2	1.3	5	16115	314.18	29.34	0.83	0.10
maths	easy	40.4	1.52	13	31066	566.78	28.31	0.93	0.23
block	easy	23.8	1.93	23	4725	80.21	25.79	0.33	0.10
block	hard	60.8	2.2	26	12207	263.79	31.36	0.17	0.10
block	med	26	1.79	33	7539	130.41	18.53	0.88	0.13
maths	easy	56.6	3.89	26	58453	1031.06	24.16	0.96	0.45
maths	hard	68.4	3.16	4	32126	560.61	19.3	0.8	0.08
block	med	47.6	3.43	32	10243	192.58	18.35	0.5	0.17
block	hard	48.8	3.27	36	13454	300.83	18.16	0.43	0.12
block	easy	55	3.8	41	6536	155.92	16.69	0.82	0.18
maths	easy	38.8	0.74	23	40254	701.62	20.62	0.96	0.40
maths	hard	90.2	1.34	5	9663	178.83	5.61	0.83	0.10
block	hard	38.6	2.39	31	10755	200.84	40.03	0.33	0.10
block	med	37.2	0.58	33	8597	151.33	25.89	0.56	0.15
block	easy	34.4	0.92	32	9625	179.53	20.55	0.56	0.15
maths	easy	78	3.01	16	31117	525.08	19.45	0.94	0.28
maths	hard	83	3.21	5	8281	142.57	8.64	0.67	0.10
block	med	48	2.41	36	8710	156.53	15.53	0.8	0.17
block	easy	48.2	2.12	37	6749	135.6	15.67	0.7	0.17

task	difficulty	NASA	WP	clicks	tot. mouse	mouse_vel	tot. hover	cor. freq	answer/min
block	hard	78	2.81	35	10140	215.58	20.73	0.57	0.12
block	easy	32.6	2.01	33	10090	191.93	16.77	0.75	0.13
block	hard	73.4	2.03	22	9158	200.53	21.82	0.2	0.08
block	med	36	0.45	36	10829	205.58	18.99	0.8	0.17
maths	easy	57.2	3.62	18	29778	521.64	18.91	0.94	0.28
maths	hard	57.8	3.79	4	8765	151.74	23.06	0.8	0.08
block	med	63	3.32	32	9931	178.22	22.52	0.89	0.15
block	hard	74.6	3.77	33	13189	226.32	19.53	0.33	0.10
block	easy	63.4	3.3	36	8542	164.61	19.02	0.78	0.15
maths	easy	30.2	0.66	22	38067	671.12	13.13	0.84	0.32
maths	hard	31	0.39	5	22076	400.28	19.92	0.4	0.08
block	med	21	0.87	30	9728	169.42	20.73	0.63	0.13
block	hard	23.6	1.27	29	14368	246.72	28.32	0.33	0.10
block	easy	12.8	0.51	3	11326	210.34	20.51	0.88	0.13
maths	easy	64	1.82	19	55677	560.49	18.58	0.95	0.32
maths	hard	75.2	2.76	9	25810	310.46	5.97	0.9	0.17
maths	easy	63	3.01	17	28746	517.14	27.63	0.94	0.30
maths	hard	64	1.72	2	8166	144.43	38.22	0.67	0.05
block	hard	49	3.01	3	10348	197.07	27.68	0.83	0.10

task	difficulty	NASA	WP	clicks	tot. mouse	mouse_vel	tot. hover	cor. freq	answer/min
block	med	45.2	3.21	35	9223	158.39	18.24	0.78	0.15
block	easy	45	3.22	33	6662	128.45	25.14	0.89	0.15

A.2 Web Application

The web-application can be accessed at: MWL Web Application

- **Front-End**: The front-end for the web-application was developed as a single-page application in ReactJS.
- **Back-End**: The back-end for the application was implemented as a RESTful API which serves JSON data to the front-end. It was developed using Ruby-on-Rails. The data was stored in a PostgreSQL database.
- **Repository**: The codebase is stored in a private GitHub repository. It can be accessed here (contact for access).

A.3 Data Analysis

Data analysis was carried out using Python3, with Pandas, SciPy and Matplotlib libraries for statistical methods and visualisation (Pandas, 2022a; SciPy, 2022; matplotlib, 2022).