Investigating The Usability of Iconography Via Brain and Key-press Responses

A THESIS SUBMITTED TO
THE UNIVERSITY OF KENT
IN THE SUBJECT OF COMPUTER SCIENCE
FOR THE DEGREE OF
MASTER OF SCIENCE (RESEARCH)

By
Daniel Thomas-Hercules
May 2024

Abstract

The purpose of this research is to better understand the circumstances in which icons are most effective in use cases that pertain to memory. To achieve this, the project aimed to investigate the research question: “How do the pre-existing icon classifications discussed by Arledge and Nielsen impact memory performance?”. The classifications in use are Solid and Outline & Resemblance and Abstract icons. Current research on the usability of these classifications suggests there is no clear superior classification in either family regarding the speed or accuracy in which they are identified. Despite this, current research regarding both the N-Back task and electroencephalogram (EEG) indicates that these tools remain effective in providing quantitative insights on memory load and the timing of memory recall. Twelve students at the University of Kent participated in an experiment where each participant conducted separate 1-back and 2-back tests (variants of the N-Back test) for each of the four permutations for both families of classifications (Solid Resemblance, Solid Abstract, Outline Resemblance, Outline Abstract). Eight tests were conducted by each participant. During each test, the reaction times of participants (in milliseconds) and the participant’s EEG data was recorded in parallel. A 32-channel wireless brain-computer interface (BCI) was used to collect the EEG recordings. The statistical analysis conducted on the data provided by the N-Back tasks and ERPs employed the Wilcoxon signed rank test to test for significant differences between the stimuli classifications used. The comparisons made on the N-Back reaction times show that Arledge’s two icon styles Solid and Outline failed to outperform each other significantly in terms of the speed at which they were recalled. Contrary to previous work favouring resemblance icons due to their suggested superior usability over abstract icons, comparisons showed that abstract icons were recalled significantly faster than resemblance icons, this observation was made when the two semantic groups were in an outline style. Dually, a comparison made showed an instance in which abstract icons in an outline style elicited a significantly smaller amplitude compared to resemblance icons in the same style, indicating an increased memory load compared to resemblance icons. User interface design principles of previous work, would suggest that Outline Resemblance stimuli had better usability than Outline Abstract stimuli, due to a significantly lower memory load. Despite these observations, overall, the ERP analysis did not reveal any significant distinction between the memory loads of the icon classifications used. To observe significant differences between the icon classifications used, this study suggests that future work explores the impact of various dimensions on memory performance.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my utmost gratitude to my supervisor Dr Palaniappan Ramaswamy for his patience, consideration, and tutelage during this research. I would like to thank the participants of this study for volunteering their time.

I would like to thank my family and friends for their support and believing in my career aspirations. I would also like to thank the NHS for supporting the swift recovery of my health during this research.

Lastly, I would like to give special thanks to my uncle, Mr Cletus Mathurin for his support and empowerment, which continued until his passing.

List of Abbreviations or Acronyms

BCI	Brain-Computer Interface
EEG	Electroencephalography
ERP	Event Related Potential
OA	Outline Abstract Icons
OR	Outline Resemblance Icons
SA	Solid Abstract Icons
SR	Solid Resemblance Icons
WM	Working Memory

1. Introduction

Icons are defined as symbols that represent data or a process in a system or interface (Gittins, 1986, p. 519). These icons support a dialogue between a computer system and its end-user, namely in fields such as Human Computer Interaction (HCI) and User Experience Design.

Solid icons are defined as icons that are filled in, taking on a bold appearance. Outline icons are defined as icons comprised of lines that make up the icons form. Curtis Arledge in a 2014 paper investigates the memory related performance of Solid and Outline icons but does not come to a finite conclusion on which of the two is recalled faster and more accurately (Arledge, 2014).

Resemblance icons are icons that depict physical objects in the real world. Abstract icons are icons that have no inherent meaning and only have meaning by convention. Jakob Nielsen introduces Resemblance and Arbitrary “Abstract” icons in a 2014 article but does not actually conduct any studies to back their claims of which style has better memory performance or usability (Nielsen, 2014).

Professor Ben Shneiderman, popular for his fundamental research in the field of HCI, instructs people in his eight rules for interface design, to lower the short term memory load required to use an interface due to humans having a limited capacity for information processing (Shneiderman, 2016). This means that to contribute to an effective interface design, icons featured should not require a large memory load, thus icons and their associated memory loads should be explored further to see how they can comply with this rule.

In addition to this, the Nielsen Norman Group partially founded by Jakob Nielsen, agrees that icons are fast to recognise if designed well. Thus, there is potential for icons to contribute to a good user interface based on their design and how recognisable they are (Harley, 2014). This study intends to explore how iconography is memorised and recognised over different classifications linking to their meaning and form. This will serve the purpose of better understanding the circumstances in which icons are most effective in use cases that pertain to memory.

1.1 Motivation

Currently a small amount of research has been conducted to show what style and semantic has the best performance in memory, i.e., which requires the least memory load to be successfully recalled and which is the fastest recalled.

A study that further investigates the memory performance of Arledge and Nielsen's icon classifications, could provide the field of HCI with more insight into the conditions in which iconography is most effective in use cases applicable to memory. This can be accomplished by using two forms of quantitative data in the form of reaction times to measure recall speed and EEG to measure the memory load of correct recalls. For example, more research on which icons elicit different levels of memory load can aid user interface designers in lowering the short term memory load required by their user interfaces (A constraint that Shneiderman advocates for in their eight rules for designing an interface). Moreover, by knowing how different classifications of icon perform (regarding the time required to recall and the memory load they elicit), this can better inform memory performance tests used in medical and commercial applications.

1.2 Research Question & Hypothesis

In order to achieve this project's purpose of better understanding the circumstances in which icons are most effective in use cases that pertain to memory, this project will investigate the research question: "How do the pre-existing icon classifications discussed by Arledge and Nielsen impact memory performance?". This will be accomplished by using these icon classifications in conjunction with the N-Back Task, a working memory performance task, specifically 1-back and 2-back variants for varying levels of difficulty. "Memory performance" will be monitored using the reaction times (i.e. The time it takes to recall) of the N-Back task. Memory performance will also be monitored using neural activity monitored during an EEG conducted in parallel with the N-Back tasks. Based on previous literature, the following hypotheses will be explored in this work:

- Hypothesis 1 – One of Arledge's classifications (solid or outline) will perform significantly better than the other regarding the speed in which the icon is recalled.
- Hypothesis 2 – Nielsen's resemblance icons will perform better than abstract "arbitrary" icons based on their ability to be associated with real world objects

(resemblance icons will be recalled faster based on their association with the real world).

- Hypothesis 3 – The memory loads elicited by each icon classification will show a clear distinction from one another.

1.3 Research Contribution

The intended contribution of this research is to supplement the fields of HCI and User Experience Design by continuing foundational research by Arledge and Nielsen in studying the memory performance of two families of icon classification. The intention was to conclude on which icons have superior memory performance over their relative. This memory performance was judged based on the time required to recall the icon from memory and the memory load the icon elicited from participants.

Results indicate that Outline Abstract stimuli was recalled significantly faster than Outline Resemblance stimuli whilst also eliciting a higher memory load. Results also indicate that in the 1-back test, Outline Resemblance stimuli had a significantly larger max peak P300 amplitude compared to Outline Abstract stimuli, this provides evidence that OR stimuli has better usability than OA stimuli as the former has a reduced memory load compared to the latter, something that is advocated for by the aforementioned rules established by Professor Ben Shneiderman. Besides this, on a macro level, each icon classification performed equally regarding their recall times and memory load elicited.

These results highlight the following contributions:

1. In a computing perspective this contributes to the reinforcement of and argument against conclusions of previous studies respectively.
2. In a psychological perspective, this contributes to current research's understanding of these icons, this conclusion of insignificance among the icon classifications indicates that the impact on user cognition is equal across Arledge's and Nielsen's classifications. The contribution to psychology is bolstered by the fact that this study has reached its conclusion using P300, which is regarding as an important ERP component for evaluating cognitive function (e.g. attention and working memory (Zhong et al., 2019, p. 4)).
3. The data collected is a contribution to the field of iconography research as the data collected could be compared to the event and EEG data of future studies to support respective further research into iconography.

1.4 Research Outline

This thesis is structured as follows: Chapter Two provides background information on the research. Chapter Three provides a literature review on the topics discussed in the background section. Chapter Four outlines the methodology used, including the experimental procedure. Chapter Five details the data analysis conducted using the data collected. Chapter Six explains the findings from the data analysis and Chapter Seven discusses these findings and their relevance to the introduction of this thesis. Chapter Eight concludes the research.

2. Background

2.1 Introduction

This chapter gives background information relating to the different classifications of icons, EEG and the N-Back task.

2.2 Solid and Outline Icons

Solid icons are fully filled shapes or images with no visible outlines, creating a bold and visually prominent appearance. In contrast, outline icons consist of only the outer lines or contours of an icon, without any interior fill, resulting in a lighter and more subtle visual style.

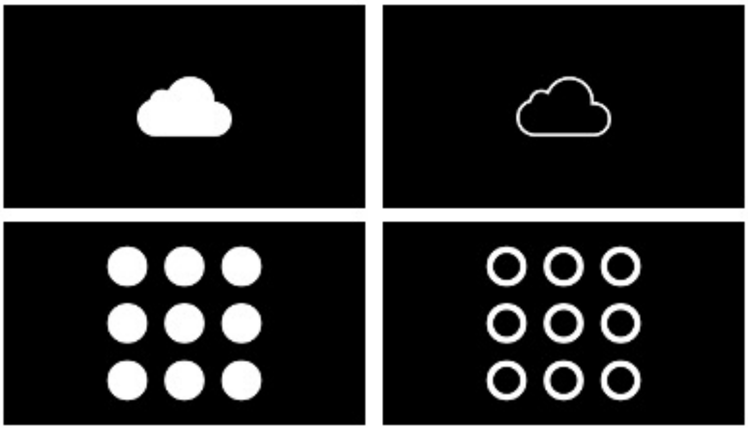

Figure 2.1: Example of solid icons (left column) & outline icons (right column)

2.3 Resemblance and Abstract Icons

Jakob Nielsen's Resemblance and Abstract icons were introduced and described by Nielsen in a 2014 article titled "Icon Classification: Resemblance, Reference and Arbitrary Icons". Resemblance icons depict physical objects in the real world. Abstract icons are described as icons that only have a meaning by convention. In other words, they are icons that do not have an inherently obvious connection to what they signify.

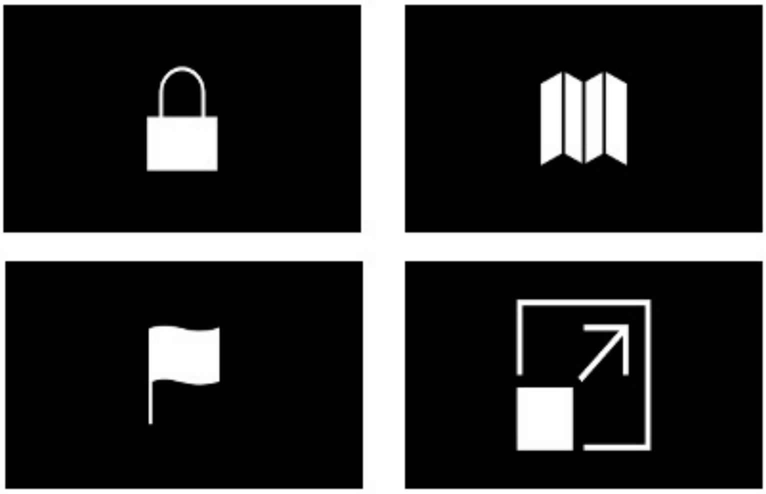

Figure 2.2: Example of resemblance icons (left column) & abstract Icons (right column)

2.4 Electroencephalogram

First explored in 1924 by physiologist Hans Berger, EEG is a neuroimaging technique used to record changes in the electrical activity within the brain. This is possible using electrodes attached to the user's scalp. This study uses a non-invasive BCI to record EEG, thus, electrodes are not inserted into the scalp but instead placed on the user's scalp.

The data recorded from each electrode represents electrical activity in a specific region of the brain; however, current technology does not have the capability to measure the output of individual neurons. The electrical activity is recorded with an accuracy range of microVolts (uV) and is presented in a computer system as a wave interpretable based on their time and frequency information. Collecting useful EEG data requires a specific preparation procedure. This may involve applying electrode gel for wet electrodes to increase conductance or using dry electrodes, which do not require gel.

2.5 Event Related Potentials

ERPs are event-related changes in the neural activity recorded by EEG. This data is often time-locked to sensory and cognitive events (Sur & Sinha, 2009), thus as a result, ERPs help capture the neural activity relating to sensory and cognitive processes. In general, the steps involved in pre-processing the ERP are as follows:

- Importing data into a development environment that can conduct the necessary data analysis e.g. MATLAB or Python, as well as libraries that are purpose built for EEG analysis such as EEGLAB and ERPLAB.
- Filtering: Application of bandpass filters to isolate a frequency range (e.g. 0.2 – 10 Hz) of interest and remove artefacts and high frequency noise.
- Baseline correction: Removing or adjusting the baseline signal level to enhance the visibility of neurological activity.
- Epoching: Segmenting the EEG data into epochs (segments of time in EEG), then averaging the neural activity across all appropriate segments. Segments included for the averaging usually all have a characteristic in common such as a specific participant or event that occurred in the segment.

2.6 The N-Back Task

The N-Back task is a continuous task that is used in research concerning psychology and cognitive neuroscience. Introduced in 1958 by Wayne Kirchner, the N-Back task involves a participant being presented with a sequence of stimuli. During this presentation, the participant is given the task of deciding whether the current stimulus being shown matches the n^{th} stimulus shown earlier (Science Direct, 2017).

For example, the N-Back task has several popular variants including:

- 0-Back: User is told to decide whether the current stimulus matches a specific stimulus they were told to look for.
- 1-Back: User has to decide whether the current stimulus shown matches the stimulus shown previous to the current stimulus (hence the name 1-Back).
- 2-Back: User has to decide whether the current stimulus shown matches the stimulus two stimuli ago (hence the name 2-Back).

The N-back task has been widely regarded as a valid measure for working memory performance, particularly in studies involving functional neuroimaging. Owen et al. (2005) states that the N-Back task is a popular experimental paradigm for functional neuroimaging studies that investigate working memory. This popularity is due to the N-Back tasks ability to provide insights on cognitive load and memory performance, a capability that is essential to this study, its motivation and its hypotheses.

More recent studies validate the N-Back task's suitability for investigating memory load using EEG, a functional neuroimaging technique. Wang et al. (2016) affirms that the N-Back task is widely used to measure working memory load (p.425) as well as demonstrates that the N-back task, combined with a brain computer interface that records EEG, is effective in measuring memory load. This is highly relevant to the current study as this current study intends to use EEG to measure the memory load of participants during the N-Back task.

Meule (2017) supports the validity of the N-Back task by stating that the N-Back task is widely used in the measuring of working memory (p. 1), while also suggesting that combining data such as the reaction times and accuracy of the N-Back task with an EEG analysis, will provide deeper insight into the N-Back task's validity as a working memory measure. The consistency between Owen (2005), Wang (2016) and Meule (2017) validates the N-Back task as a reliable method of measuring memory performance, justifying the use of the N-Back task in the current study. Furthermore, the N-Back task's validated ability to measure memory performance, clearly aligns with the research question which aims to investigate the memory performance of different stimuli, this further demonstrates why the N-Back task is appropriate for use in this current study.

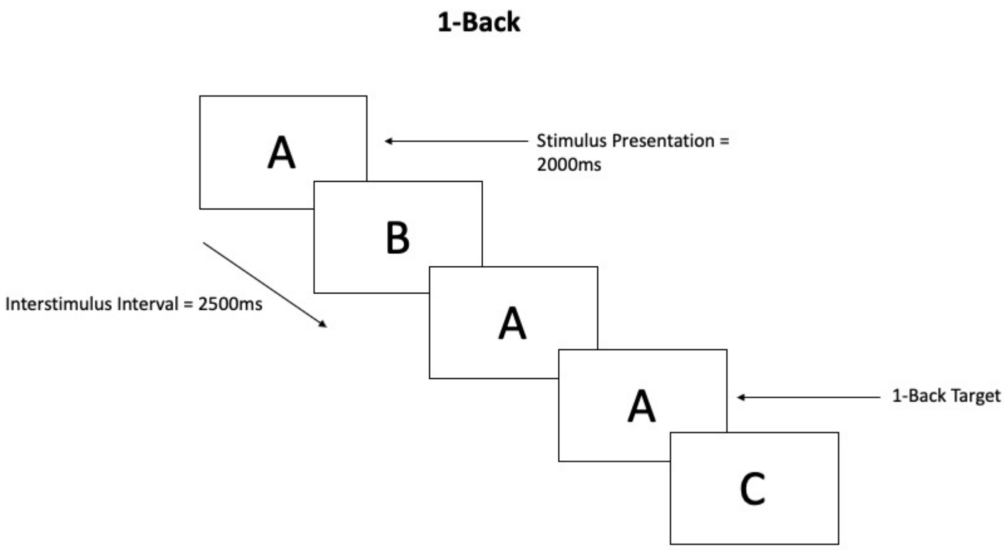

Figure 2.3: Example of a 1-Back Task

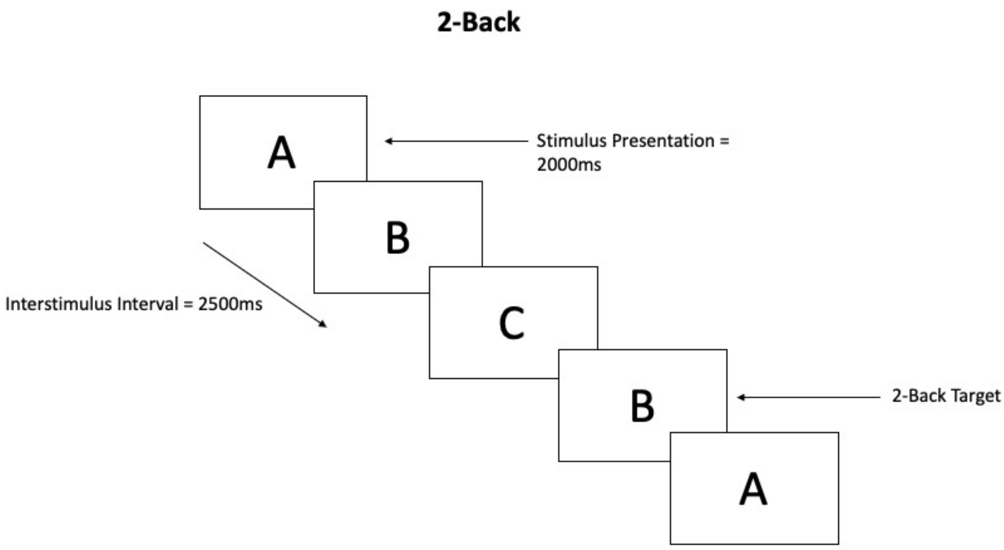

Figure 2.4: Example of a 2-Back Task

3. Literature Review

3.1 Introduction

This chapter discusses current research relating to icon recognition, the N-Back task, N-Back timing, ERPs and EEG. This section elaborates on the relevance these concepts have on the research aim and motivation. This literature review's inclusion in this thesis is important as it justifies both the decisions made in the project's methodology of experimentation as well as justifies the hypotheses created in response to previous work discussed in this section.

3.2.1 Arledge's Work

The connection between the style and form of iconography has been explored in previous research, memory is defined as the ability to recall information (Harvard Medical School, 2022). Starting with "Filled-in vs Outline Icons: The Impact of Icon Style on Usability" (Arledge, 2014), Arledge's study attempts to come to a conclusion on if single-colour, "flat" icons are selected faster and more accurately when shown in a filled-in or outline style (p. 1). In the study, Arledge gives examples of what a "Filled-in" and "Outline" icon looks like. Arledge also gives the reader synonyms for "Filled-in" icons, these icons are also referred to as "solid" icons and outline icons are also referred to as "hollow" icons.

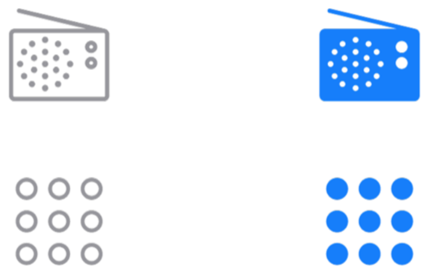

Figure 3.1: Example given by Arledge (Outline icons on the left and Filled-in icons on the right)

Arledge developed a program that required the study's participants to take a test that measured the speed and accuracy of responses given when the participant was told to select a specific icon among other icons i.e. "distractors" - The study used 20 unique icon forms. Arledge's study had 1,260 participants. Outline icons in white with a black background took the longest time to be identified, Arledge reported a 170 millisecond difference between these icons compared to the other style colour variations of icons used – Arledge reported that these other variations did not have any significant differences in comparison to each other. Arledge reported that the effect of icon style varied from icon to icon, concluding that the "filled in" icon style did not perform objectively better than the outline style, Arledge argues that the icons form has a greater influence on usability compared to the style or colour of the icon (p. 53). Because of this, Arledge concludes that one icon style was not objectively better than the other (p. 1).

Arledge's study has yielded a large following online with multiple websites writing articles on the paper and presenting the ideas discussed as fact. For example, a 2019 article by uxmovement (uxmovement, 2019) cites Arledge's paper and makes statements on the recognisability of both Solid ("Filled-in") icons and outline icons, claiming that this is partly affected by "characteristic cues", parts of the icons form that users use to identify an icon. The concept of identifiability enabled characteristic cues or resemblance is something that could be used towards future research (more on this later). Although containing a potentially useful insight into the usability of icons, the article has the significant limitation of not having made any attempt to replicate Arledge's findings to back their statements, a trend that occurs across many articles that reference Arledge's work.

A 2020 article titled "Outline Icons vs Solid Icons: Which are faster to recognise and when?" (KrishaWeb, 2020) also cites Arledge's 2014 paper and concludes that Solid icons are preferred, as they are more recognisable than hollow icons, despite Arledge's original paper stating quite the opposite: Given the results of Arledge's study, Arledge could not come to a conclusion on if filled-in or outline icons were selected faster or more accurately than one another. Arledge goes on to say that because of this result, designers should not view outline icons as less usable than filled-in icons (Arledge, 2014, p. 50). Because of this, Arledge's paper states a conclusion that does not represent or agree with the articles it has inspired (uxmovement, 2019; KrishaWeb, 2020), this represents an inconsistency between the original study and sources that attempt to reiterate the result of its work whilst also providing no attempt to replicate Arledge's conclusion. Alongside these two articles, more articles reference Arledge's study (liu, 2017; Wake, 2021), without repeating the study.

Furthermore, when looking at Arledge's study on Google Scholar, the search engine claims it has been cited in four papers (Rossi & Lenzini, 2021; Perrier et al., 2022; Eide et al., 2021; Rossi & Lenzini, 2021) recently published between 2021 - 2022. After reading these four papers, we can conclude that although these papers reference Arledge's study and were published quite a while later, the work done in these papers do not continue or rival the work done, meaning no visible attempt at future work has currently been conducted.

3.2.2 Iconography Studies Related to Memory and Cognition

Beyond the work of Arledge, other studies have also explored the recognisability of icons beyond their style, be that solid, outline etc. Guo (2016) attempted to identify the point at which users are unable to recognise interface icons, featured on a smart phone, based on the simplicity levels of the icons (p. xii). The study is comprised of three survey parts that test the icon recognition of "differently simplified icons".

The study had two phases:

- Phase One – In Part one, participants had to indicate which icons of the five simplicity levels matched the definition given to them. The most recognisable simplicity level and the simplest level of icons recognisable by participants, were to be used in Phase 2.
- Phase Two – In Phase two, the most recognisable simplicity level was used for Part two, and the maximum simplified icons for each definition of icon (e.g. Message, Email, Camera, Browser, etc), was used for the survey in Part 3. In Part two and Part three, the participants had to guess and state the definition of each icon they saw. Participants also had to use a slider to indicate how confident they felt about their answers.

Relevant to the work of Arledge, Guo reported that in "Phase 1", outline style icons (the "second most" simplicity level according to Guo) were the most recognisable for most participants and that "silhouette" i.e. shapes filled with one single solid colour (the highest level of simplicity according to Quo) (p. 30) icons had the lowest recognition rate.

Punchoojit and Hongwarittorrn (2018) investigates if an icon's background colour and the icon's symbol itself had an influence on menu selection time. The paper had two types of stimuli, the first were a range of icons (The "Pictorial Icons"). The second set of stimuli were the matching words that described the icons (the “textual” icons). The example of these two stimuli types was given as an icon that represented a picture of a clock (The Pictorial icon) and the word “clock” in the participant’s native language (The textual icon). The paper concluded that icons with a multicoloured background and pictorial symbol had the fastest menu selection time. The future research section suggests extending the current research in order to gain more understanding on the influence of the icon background and the icon symbol on icons and menu designs. Punchoojit and Hongwarittorrn proposes future work concentrated on the icon's pictorial symbol and creating situations that add an element of abstractness to the stimuli. Punchoojit and Hongwarittorrn indicate this when they state that if the icons appearance is abstract or not familiar to the user, the textual symbol could produce a faster selection speed. This comment invites this current study to question if future work on icon recognition should focus on varying levels of icon abstractness to potentially observed varying levels of selection speed.

Legleiter and Caporusso (2020) investigates the objectives: 1) Evaluate whether there is a case for universal icon meaning for icon sets and 2) gain insight on the factors that affect recognition of typographic icons. Two surveys were conducted:

- Survey A: Icon Meaning – The purpose of the questionnaire was to evaluate icon recognition in the absence of context. Participants had to select the meaning of an icon from a predetermined list of choices with one choice being the correct answer.
- Survey B: Icon Contextual – Survey B’s purpose was to evaluate icon recognition in the presence of context. Participants had to select an icon among the four giving the best described based on the description given.

The questionnaires given were spread on social media and completed independently by participants without supervision. After completing the part of the section on icon meanings, participants answered questions about their degree of familiarity with the symbols, preceding interactions with the icons as well as the confidence levels in their responses.

The 17 icons used were classified under three different semantic classifications: “Arbitrary”,” Reference” or ” Resemblance”. The authors justify this classification by adding that the icons were classified under the three semantic classifications in order to be consistent with prior research (p. 4). Legleiter and Caporusso cite Nielsen (2014) when making this statement.

Legleiter and Caporusso's study concludes that despite some icons supporting the theory of a universal meaning, symbols that all represent a common function, might still be perceived as not representing anything by other users. The paper also states that the results show that icon recognition is correlated with human factors both intrinsic and extrinsic, the paper highlights age and education level specifically. Legleiter and Caporusso also explain that in regard to the study's results, proficiency with technology such as the internet and mobile phones, had no influence on the individuals' ability to recognise the meaning of the icons shown (p. 7).

3.2.3 Nielsen's Work

Focussing on the source used by Legleiter and Caporusso (2020) to classify its icons, Nielsen (2014) introduces the reader to the notion that although visual designs vary vastly from one another, there are classifications of icons that can help users better understand what makes icons easier to understand. Nielsen gives these icon classifications alongside their descriptions:

- Resemblance Icons: Icons that depict physical objects in the real world, such as a lock icon depicting a physical lock.
- Reference Icons: Icons that depict an object meant to reference or represent a concept, such as a floppy disk icon representing a save function.
- Arbitrary Icons: Icons that have no inherent meaning.

After defining these icons, Nielsen adds that the classification of an icon can vary with changes in context and people. Nielsen gives the example of people of different age groups regarding icons that resemble technology of a specific time period to be either arbitrary or resemblance depending on the persons age. This indicates that the inclusion of Nielsen's icons in an experiment will depend on what classification the person(s) conducting the experiment think is the most appropriate for each icon. This is supported by the fact that Legleiter and Caporusso (2020) classified their icons under the guidance of Nielsen's definitions. Furthermore, an example of these different classifications is dependent on any one person's perception. Nielsen's article concludes by stating that Resemblance icons have the best usability.

3.2.4 The Use of EEG in Iconography Studies Related to Memory and Cognition

EEG has been featured in the research field of memory studies for iconography, this technology has an already established relevance to studies investigating Icons. For example, Xiao-Fan et al. (2014) uses ERPs to explore how the concreteness of an icon impacts the icons recognisability. In the study, Xiao-Fan defines "Concrete Icons" as icons that are intended to look like objects in the real world. Xiao-Fan's definition is similar to Nielsen's (2014) definition of resemblance icons - both definitions describe icons that represent tangible objects in the real world. The same can be said for Nielsen's definition of arbitrary icons and Xiao-Fan's definition of abstract icons i.e. icons that do not have real-world references. Both definitions arguably describe icons that have no inherent meaning.

However, despite any common definitions this paper might share with Nielsen (2014), Xiao-Fan's study still represents an example of where EEG has been used to explore how specific icon classifications (in this case "concrete and abstract icons") are recognised in comparison to each other. This is similar to Arledge's goal of determining if single-colour, "flat" icons are selected faster and more accurately when shown in a filled-in or outline style. Both experiments were conducted to explore the recognition of classified icons, Xiao-Fan's study uses "concrete and abstract" and Arledge's study uses "filled-in and outline". EEG recording being used to previously test the cognition of classified icons makes this data collection process appropriate for use in the current study as this current study also intends to test the cognition of its classified icons, the purpose of this being to investigate the memorability and thus the usability of each classification used (Filled-in, Outline, Resemblance and Arbitrary).

Expanding on the usage of EEG in previous studies that focus on iconography, Cherng et al. (2016) explored the implications of icon design in connection to icon semantic using an EEG-based approach. Cherng's study explores the concept of semantic distance, a measurement that defines the level of association between an icon's appearance and the function the icon is intended to represent (p. 1). The study goes on to define its two classifications of semantic distance as "close" and "far", with the former describing icons that help users to recognise the function of the icon more quickly (p. 2) and the latter being icons that are recognised less quickly. Cherng's paper like Xiao-Fan et al. (2014), represents an attempt to observe the differences in user interaction and cognition between semantically varying stimuli by using EEG. Cherng et al. (2016) supports this conclusion as their study was successfully able to find that “close” icons better attracted attention. Cherng’s paper also concludes that an ERP was successfully used to determine the semantic distance of icons. This not only shows that EEG has been used once again in a study evaluating the usability of icons with varying icon design thus boosting its popularity regarding this particular use case, but also shows the validity of EEG as a tool for evaluating the usability of iconography, this conclusion represents the success of EEG as a tool of measurement and investigation.

Yeh et al. (2013) continues the narrative of EEG being a valid tool when evaluating the usability of icons. In their study, Yeh explores the effect of varying colour combinations and icon exposure times on icon legibility and EEG responses. The study led to Yeh reporting back that a significant effect was observed and that this result was ‘confirmed’ in EEG analysis. This paper expands on the importance and suitability of EEG to studies focussing on the cognition of iconography as unlike Xiao-fan and Chreng’s previously mentioned studies, this paper does not focus on using EEG to show varying levels of cognitive performance through the use of varying semantics in the design of its stimuli but instead focusses on showing a difference in cognitive performance using other variables such as the colour of the icon, its background and the icon’s exposure time.

3.3 N-Back Timing

This section of the literature review explores the stimuli and interstimulus interval (ISI) timings used in previous studies that use the N-Back Task. The following discussion on previous literature will justify the timings used in the current study.

Nawaz et al. (2023) assesses working memory performance using single icons in the form of letters, from the English language, as its stimuli. This study used the N-Back task, specifically the 1-Back and 2-Back tasks (p. 5). Additionally, this study used EEG as a method of data collection. In terms of relevance, these two factors make Nawaz et al. (2023) relevant to this literature review and this current study, not only because the paper uses EEG but also because it uses the N-Back task. This study used 1500 ms as its stimuli presentation time and 500 ms as its interstimulus interval for both N-Back tasks. Similarly, Grissmann et al. (2017) uses EEG in a study that investigates working memory load. Grissmann also uses the 1-Back and 2-Back tasks with a 1500 ms stimuli presentation timing. Grissmann justifies using this timing by putting forward the argument that this specific timing ensures that enough trials can be recorded to get reliable estimates for the EEG recording (p. 4).

This study also argues that the ISI of the N-Back task should be varied, to avoid periodic responses in the EEG data recorded, Grissmann et al. (2017) used an ISI of 1500 ms with 1 to 500 ms of jitter at the end of the ISI (p. 4). Given this, this current study will use a range of timing for its ISI to prevent these "periodic responses". By eliminating periodic responses, this prevents the participants brain from getting into a rhythm or pattern that might interfere with their N-Back responses.

Looking further into current research that has been done concerning the specific stimuli time of 1500 ms, Grissmann et al. (2017) has inspired more recent work in the form of Ece Aksen (2023) which notably investigates the effect of working memory as well as emotion valence on state impulsivity. In this study, the stimulus presentation time of 1500 ms used by Grissmann et al. (2017) is used by the N-Back tests in this study, one of which being the 1-Back task. Ece Aksen (2023) labels this decision as consistent with previous work (p. 18), whilst citing Grissmann et al. (2017). The inclusion of this stimulus presentation time due to Grissmann et al. (2017) as well as this paper being a much more recent study, contributes to the trend of 1500ms being a popular stimulus presentation time in working memory studies.

This trend continues in a study that did not have the sole purpose of investigating working memory. Lahr et al. (2018) evaluates Huntington's Disease related changes in the neural network that enables working memory. The study uses N-Back tasks such as the 1-Back and 2-Back, while using a 1500 ms stimuli presentation time (p. 3). This confirms that even outside of the field of monitoring visual recognition for the sole purpose of studying memory, this specific timing is still being used, thus supporting a trend of using 1500 ms for the presentation of stimuli during an N-Back task. Furthermore, this paper shows that this specific timing has utility even for purposes outside of the aforementioned studies in this review.

This review pivots to literature based on the timings used for the ISI. The ISI, alongside the stimuli presentation time, could impact the data recorded as well as the voluntary data given by participants (i.e., the number of answers when conducting the N-Back task and the accuracy of their input).

Regarding studies that mention what ISI has been used, specifically in an investigation regarding working memory and EEG, Pergher et al. (2018) uses 2000 ms as the ISI in their use of the N-Back task. Pergher et al. (2018) focusses on the decline of cognitive performance caused by the progression of age. In order to mitigate this, this study uses the N-back task to verify if it improves both trained working memory and untrained cognitive functions.

This study compares the cognitive performance amongst a diverse group of adults (i.e. young and mature, healthy adults) (p. 1). Relevant to the current study, this cognitive performance is based on the following: N-Back tasks including two types that will be used in this study (1-Back and 2-Back) as well as EEG (specifically the P300 ERP component), reaction times and accuracies recorded during the N-Back tasks. Pergher et al. (2018) uses an N-Back ISI of 2000 ms, giving a relevant ISI time to this current study. This timing is relevant to the current study as supported by this paper's use of experimental components that this current study also intends to use, such as EEG, ERPs, and a wide range of adults.

Although Pergher et al. (2018) uses 2000 ms for its ISI, this paper alone is not sufficient to conclude on what ISI should be used for the current study. Considering this, another paper also uses an ISI of 2000 ms. Pergher et al. (2019) suggests the use of this ISI when investigating the differences in younger and older adults. Moreover, this paper focusses on differences in EEG (specifically the P300 ERP component) as well as differences across a range of N-Back tasks. These N-Back tasks included the 1-back and 2-back variants this current study intends to use. It is during these N-Back tasks that Pergher et al. (2019) uses an ISI of 2000 ms. Pergher et al. (2019) represents another example of a paper using this specific ISI for these N-Back tasks.

In addition to this, the latter paper (Pergher et al., 2019) does not cite the work of Pergher et al. (2018). Both papers also have two different bibliographies. It is because of these two facts that this literature review can assume that these two papers by Pergher, are independent from one another. Because of this, Pergher et al. (2019) represents an unbiased agreement with Pergher et al. (2018) on the use of 2000 ms as an N-Back ISI timing.

3.4 Event Related Potentials

This study will use EEG data recorded from participants while taking the N-Back task. This data will contribute towards the study's conclusion alongside the reaction times recorded from the

N-Back task. When looking at the EEG data recorded, this study will use ERPs that can give further context to the inner workings of each participant's neural activity when participating in the study.

When focussing on the ERPs this study will use when analysing the EEG data recorded, the first source looked at was an overview of ERPs. This overview acted as an introduction/indication towards the ERPs that could be isolated as appropriate to this current study. Sur and Sinha (2009) gives the reader an overview of different ERPs (p. 70). The following ERPs were identified based on their description given by the author. These ERPs were picked based on their potential to aid the analysis of the current study's data.

N100 - Described by the author as recorded when a user encounters an unexpected stimulus, the N-Back task will use perceivably unexpected stimuli to both elicit reaction times from the participant thus immediately showing N100's relevance to this current study's methodology.

P300 - Sur states that greater levels of attention, produce larger P300 amplitudes. This ERP is of interest to the current study as if P300 and the level of attention given by participants are correlated, then attention levels can be measured in this study. The P300 amplitude is said to be an important ERP component for evaluating cognitive function (e.g. attention and working memory (Zhong et al., 2019, p. 4)). P300's relevance to working memory is further emphasised in the following papers.

Focussing specifically on P300, we move on to Scharinger et al. (2017). The intention of this study was to examine typical EEG correlates of working memory processing load like the P300 amplitude (p. 3). Scharinger justifies the use of P300 by stating that the analysis of P300 provides insight into working memory performance (Scharinger et al., 2017, p. 3). In the same section, Schringer also says that in N-Back task, the P300 amplitude decreases with an increasing working memory load (p. 3), this point is especially useful as in the data analysis of the current study, this study also intends to use N-Back as well as EEG to examine the memory loads of participants in an effort to conclude on memory performance. Knowing that a previous study has reported this correlation will give context to the EEG data this study records as well as give evidence towards the result of this current study. These statements are reinforced by the discussion section where Scharinger reports that as a result, based on all tasks used (including the N-back task): The P300 amplitude decreased with an increasing working memory load (p. 15). This reinforcement continues in the "P300 Amplitude" section of the discussion where Scharinger adds that the P300 amplitude was a good measure of changes in the overall WM-load based on the study's results (p. 16).

Looking at this paper, despite it reporting a correlation between P300 and memory load as a result of using the N-Back task, a finding that will give this current study's data more context, this study also uses other memory tasks that contribute to Scharinger's comment on the correlation between P300 and memory load. The current study does not intend to use the other tasks mentioned (Ospan and Dspan), making Scharinger's overall result only partially relevant to this current study as only part of Scharinger's study uses the N-Back task to explore working memory. Although this does not invalidate the relevance and usefulness of the study's conclusion to the current study, we move on to other literature to further prove the relevance of P300 to this current study.

Ren et al. (2023) investigated the neuropsychological activity of working memory load using the N-back task. As a result, Ren's study showed that in terms of memory load, a higher memory load caused a smaller P300 amplitude during the N-Back task, whilst a smaller memory load caused a higher P300 amplitude. (Ren et al., 2023, p. 1). In addition to this point, this paper cites the previous work of Scharinger et al. (2017). Ren et al., 2023 citing Scharinger et al. (2017) in their work as well as Ren echoing the same point made by Scharinger on the corelation between P300 and memory load could be argued to reinforce the trend of there being a correlation between a decrease in the amplitude of P300 in response to an increased working memory load in the N-Back task.

3.4.1 P300 and Memory

Adjusting focus to literature that validates the relationship between P300 and memory, this is supported by Polich & Kok (1995), confirming that P300 is thought to be related to both attention allocation and the activation of memory (p. 1). Polich continues this train of thought in a chapter of (Luck & Kappenman, 2013). It is in this chapter that Polich discusses the relationship between memory and P300, claiming that that an increase in memory load reduces the P300 amplitude (p. 164).

Amin et al. (2015) argues that there is a relationship between P300 and memory in a study that aimed to find the relationship between ERPs and memory recall, as well as predict a memory recall score using P300. Participants were split into two (high ability or low ability) based on their “Raven’s Progressive Matrices” result, this is a test used to measure general cognitive ability (Pearson Talent Lens, n.d.). Amin reported the “high ability” group recalled 10.89% more information than the “low ability” group, whilst also adding that the study’s ERP results showed that the high ability group had a P300 amplitude that was larger in comparison to that of the low ability group. Amin concluded that these results suggest an association between the P300 component and memory recall, arguing that because of this, the P300 amplitude could be used to support psychometric tests in assessing memory recall ability. In addition to Amin establishing a connection between memory and P300, similar to previous papers, Amin also comments on a link between P300 and attentional resource allocation by saying that the two are “Closely related” (p. 2).

Simpson and Rafferty (2021) also supports the argument that the two are related by providing evidence of this. Their study aimed to investigate a relationship between P300 and memory in virtual training environments. When observing P300, results showed a significantly larger peak-to-peak amplitude as well as a significantly greater peak latency when comparing high memory retention to low memory retention. Based on the study’s results, Simpson in their conclusion states that both the amplitude and latency of P300 is impacted by memory retention (p. 11). Simpson supports this conclusion in the abstract where they describe the relationship between P300 and working memory as “evidenced” due to their findings.

3.5 Electroencephalography

A previous discussion in this review looked at the N-Back task and how it was used to successfully investigate working memory in Wang et al. (2016). Secondly, the discussion on Event Related Potentials (Chapter 3.5) reached a conclusion that a particular ERP, P300, had a clear correlation with working memory load, this was indicated by previous papers (Scharinger et al., 2017, p. 3; Ren et al., 2023, p. 1; Luck & Kappenman, 2013, p. 165). These two points involved the utilisation of EEG in their respective studies in order to reach their respective conclusions.

This section will further explore the relevance of EEG to memory in order to explain why this technology is being used in this study. As of 2023, the University of Kent currently does not possess a functional magnetic resonance imaging device (fMRI), thus the alternative had to be used to conduct this experiment – a brain computer interface device capable of measuring EEG data. Despite this, this section of the literature review will discuss why EEG is a suitable choice for this study.

Returning to Wang et al. (2016), Wang states that EEG signals can be useful in measuring different working memory levels (p. 434). Wang et al. (2016) also argues that based on the study's outcome, a wireless brain computer interface that collects EEG, can monitor and assess memory load (p.434). These statements make EEG and the brain computer interface currently in possession at the University of Kent (i.e. The Starstim 32), suitable for use in this study's investigation of Hypothesis 3, which intends to show a clear distinction between the memory loads of its stimuli.

Returning to the literature review made on ERPs and their relevance to this study (established earlier in the "Event Related Potential" section), we return to its definition – ERPs, are event-related changes in the neural activity recorded by EEG (Sur & Sinha, 2009) – in this sense, ERPs can be seen as a form of EEG. Due to this definition, the relevance of ERPs to this study discussed earlier in this review also argues the relevance of EEG to this study.

Keeping the definition of ERPs in mind we move on to the subject of EEG compared to fMRI. A stark comparison between the two technologies, is the resolution provided. In Sturzbecher and de Araujo (2012), the authors explore the difference in the resolution between the two technologies, stating that EEG has great temporal resolution and can record neural activity with millisecond accuracy, whilst fMRI has great spatial resolution (p. 199). Temporal Resolution is defined as how closely the measured activity corresponds to the timing of the actual neuronal activity (Lystad & Pollard, 2009).

Applying the statement made by Sturzbecher and de Araujo (2012) and the definition of temporal resolution to our previous discussion on ERPs, P300 is an ERP component described by sources as important to evaluating working memory load. EEG has good temporal resolution, enabling neural events such as ERPs and the P300 component to be recorded in milliseconds rather than seconds. Using EEG rather than fMRI will allow the data recorded on neural events such as ERPs to be more time-specific and therefore more accurately recorded.

In addition to this, when looking at the definition of temporal resolution as well as the papers that discussed the timing of the N-Back task, the papers discussed describe the stimuli presentation time and ISI used in the N-Back task as configured using milliseconds.

This matches the unit of time the EEG is recorded in (milliseconds). This means that if EEG is used in this study, this will enable a good temporal resolution between the two data types. Keeping in mind that temporal resolution is defined as how closely the measured activity corresponds to the timing of the actual neuronal activity (Lystad & Pollard, 2009), In this current study, the measured activity (the N-Back task and its reaction times) will correspond exactly to the timing of the neural activity (ERPs) as the two will be recorded using the same unit of time (milliseconds), avoiding the need for estimates to be made on how the two relate to each other regarding their timing, something that might need to be done if fMRI is used as an alternative technology. In summary, fMRI records its data in seconds whilst EEG will afford a much faster and accurate recording of data as it will record in milliseconds.

A good temporal resolution will be enabled as the two data types will be recorded in the same unit of time, this will enable each data type to support each other due to a lack of ambiguity surrounding when they happen in relation to each other.

Regarding the work of Sturzbecher and de Araujo, although providing a clear clarification between EEG and fMRI, the current study is being conducted in 2023-24. Despite highlighting a technology that can support the experimental design of this current study, as well as lead to more accurate results, Sturzbecher and de Araujo's work was published in 2012, so a more recent agreement of their statement needs to be found, in order to further validate the accuracy of EEG data.

Mele et al. (2019) reviews the use of EEG and fMRI simultaneously, yet also boasts the same temporal resolution mentioned in the work of Sturzbecher and de Araujo (2012). When comparing the two in the article's introduction, once again, the same clarification is made between EEG and fMRI, describing EEG as the superior technology over fMRI regarding its ability to offer an excellent temporal resolution. Unlike Sturzbecher and de Araujo's work, this paper is a more recent publication, providing more temporal validity, i.e. two separate publications seven years apart both claim the same conclusion - the temporal resolution of EEG is excellent.

Revisiting Wang et al. (2016) to investigate this trend in assertions made, when comparing the different neuroimaging techniques, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and functional near-infrared spectroscopy (fNIRS), Wang states that similar to FMRI, fNIRS has a poor temporal resolution. Wang also states that EEG has a high temporal resolution and is recorded in milliseconds. Because of this, Wang concludes that EEG is an appropriate tool for capturing fast and changing neural activity. This assertion on the temporal resolution concurs with the trend highlighted in the two previous publications whilst also supporting the point that a high temporal resolution will facilitate the efficient recording of neural activity.

3.6 The Present Study

This section will explain the methodical decisions made in relation to the previously discussed literature. This section will explain the reasoning behind this study's use of iconography, timing for the n-back task, electroencephalography and event related potentials.

Returning to Arledge's 2014 study, this current study will attempt to partially replicate Arledge's study to add to this area of research. This current study will attempt to continue Arledge's study and come to a conclusion on if single-colour, "flat" icons are selected faster and more accurately if shown in a filled-in or outline style (Arledge, 2014, p. 1) , through the exploration of Hypothesis 1.

The word "partially" is used as this study will also employ an experimental methodology that is partly novel to the work of Arledge – E.g., the use of EEG and N-Back to test the speed and accuracy of participants interaction and reactions to the "filled-in" or "outline" icons. Contributing to Arledge's 2014 paper is important given the fact that there are currently no other studies that have expanded upon the research discussed in this paper and therefore it could be argued difficult to establish the reliability and validity of their results. Many recent articles, spanning the last 10 years, still reference this study with no review of the reliability of this paper therefore, replicating this study will enable readers of the aforementioned articles to reach a more informed and reliable conclusion.

Moreover, with the range of articles published with no counterargument to the 2014 paper spanning ten years, there is limited research that attempts to falsify or further prove the findings of the 2014 paper, hence why it's a good decision to partially replicate this study to reach the same or a different conclusion so that the research area can be informed by more studies. In addition to their being no counter argument to Arledge's 2014 paper, the website the study is hosted by specifically states that "This work has no parents" (Arledge, 2014), combining this with the fact that as of August 2023, Arledge's paper has only been cited by four sources that do not continue or rival Arledge's work, this means that Arledge's study represents a new body of work that is yet to be continued through further research. This study intends to change that by continuing this research with a partially new experimental design whilst using icons from the same collection of icons used in the 2014 study.

To partially replicate Arledge's Study, icons from a set of IOS version 7 operating system icons will be used, the Apple icons featured in Arledge's work also belong to this set - this set remains the only set of icons from Arledge's study that remains available. It is important that the same collection of icons featured in this study is used where possible so this current study's relevance to Arledge's work can be maintained.

This study will also involve the concept of identifying icons through their resemblance proposed by the frequent mention of "characteristic cues" from the articles surrounding Arledge's study. Characteristic cues mentioned by the aforementioned uxmovement (2019) and KrishaWeb (2020), have been described as unique design components that help users to identify an icon (KrishaWeb, 2020). As these characteristic cues seem to be design features varying from icon to icon, more universal definitions of icon semantic will be used in this study. This approach has been chosen as comparing 2 or more classifications of icon semantic will provide a clearer and more structured analysis of user interaction compared to characteristic cues as this would entail the comparison of several design features across several icons, such a comparison would result in a convoluted examination, obscuring any conclusion to be had on the impact of varying levels of identifiability on icon recognition. Supporting this decision for the use of varying levels of semantic came in the form of the future work proposed by Punchoojit & Hongwarittorrn (2018) which as discussed implied the use of varying levels of abstractness to measure a user's interaction with icons.

As a result, in addition to Solid and Outline styles of icons being used in the current study, this study will also use the classification approach discussed in Legleiter and Caporusso (2020) and Nielsen (2014), in an effort to contribute to the papers not only discussed on the style of icons but the form of icons being used as stimuli. In addition to contributing to the work done on the memory performance of the style and form of an icon, this decision has also been made to investigate and thus contribute to Nielsen's conclusion that Resemblance icons have the best usability. This will be done by directly comparing resemblance icons to arbitrary icons in terms of their recognition performance.

One may ask “Why will this study use Resemblance and Arbitrary icons and not Reference icons?”, This has been done mainly for simplicity. Resemblance icons have been said to have the best usability so this current study will attempt to test that by comparing them directly to another classification of icons. Connecting the work of Arledge and Nielsen, using this classification approach proposed by Nielsen will also bring novelty to the partial replication of Arledge’s study - a study that did not explore the semantic classification discussed by Nielsen but simply just the visual appearance of the icons used. It is important to note that although this study intends to use Nielsen’s classification just like Legleiter and Caporusso’s study, unlike Legleiter and Caporusso’s study as well as Arledge’s work, this current study will use an EEG recording alongside an N-Back task instead of a survey when evaluating recognition performance. To expand on this design choice for the current study’s experiment, EEG and N-Back conducted simultaneously will be used to test the memory performance of the icons when grouped into specific classifications based on Arledge and Nielsen’s work. The relevance of memorability to usability is given by Nielsen, J (2012) where memorability is referred to as one of five quality components usability is defined by. Memorability is the ease of which something is remembered (Cambridge Dictionary, 2024). Memorability will be tested using N-Back, a measure that is used for the assessment of working memory function (Meule, 2017).

Investigating the usability (specifically the memorability) of semantic classifications by including them in this study, will also contribute to Arledge’s conclusion on the usability of icons. This study concludes that the filled-in icon style is not objectively better than the outline style. The conclusion also goes on to say that the form of the icon has a greater influence on its usability rather than its style or colour. This current study will add to this conclusion as the forms defined in Nielsen’s classifications and the styles defined by Arledge’s paper will be compared together and separately in order to firstly reveal the icon type with superior usability in their respective families (Hypothesis 1 & 2) as well as observe any impact the two families of icons have on each other (Hypothesis 3).

To summarise on iconography, based on the literature discussed on cognition studies for iconography, this study will use the same set of icons used in Arledge’s study as well as their style (solid and outline) but will classify those icons under the direction of Nielsen’s work, specifically classifying the solid and outline icons under resemblance or arbitrary, representing a novel, hybrid classification, informed by the previous work of Arledge and

Nielsen. As a result of this, the current study will conclude on the memory performance of solid and outline icons, therefore continuing Arledge's work and enabling a conclusion to be made on hypothesis 1. As well as this, the study will also conclude on the memory performance of resemblance and arbitrary icons therefore also continuing Nielsen's work and contributing to their conclusions on usability, allowing this study to conclude on hypothesis 2. In addition to this, the literature related to EEG has shown the popularity of the tool when investigating differentiation in cognitive performance, in response to a varying icon design, including a specific design variable important to this current study - the semantic of the icon. It's because of this and the tool's ability to produce significant findings, in the specific use case of evaluating cognitive performance relating to varying icon design, that EEG will be used as a tool of measurement in this study.

Regarding the configuration this study will use in the N-Back task, the decision has been made to use the 1-Back and 2-Back tasks due to their popularity across literature that uses the N-Back task (Nawaz et al., 2023; Grissmann et al., 2017; Ece Aksen, 2023; Lahr et al., 2018; Pergher et al., 2018; Pergher et al., 2019). Due to their inclusion of the N-Back task, these studies by extension, also explore working memory (Owen et al., 2005; Meule, 2017). In addition to this, the decision has been made to use the stimuli presentation time of 1500 ms. This specific timing has been chosen because of a trend observed in working memory investigations which entails 1500 ms being used frequently as a stimuli presentation time in an N-Back task (Nawaz et al., 2023; Grissmann et al., 2017; Ece Aksen, 2023). One study in particular, provides the argument that this has been done to gain reliable responses (Grissmann et al., 2017).

When looking at what ISI to use, based on the aforementioned papers, the decision has been made to use an ISI of 1500 ms to 2000 ms. Firstly, this decision has been made because 1500 ms as an ISI has been used in a study on working memory (Grissmann et al., 2017), as well as the fact that the source demonstrated its reliability by mentioning how and why to avoid periodic responses. Secondly, the decision to use an ISI of 1500 ms to 2000 ms has been made because of the popularity of the 2000 ms timing across studies similar to this current study (Pergher et al., 2018; Pergher et al., 2019). This current study intends to collect reaction times, accuracies, and EEG during the N-Back task which these papers have also measured. Thirdly, neither papers report any issues caused by the 2000 ms timing in their respective studies. Moreover, 1500 ms to 2000 ms is being used due to the importance of avoiding periodic responses as mentioned. The point of avoiding periodic responses, was made in Grissmann et al. (2017). Ignoring the contingency of periodic responses is a risk that this current study cannot afford to take if it intends to produce accurate, quantitative findings. An ISI of 1500 ms to 2000 ms does represent a novel approach influenced by the previous literature (Grissmann et al., 2017; Pergher et al., 2018; Pergher et al., 2019). Any potential conclusions on the effectiveness of the ISI, provided by the results of this study, will add to the work of this previous literature.

Based on the publications discussed on EEG and ERPs, the following has been confirmed: EEG has a temporal resolution that is much better than its alternative, fMRI. The temporal resolution of EEG will allow the data recorded from participants to accurately correspond with the reaction times recorded during their participation in the N-Back task; both sets of data will be recorded in milliseconds thus allowing an accurate correspondence between the two. This conclusion is supported by the definition of temporal resolution in the context of measuring neural activity (Lystad & Pollard, 2009). This conclusion on the strength of EEG's temporal resolution has been boasted by previous literature. In addition to this, specific ERP components that link closely to working memory have been identified (P300 and N100) and regarded as relevant to working memory by previous literature (Scharinger et al., 2017; Ren et al., 2023; Luck & Kappenman, 2013; Wang et al., 2016). As ERPs are a product of EEG, this by default makes EEG suitable to measure working memory. It is for these reasons that EEG and by extension, ERP component P300, will be used as a tool, in this study to investigate the memory loads of different icon classifications, in the exploration of Hypothesis 3. Due to time constraints, N100 analysis is left as a future work. This study will focus on P300 due to it being used by previous work in conjunction with the N-Back task as well as the fact that multiple sources state that it has a relationship with memory load. This relationship will enable this current study to measure the memory load of its stimuli through the recording of EEG and observation of the P300 data produced.

4. Methodology

4.1 Introduction

This chapter presents a comprehensive overview of the methodology employed during this study. Specifically, this section explains the task design used, the participant selection process, the equipment utilised, the experimental procedure used and any issues encountered. This framework represents the environment in which this study's findings were created.

4.2 Ethics Approval

This experiment was conducted at the University of Kent and required human participants. Because of this the study had to apply for ethical approval from the Central Research Ethics Advisory Group (CREAG) at the University of Kent. The following documents were submitted to CREAG in order for the experiment to be allowed to begin:

- Full Ethics Application for Research
- Ethics Review Checklist for Research
- Research Proposal
- Participant Information Sheet (To be given to participant)
- Consent Form (To be given to participant)
- Advertisement for Research Project (Purposed with recruiting participants)

The participant information sheet, Consent Form and Advertisement was only used once the experiment was ethically approved by CREAG. After rounds of revisions, CREAG approved the application on 04/01/2023 (CREAG Reference Number: CREAG017-11-22).

4.3 Task Design (Conceptual Overview)

In the experiment, the task design was based on the N-Back task in order to investigate the memory performance of participants while using the following icon combinations:

1. Outline Resemblance Icons
2. Outline Abstract Icons

3. Solid Resemblance Icons
4. Solid Abstract Icons

Each participant completed a total of 8 N-Back Tasks – One for each permutation of icon combination and N-Back Type (1-Back and 2-Back). Each N-Back task had the participant monitor a sequence of icons and press the SPACE if the current icon matched the icon shown n places earlier. The 1-back and 2-back tasks were used to vary the memory loads and reaction times under different levels of difficulty. Each icon was presented for 1500 milliseconds followed by an interstimulus interval of 1500-2000 milliseconds in order to avoid periodic responses. This task design was implemented using Psychtoolbox 3 in MATLAB (Kleiner et al., 2018b). The several functions provided by this software tool enabled the precise presentation of the stimuli (i.e. Icon combinations) and the recording of this study's data. See Appendix C for an overview of the technical implementation of this task design. As a result of this task design, the following data produced will help this study achieve its research objective of investigating the memory performance of the icon combinations:

- Participant's Reaction Times
- Participant's Correct Answer Accuracy
- Participant's EEG data recorded simultaneously during the N-Back task.

4.3.1 Experimental Setup

Before the task design commenced, the Lab Streaming Layer (LSL) was enabled in MATLAB to enable the synchronisation of the participant's N-Back performance and the EEG data collected during the execution of the N-Back task. For example, events that occurred in the N-Back task such as the participant getting a correct answer, was timestamped, enabling the researcher to observe when the event happened in relation to the EEG data.

4.3.2 Participants Interaction

Before the N-Back task, participants were instructed to read the participant information sheet provided and ask the researcher present any questions if the participant deemed it necessary. Once reading the participant information sheet, if the current session was the participant's first N-Back task, they were asked to read, complete and sign the consent form - The following steps were not carried out until this requirement was met. Once the participant gave their consent for the experiment to move forward, the researcher present set up the experiment (further details in 4.6 The Experimental Procedure). During the N-Back Task, participants are shown a welcome screen that briefly explains the experiment about to take place and instructs them to begin the N-Back task when they are ready by pressing the SPACE button. During the N-Back task the participant presses the SPACE button when they think a match is present on screen. Feedback for Correct or Incorrect answers was given in the form of a visual cue. A green box for correct answers or a red box for incorrect answers was shown in the bottom left corner of the screen. Once the N-Back task had concluded, an end screen is shown prompting the participant to wait to be instructed by the researcher present. Once the experiment had ended, the participant was instructed to allow the researcher to safely remove the EEG equipment worn by the participant.

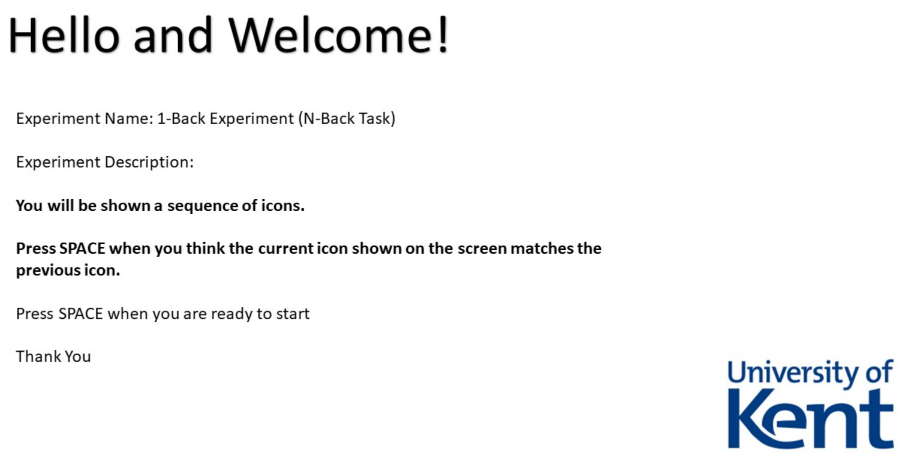

Figure 4.1: The welcome screen when the participant was conducting a 1-Back task.

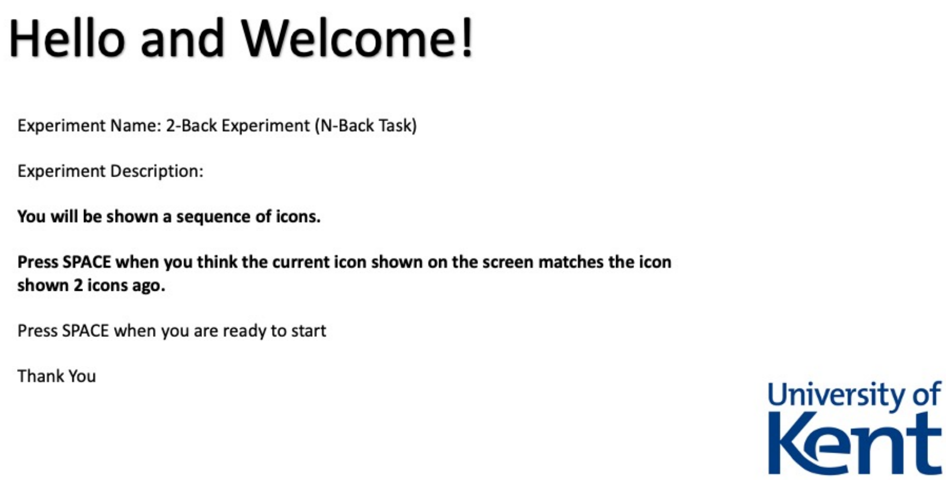

Figure 4.2: The welcome screen when the participant was conducting a 2-Back task.

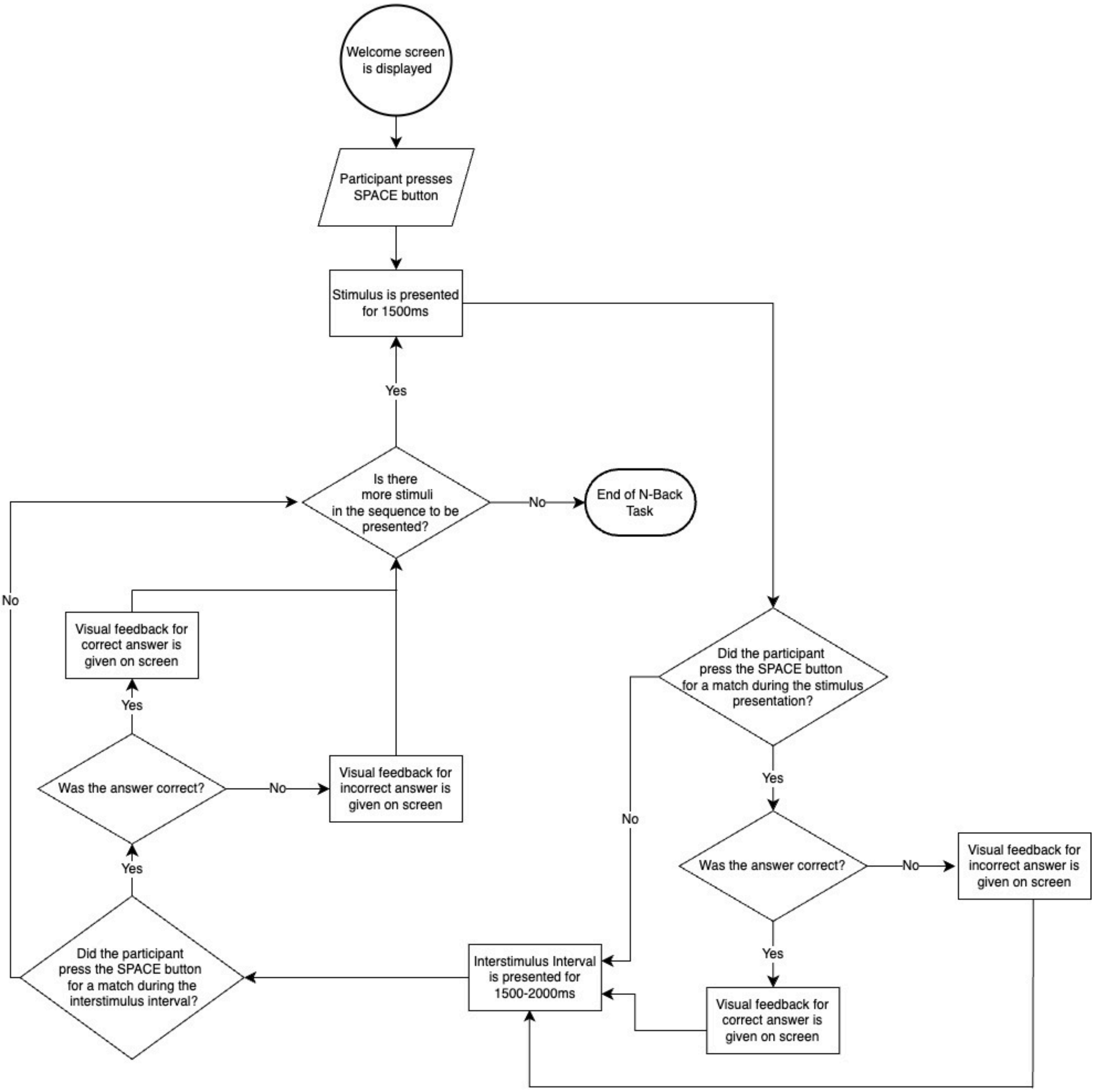

Figure 4.3: A flowchart that illustrates the experiment as seen by the participants.

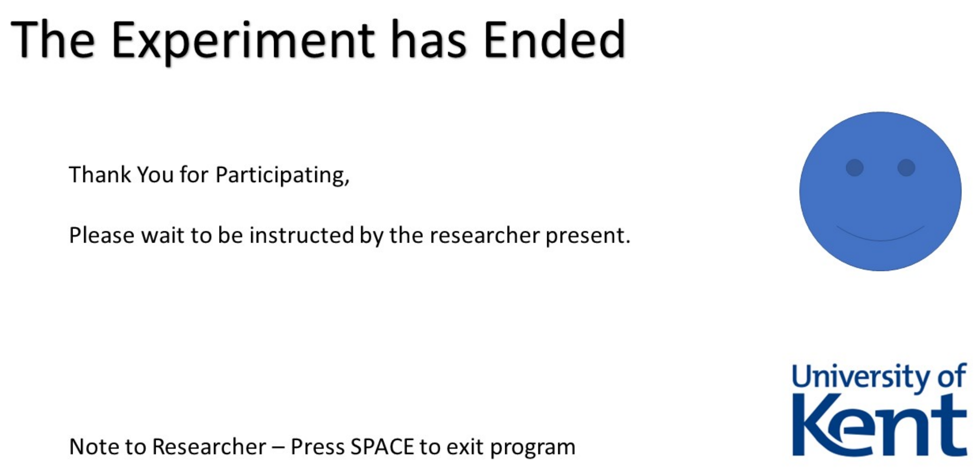

Figure 4.4: The end screen of N-Back task.

4.4 Participant Selection

Participants were recruited at the University of Kent via an advertisement being placed around the Canterbury campus. Those that participated were interested in the experiment and thus wanted to contribute their time to its conduction. Overall, 12 participants were recruited, of those 12 were six males and six females. The age range between the subjects were 21 and 24 years of age (Mean = 22.25, SD = 0.92). Each participant had moderate experience with icons featured in examples such as smart devices, computer systems and iconography in modern society. The advertisement that recruited these participants detailed that the experiment required the following out of any potential participants interested:

- Age 18 – 60, male or female.
- Participants who are able and willing to sit in concentration for the duration of the experiment.
- Participants who are not taking any medication.
- Participants with the ability to read and understand English.
- Participants who are not visually impaired (corrected vision with glasses or contacts lens is fine).
- Participants with no medical history of stroke, epilepsy, or any other neural disorders.

4.5 Equipment Used

The following resources were used in this study:

- Starstim 32: A wireless, non-invasive, 32-channel BCI device. This device was used to record the EEG from participants.
- NG geltrodes: A combination of an electrode & an electrode holder to inject electrode gel into and screw the electrode on to.
- The reference electrodes and accommodating ear clip: The reference is used to define the level of zero voltage and all other channels are expressed in relation to it. This is often referred to as the component that grounds the amplitudes of the EEG channels (Leuchs, 2019). The reference electrodes used with the Starstim were the cables labelled Common Mode Sense (CMS) and Driven Right Leg (DRL). These electrodes were connected to an ear clip also designed by the manufacturer to hold the electrodes in place on a participant's ear (Neuro Electrics, n.d.-b).
- Neoprene headcap (Starstim 32 accessory): A neoprene cap that accommodates the Starstim 32 device and the NG Geltrodes.
- Neuroelectrics Instrument Controller (NIC version 2): An integrated software environment for end-to-end management of the Starstim 32. This software was installed on the laptop used in this study. This software allowed the researcher to check that the EEG channels were reading correctly (with appropriate impedance) before conducting the experiment for each participant.
- Signa-gel: The electrode gel used in combination with the BCI and its electrodes (Signa-gel is recommended by the manufacturer of the BCI device).
- Asus Notebook PC (laptop): To conduct the experiment, the programmed N-Back task had to be shown on a screen so participants could interact with it, alongside this their EEG needed to be recorded whilst undergoing this task. Both constraints were fulfilled by the laptop used in the experiment.
    - At no point in the experiment was the laptop discussed connected to the internet, this was done to eliminate all possible distractions that could occur during each EEG recording, this was important as such a contingency could disrupt the participants focus, making the data collected in the participants EEG

session whilst conducting the N-Back task useless, these distractions were notifications such as the availability of a software update.

- The technical specifications of the PC were the following:
  - Processor - Intel Core I7 - 7500U CPU @ 2.70GHz
  - Installed RAM - 16GB (15.9 Usable)
  - Operating System - Windows 10 Pro

- Lab Streaming Layer: A system designed for the unified collection of real time data published on the website “GitHub”: by third parties, intended for public use. The LSL lab recorder provided by this software allowed the researcher to program timestamps into the N-Back task used so the participants reaction times and EEG data could be recorded simultaneously and logged into the same XDF file (the file format of the EEG data recorded).
- MATLAB: The N-Back task used was programmed using the programming language MATLAB, specifically version R2020a.
- Psychtoolbox 3: Psychophysics Toolbox Version 3 (Psychtoolbox 3) is a set of free MATLAB functions for neuroscience research. This tool set in conjunction with MATLAB was how the N-Back task was successfully programmed. Psychtoolbox 3 was a necessary choice for this study as the MATLAB functions it provides enables the presentation of accurately controlled stimuli (Kleiner et al., 2018b). This accuracy is enabled via the toolbox’s `WaitSecs` and `GetSecs` functions. `WaitSecs` enables the programs execution to be paused for a specific time. The authors of Psychtoolbox 3 claim the timing precision of this function is accurate to 1 millisecond (Kleiner et al., 2018c). The `GetSecs` function returns the time in seconds. The time returned uses the real-time clock in the operating system of the computer being used, with microsecond accuracy (Kleiner et al., 2018a). Both functions utilise `QueryPerformanceCounter()`, a windows function that retrieves a high resolution (less than 1 microsecond) time stamp, which can be used for time interval measurements (Microsoft, 2024). With these functions providing the appropriate timing accuracy that would allow this study to display stimuli using presentation times as specific as milliseconds, Psychtoolbox 3 in MATLAB was chosen to use when developing this study’s experimental procedure.

4.6 The Experimental Procedure

The experimental procedure was conducted as follows:

1) Give Participant information Sheet to Participant.
2) Give Participant Consent Form to Participant.
3) Make sure that all electronic devices that are not required to be on during the recording of EEG are switched off - (Refer to “4.8.3 Noise in the EEG” section in “4.8 Issues Encountered During Experiment" section).
4) Ensure electrode holders have been fitted into the cap.
5) Fit cap on to participant.
6) Apply electrode gel from scalp to top of the electrode’s holder using a syringe - a small portion of conductive gel should be injected using the plastic syringe. The syringe is blunt, injecting gel into the cap will be completely painless for the participant.
7) Screw on electrodes.
8) Repeat 6 & 7 until all electrodes have been screwed on.
9) Turn on and connect Starstim 3.
10) Fit matching cables to the electrodes – after fitting each cable, check that the electrode is constantly green on the Starstim user interface, reapply gel if this is not the case.
11) In addition to this, clean participants right ear lobe, stretch ear lobe and apply the earlobe electrodes (i.e., the reference, insert gel between the grips and apply).
12) If electrodes are constantly green – continue, else reapply gel.
13) Start LSL:
    a) Whilst in the LSL recording mode, ensure the correct Participant ID (%p), N-Back type (%s) and stimuli type (%a) (In that order so it complies with a file naming scheme) is given in LSL before the recording is started.
    b) Run MATLAB pre-experiment (this enables LSL timestamps).
    c) Press the update button and select all streams – ensure all streams required for experiment have been ticked.

Before starting, give participant advice on how to behave during the recording to minimise EEG noise such as: not to clench teeth, breath through the mouth and importantly, to stay as still as possible.

14) Start LSL recording.

15) Start StarStim EEG recording in the Starstim user interface

16) Start MATLAB experiment code.

17) After each recording, ensure that the recording is stopped and stored correctly.

18) After a break for each recording – Go to step 12.

After all recordings have concluded do the following:

a) Remove cap from participant.

b) Clean and dry cap.

c) Clean and dry electrodes.

d) Charge laptop.

e) Charge StarStim 3.

The icons used were classified and labelled using classification approaches used in previous studies (Arledge, 2014; Nielsen, 2014):

- Resemblance icon – an icon that depicts a physical object which the icon is intended to represent.
- Abstract icon – an icon that only has meaning by convention.
- Solid icon – icon that uses the “filled-in” style detailed in Arledge’s study.
- Outline icon – icon that uses the “outline” style detailed in Arledge’s study.

Each experiment represents an EEG recording session as shown in Table 4.1:

Table 4.1: ***Each style and classification combination used in each experiment.***

		Icon Style	
		Solid	Outline
Icon Classification	Resemblance	Experiment 1	Experiment 2
	Abstract	Experiment 3	Experiment 4

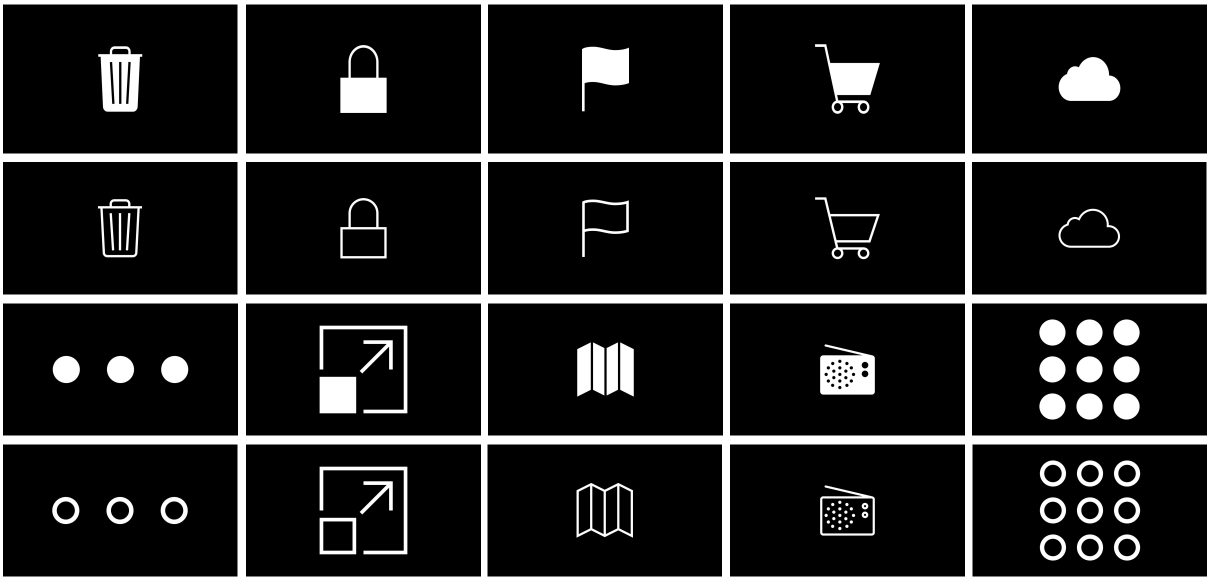

Figure 4.5: All Icons used as stimuli in the N-Back tasks (1-Back & 2-Back)

Further details of the experiment are given in the form of a question and an answer:

Why was this done? - Every participant was recorded for EEG using a unique sequence of the four experiments (Solid Resemblance, Outline Resemblance, Solid Abstract and Outline Abstract). This was done to avoid any similar habituation trend in the data collected. Such a contingency could invalidate the data collected, if each participant was exposed to each experiment in the same order.

Why was a break taken after two participants? - After two participants had been recorded using the 1-Back and the 2-Back task for all experiments, a short break of a few days was taken in order to ensure the data was collected properly during the beginning of the experimentation. Luckily no major issues were found in the EEG data collected from the first two participants, thus the remaining 10 participants were recorded using the schedule above.

What is the “stimuli order”? - Each “stimuli order” is a unique sequence of stimuli used for each participant's N-back test. Each sequence was comprised of 210 random numbers between 1 to 5, representing the 5 possible icons to be shown on screen. Each sequence featured 50 stimuli that matched the stimulus shown N places ago i.e. representing 50 opportunities for the participant to submit a correct answer.

Regarding the participants involved, each participation was arranged in the following structure (Table 4.2):

Table 4.2*: Schedule of N-Back experimentation trials.*

Participant ID	Day	Morning	Afternoon	StimuliOrder .mat files
P1	1	1, 3, 2, 4 (1 back)		`StimuliOrderS1N1.mat`
	2	1, 2, 4, 3 (2-back)		`StimuliOrderS1N2.mat`
P2	1		1, 2, 3, 4 (1-back)	`StimuliOrderS2N1.mat`
	2		1, 3, 4, 2 (2-back)	`StimuliOrderS2N2.mat`
BREAK				
P3	3	1, 4, 2, 3 (1-back)		`StimuliOrderS3N1.mat`
	4	1, 4, 3, 2 (2-back)		`StimuliOrderS3N2.mat`
P4	3		2, 1, 3, 4 (1-back)	`StimuliOrderS4N1.mat`
	4		2, 1, 4, 3 (2-back)	`StimuliOrderS4N2. mat`
P5	5	2, 3, 1, 4 (1-back)		`StimuliOrderS5N1.mat`
	6	2, 3, 4, 1 (2-back)		`StimuliOrderS5N2.mat`
P6	5		2, 4, 1, 3 (1-back)	`StimuliOrderS6N1.mat`
	6		2, 4, 3, 1 (2-back)	`StimuliOrderS6N2.mat`
P7	7	3, 1, 2, 4 (1-back)		`StimuliOrderS7N1.mat`
	8	3, 1, 4, 2 (2-back)		`StimuliOrderS7N2.mat`
P8	7		3, 2, 1, 4 (1-back)	`StimuliOrderS8N1.mat`
	8		3, 2, 4, 1 (2-back)	`StimuliOrderS8N2.mat`
P9	9	3, 4, 1, 2 (1-back)		`StimuliOrderS9N1.mat`
	10	3, 4, 2, 1 (2-back)		`StimuliOrderS9N2.mat`
P10	9		4, 1, 2, 3 (1-back)	`StimuliOrderS10N1.mat`
	10		4, 1, 3, 2 (2-back)	`StimuliOrderS10N2.mat`
P11	11	4, 2, 1, 3 (1-back)		`StimuliOrderS11N1.mat`
	12	4, 2, 3, 1 (2-back)		`StimuliOrderS11N2.mat`
P12	11		4, 3, 1, 2 (1-back)	`StimuliOrderS12N1.mat`
	12		4, 3, 2, 1 (2-back)	`StimuliOrderS12N2.mat`

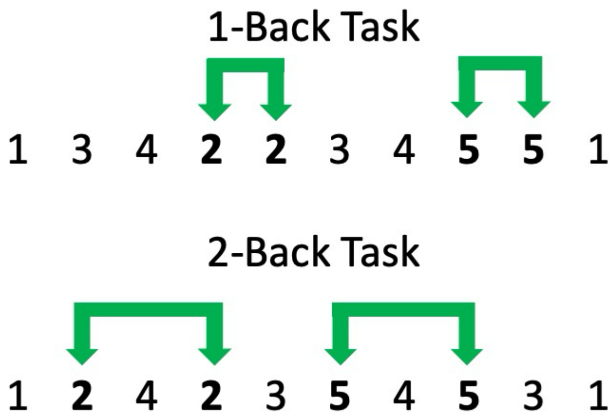

Figure 4.6: An example of a stimuli order for the 1-back task (Top) and an example of a stimuli order for the 2-back task (Bottom), the green arrows indicate the matches.

What was the Electrode Placement? - The Neoprene headcap that accommodates the "NG Geltrodes", used the following arrangement of electrodes (Neuro Electrics, n.d.-a):

Table 4.3: The electrodes accommodated by the Neoprene headcap.

Brain Region	EEG Electrodes
Frontal-polar area	Fp1, Fpz, Fp2
Anterior-frontal area	AF7, AF3, AF4, AF8
Frontal area	F7, F3, Fz, F4, F8
Fronto-central area	FC5, FC1, FC2, FC6
Central area	C3, C1, Cz, C2, C4
Temporal area	T7 (T3), T8 (T4)
Central-parietal area	CP5, CP1, CP2, CP6
Parietal area	P7, P3, Pz, P4, P8
Parieto-occipital area	PO7, PO3, PO4, PO8
Occipital area	O1, Oz, O2

Despite naming 39 different electrodes, the electrodes used in this study are as shown in Table 4.4:

Table 4.4: EEG Channels in study and their channel numbers

Channel Description	Channel Location	Channel Description	Channel Location
P8	1	PO3	17
T8	2	O1	18
CP6	3	Oz	19
FC6	4	O2	20
F8	5	PO4	21
F4	6	Pz	22
C4	7	CP1	23
P4	8	FC1	24
AF4	9	P3	25
Fp2	10	C3	26
Fp1	11	F3	27
AF3	12	F7	28
Fz	13	FC5	29
FC2	14	CP5	30
Cz	15	T7	31
CP2	16	P7	32

4.7 Issues Encountered During Experiment

4.7.1 EEG device battery

Regarding the device used to record EEG from participants, the Starstim 3, due to this device being wireless it had to be recharged after each participant's recording session as a full recording session with a participant would diminish the device's charge to <20%.

To expand further, each participant recording session would consist of four separate EEG recordings for each of the types of stimuli being shown (Solid Resemblance, Outline Resemblance, Solid Abstract and Outline Abstract) and each separate EEG recording would use on average 20% charge, in addition to this setting up the EEG device and checking that each channel was correctly calibrated used around 20% (hence each participant recording session diminishing nearly the entire charge of the device). Charging the device to 100% required an average of three hours, making it only possible for the researcher to conduct a maximum of two participant sessions a day. In total 24 separate participant recording sessions were carried out, two for each participant (12 participants in total were used, one session was used for a 1-Back session and one session was used for a 2-Back session). Despite the logistics of the experiment in connection to the Starstim 3, the experiment did not run into any issues regarding the charge of the device.

4.7.2 Shape of Participant Earlobe

As previously discussed, this experiment used electrodes to record EEG, among those electrodes, was the plastic ear clip that accommodates the reference electrodes, this component was required to ground the signals collected by the electrodes. To be used, the reference had to be attached to the right earlobe of the participant. This is because the wire used by the reference was only long enough to be attached to the right earlobe. Alongside this, during setup and testing, the earlobe was the most secure part of the ear when using the required electrode gel - The electrode gel would make the reference slip off the participants ear if the reference was not securely attached. The issue encountered was that a portion of participants initially recruited had an earlobe shape that could not afford a stable grip on the participants earlobe (‘lobule’) due to their shape (Neuro Electrics, n.d.-b; McGovern Medical School, n.d.). To resolve this issue, only participants with a pronounced earlobe were used in this study. This was done to ensure the EEG data collected was not ruined by the reference being attached to a part of the ear that was not secure.

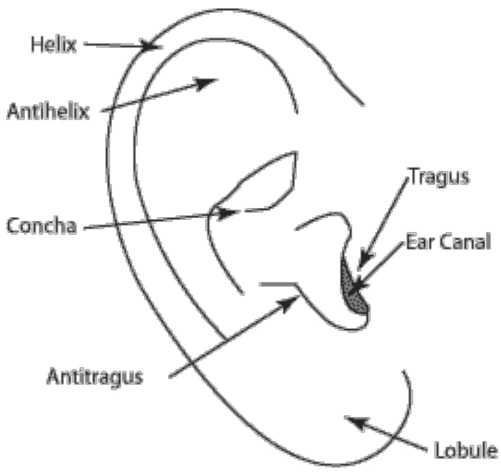

Figure 4.7: Diagram of Ear featuring ear lobule (McGovern Medical School, n.d.).

4.7.3 Noise in the EEG data

The following actions and components of the study created noise in the EEG data (this noise had to be filtered out in the data analysis of this study):

- Participants clenching their teeth.
- Participants blinking.
- The electronic noise of the lights in the lab.
- The electronic noise of the laptop used.
- The electronic noise of the devices owned by participants - because of this known risk, participants were asked to either not bring any electronic devices to their participation, or if they did, those devices were switched off and put in a different room to prevent any potential noise produced by these devices from affecting the outcome of the study.

4.7.4 Lack of soundproofing

Regarding the location of the lab at which the experimentation took place, this lab was in the school of Computing at the University of Kent, this building is available for public use by students and members of staff. The corridor that led to the lab could and was entered many times during a participant's experiment as it led to other rooms that people presumably needed to access.

Despite this, the lab’s privacy was ensured as it was only accessible using a lock and key, of which the researcher had the key in their sole possession for the duration of the experimentation component of this study, meaning there were no incidents of the experiment being interrupted due to someone walking into the lab or knocking on the door as through the use of a sign, people were notified that an experiment was taking place and to not disturb anyone in the lab at the time. Despite this notification to the public and the public respecting the constraints of the study by A) Not disturbing the people in the lab at the time and B) Being completely silent when walking past the door of the lab, the lab itself did not have any measures of soundproofing. However, there were no incidents in which the researcher or participant were disturbed due to a noise incident.

5. Data Analysis

5.1 Introduction

The methodology established was used for the purpose of generating data that would be relevant and useful towards the investigation of this projects research question and accompanying hypotheses. However, to make this data intelligible and effective towards the study's conclusion and any discussion that can be had in relation to other studies, its processing and analysis is required. This section details the data used, its purpose, the tools used to analyse this data and the analysis techniques employed to reach an intelligible result on what narrative the data conveys. It is important to note, that in the case of the data analysis, only the stimuli that caused 'Correct' answers were analysed for their timing (i.e. Reaction Times and Accuracies) and EEG data. This decision was made as these stimuli represent the target stimuli required to elicit P300. Taking this into account, this would make the ERP analysis of stimuli that caused incorrect answers or correct answers to be missed, irrelevant to the P300 component, invalidating any connection that could be argued between the ERP analysis and any significant result found when analysing the matching timing data.

5.2 Methodology

5.2.1 Datasets

The experimental procedure used produced the following datasets:

- Electroencephalogram data from the recording sessions.
- The timings of each event that occurred during the recording sessions.

5.2.1.1 Recording Session – EEG

EEG data reflects the electrical activity of each participant's brain during each recording session. Each participant undertook eight separate EEG recording sessions in total based on a different combination of icon classifications and versions of the N-Back task. During each EEG recording, 32 channels of EEG were recorded simultaneously, lasting 15-minutes on average. The direct purpose of this data is to contribute to the investigation of hypothesis 3. Due to the data collected (EEG) being that of a metric nature, as well as facilitating the measurement of memory loads via P300, this data thus also promotes the exploration of the research question which sets out to explore how the different icons impact memory performance.

5.2.1.2 Event Timings

Much like the EEG data, the event timings of the N-Back task are metric in nature. The recording sessions also entailed the creation of timestamps. These timestamps would be created in response to one of six events:

- The program starting.
- Each time a stimulus is shown, along with a number indicating the stimulus's place in the larger stimuli order (e.g., 'Stimuli Number: 1', 'Stimuli Number: 2').
- ISI – Interstimulus Interval.
- A correct answer being submitted.
- An incorrect answer being submitted.
- The program ending.

P01N1OA

1x633 struct with 3 fields

Fields	type	latency
1	"0"	1
2	"Program Started (Welcome Screen Exited)"	41854
3	"StimuliNumber:1"	46861
4	"ISI"	47624
5	"Answer Not Submitted"	48483
6	"StimuliNumber:2"	48485
7	"ISI"	49238
8	"Answer Not Submitted"	50162
9	"StimuliNumber:3"	50163
10	"ISI"	50914
11	"Answer Not Submitted"	51854
12	"StimuliNumber:4"	51855
13	"ISI"	52606
14	"Answer Not Submitted"	53465
15	"StimuliNumber:5"	53467
16	"ISI"	54219
17	"Answer Not Submitted"	55133
18	"StimuliNumber:6"	55135
19	"ISI"	55886
20	"Answer Not Submitted"	56664
21	"StimuliNumber:7"	56665
22	"ISI"	57416
23	"Answer Not Submitted"	58400
24	"StimuliNumber:8"	58401
25	"ISI"	59152
26	"Answer Not Submitted"	59949
27	"StimuliNumber:9"	59951
28	"ISI"	60702
29	"Answer Not Submitted"	61518
30	"StimuliNumber:10"	61520
31	"Correct"	61734

Figure 5.1: Visual Representation of Event Data for EEG Recording - (Participant 1, 1-Back, Outline Resemblance Stimuli)

Each timestamp would be labelled with a description of the event that triggered its creation. Unlike EEG data, this temporal data supports the projects means of reaching conclusions on hypothesis 1 and 2, both involving comparisons specific to time. For example, the time difference between a stimulus being shown and a subsequent correct answer being submitted, represents the time taken for a participant to submit a correct answer upon seeing the target stimulus i.e. the reaction time. Once all reaction times for all correct answers are calculated and analysed, the data will provide a conclusion on hypothesis 1 and 2.

5.2.2 MATLAB

MATLAB (Matrix Laboratory) is a high-level programming language. It enables its users to develop algorithms and manipulate large datasets in the form of matrices. MATLAB has already proved a sufficient tool for the development of the projects experimental design. In addition to this, MATLAB also provides an extensive set of tools that would enable the ease and practicality of the projects data analysis.

Firstly and logistically speaking, EEGLAB, an extension of MATLAB affords the importing of ".xdf" files, this is paramount to the data analysis as this is the file format used by the Starstim 3 device to store the EEG recordings conducted, hence an appropriate tool in the form of MATLAB was required to manipulate this data. Although the conversion of an .xdf file is likely possible for the sake of using the EEG recordings in an alternative programming language, this could have exposed the data to the risk of being corrupted or inaccurately translated, making the conclusion to this thesis invalid, hence a language that afforded the direct translation of the data recorded was required. Unlike the event related data recorded in the EEG recording session, the EEG recorded required a specific tool to manipulate the data. Named "EEGLAB", this interactive MATLAB toolbox provided the necessary tools for processing EEG data. As discussed, EEGLAB enabled the importing of the EEG recorded in its original format, EEGLAB also provided functions that were required in analysing the EEG data (See Table 5.1).

Table 5.1: EEGLAB (MATLAB) Functions used in EEG Data Analysis.

Function	Purpose	Use
`pop_select`	Selecting specific channels by giving the specific channel number.	Used in selecting channel 22 (Pz) to analyse potential occurrences of P300.
`pop_eegfiltnew`	Apply a digital filter to EEG data to remove unwanted frequencies such as noise or artifacts.	Used in filtering EEG data between 0.2 and 10 Hz (Zhang et al., 2023).

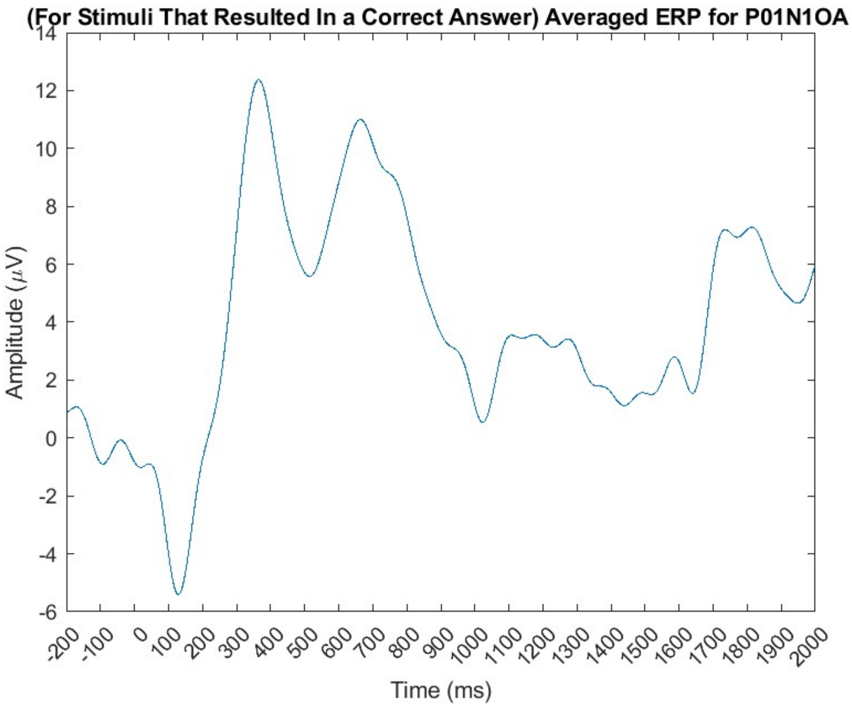

Figure 5.2: An example of an ERP plot generated by MATLAB, this plot is the averaged event related potential for Participant 1's 1-Back, OA recording. This plot displays neural responses to stimuli that resulted in a correct answer. The curves represent averaged EEG epochs, time-locked to the onset of the stimulus (0ms) and filtered to remove noise. Each epoch was baseline corrected using the pre-stimulus period (-200 to 0ms). Peaks such as the P300 at 300ms was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

The ‘Signal Processing Toolbox’ in MATLAB, enabled the presentation of the EEG data once processed correctly. This toolbox enabled plots of event related potentials to be made.

Specific to the statistical analysis of the thesis, MATLAB provided a Wilcoxon Signed Rank Test in the form of the `signrank` function. The Wilcoxon signed rank test was the statistic test chosen for the data collected, MATLAB providing this operation in the form of a function enabled repeatable and appropriate quality code to be generated in a timely and concise manner. In addition to this, with code being made repeatable with little effort, this made error prevention a lot easier, as instead of writing multiple lines or sections of code that all conduct the same purpose, a centralised body of code could be produced, enabling the source of errors to be more easily identifiable.

5.3 Data Processing

This section provides a step-by-step account of the data cleaning process, ensuring reproducibility. This section will elaborate on the data transformation techniques of the event data and EEG data in detail.

5.3.1 Event Data Analysis

5.3.1.1 Importing Event Data

Firstly, each event and their specific timestamps are extracted from their own unique EEG recording / .xdf file. The experiment that took place, required 96 separate sets of EEG recordings to be created, each set represented a specific combination of:

- The Identification Number of the Participant.
- The N-Back Type Used.
- The Stimuli Type Used (In acronym form)

For Example:

  - ‘P01N1OA’: Participant 1 - 1-Back - Outline Abstract Stimuli
  - ‘P12N2SR’: Participant 12 - 2-Back - Solid Resemblance Stimuli

5.3.1.2 Calculate Totals and Accuracies of Each Recording

For each set of data (e.g., P01N1OA), the following data was calculated:

- The number of correct answers submitted.
- The accuracy/percentage of correct answers scored.

The following formula was used:

$$\frac{Amount\ of\ Correct\ Answers\ Submitted}{Total\ Amount\ of\ Correct\ Answers\ Possible} \times 100 \tag{1}$$

Each experiment had 210 stimuli shown, 50 of the stimuli being the target of the 1-Back or 2-Back being conducted.

5.3.1.3 Calculate Reaction Times

For each set of data (e.g., P01N1OA), the “Reaction Time” for each correct answer is calculated. This is done using the following formula:

$$\frac{(Time\ of\ Answer - Time\ of\ Latest\ Stimulus\ Presentation)}{Sampling\ Rate} \tag{2}$$

The reaction time is calculated by taking the time recorded when the answer was submitted and subtracting the time recorded when the latest stimulus to be shown was presented, this total is then divided using the sampling rate.

The sampling rate is the frequency at which the EEG recording device records data points per second. The sampling rate of the EEG headset is 500 Hz (Neuroelectrics, 2020), meaning that 500 samples are recorded every second, this is equivalent to one sample every 0.002 seconds. By dividing the difference between the time of answer and the time of stimuli presentation, the participants’ reaction time is revealed, the reaction time’s intelligibility is revealed by dividing by the sampling rate as the raw data does not suffice for data analysis while the raw data converted into seconds/milliseconds via this division allows for any meaningful differences to be observed thus justifying why this operation is a necessity. The reaction times were calculated and grouped for each participant’s individual N-Back test and stimuli type combination. For each combination, the reaction times for the correct answers were calculated separately.

An example of the sets produced:

- “P01N1OACorrectReactionTimes”
- “P02N1OACorrectReactionTimes”
- “P03N1OACorrectReactionTimes”
- “P04N1OACorrectReactionTimes”
- “P01N2OACorrectReactionTimes”
- “P02N2OACorrectReactionTimes”
- “P03N2OACorrectReactionTimes”
- “P04N2OACorrectReactionTimes”
- “P01N1SRCorrectReactionTimes”
- “P02N1SRCorrectReactionTimes”
- “P03N1SRCorrectReactionTimes”
- “P04N1SRCorrectReactionTimes”
- “P01N2SRCorrectReactionTimes”
- “P02N2SRCorrectReactionTimes”
- “P03N2SRCorrectReactionTimes”
- “P04N2SRCorrectReactionTimes”

5.3.1.4 Calculate Average Reaction Times

For each set of reaction times calculated, the average reaction time was calculated. Each average reaction time is then grouped with other averaged reaction times based on their N-Back type and stimuli type. So as a result, a set would be comprised of each participants average reaction time for that specific combination. These sets of averages are created for the purpose of conducting statistical analysis.

For example:

$$\{\mu(\mathrm{P01N1OACorrectReactionTimes}), \mu(\mathrm{P02N1OACorrectReactionTimes}), \ldots \quad (3)$$

An example of the sets produced:

- “AverageN1OACorrectReactionTimes”
- “AverageN2OACorrectReactionTimes”
- “AverageN1ORCorrectReactionTimes”
- “AverageN2ORCorrectReactionTimes”
- “AverageN1SACorrectReactionTimes”
- “AverageN2SACorrectReactionTimes”
- “AverageN1SRCorrectReactionTimes”
- “AverageN2SRCorrectReactionTimes”

5.3.1.6 Grouping Accuracy Values

The values describing each participants accuracy for correct answers are grouped based on their N-Back Type and Stimuli Type. As a result, each group would be comprised of each participant's accuracy for that specific combination.

The sets produced:

- AllN1OACorrectAnswerAccuracy
- AllN2OACorrectAnswerAccuracy
- AllN1ORCorrectAnswerAccuracy
- AllN2ORCorrectAnswerAccuracy
- AllN1SACorrectAnswerAccuracy
- AllN2SACorrectAnswerAccuracy
- AllN1SRCorrectAnswerAccuracy
- AllN2SRCorrectAnswerAccuracy

5.3.1.7 Statistical Analysis

The statistical analysis of the event data saw the Wilcoxon Signed Rank Test applied to comparisons of two sets of data. Groups of data were compared based on their stimuli type and N-Back type. The following testing scheme (Table 5.2) was used to compare the reaction times and accuracies of each stimuli type and N-back combination.

As some groups would have less values then others, the "1-Back and 2-Back" comparisons would entail the 1-Back and 2-Back data from each stimuli type to be averaged together, in order to avoid an inaccurate result when comparing with another stimuli type.

5.3.2 EEG Analysis

5.3.2.1 Finding the stimuli numbers that resulted in correct answers.

The EEG recording involved timestamps with labels being created in the event of any of the following events:

- Stimuli being presented alongside the number of the stimulus.
- The Interstimulus Interval.
- A Correct Answer being given.

Table 5.2: The comparisons made during the data analysis (Group 1 being compared to Group 2 in each instance of the N-Back Task).

N-Back Type	Stimuli Type	
	Group 1	Group 2
1-Back	Outline Resemblance	Outline Abstract
2-Back	Outline Resemblance	Outline Abstract
1-Back and 2-Back	Outline Resemblance	Outline Abstract
1-Back	Solid Resemblance	Outline Resemblance
2-Back	Solid Resemblance	Outline Resemblance
1-Back and 2-Back	Solid Resemblance	Outline Resemblance
1-Back	Solid Resemblance	Outline Abstract
2-Back	Solid Resemblance	Outline Abstract
1-Back and 2-Back	Solid Resemblance	Outline Abstract
1-Back	Solid Abstract	Outline Resemblance
2-Back	Solid Abstract	Outline Resemblance
1-Back and 2-Back	Solid Abstract	Outline Resemblance
1-Back	Solid Abstract	Outline Abstract
2-Back	Solid Abstract	Outline Abstract
1-Back and 2-Back	Solid Abstract	Outline Abstract
1-Back	Solid Resemblance	Solid Abstract
2-Back	Solid Resemblance	Solid Abstract
1-Back and 2-Back	Solid Resemblance	Solid Abstract

For each EEG recording session, a search is made for each instance when a correct answer was given. Once the correct answer event has been found, the latest preceding event that details a stimuli number is recorded. This event represents the stimuli number that caused the correct answer to be made. Knowing the stimuli numbers that caused each of the correct answers made, enables the analysis of P300, a marker presumed to occur in the presence of a target stimulus.

5.3.2.2 Creating ERPs for Statistical Analysis.

An ERP is calculated for each EEG recording session. An ERP is a set of data that describes the average EEG data recorded during a specific event.

Each ERP is averaged using epochs. In the case of this study, an epoch is a set of data that represents a specific time frame in which a stimulus that resulted in a correct answer, was shown. For example, if during recording session, “P01N1OR”, Participant 1 gave 50 correct answers, the ERP for the correct answers made during P1N1OR would be the average of 50 epochs.

In order to create each epoch, the following operations are performed on the EEG data.

5.3.2.2.1 Selecting Specific Channels of EEG

In this study, the EEG channel “Pz” is being used for the purpose of analysing P300, previous studies have deemed this channel appropriate for analysing this ERP component (Zhang et al., 2023). Because of this focus, the data from the other 31 channels recorded were not used in this study. In EEGLAB (MATLAB), using the pop_select function, channel 22 (Pz) is retained as the only channel to be used from the EEG data.

5.3.2.2.2 Filtering EEG

In MATLAB using the pop_eegfiltnew function, a filter is used to remove unwanted noise from the EEG data with the aim to extract all data that will be relevant to the analysis of P300. For P300, the recommended settings of 0.2 (High-pass) and 10 Hz (Low-pass) are used to discard all other EEG data within the sample (Zhang et al., 2023).

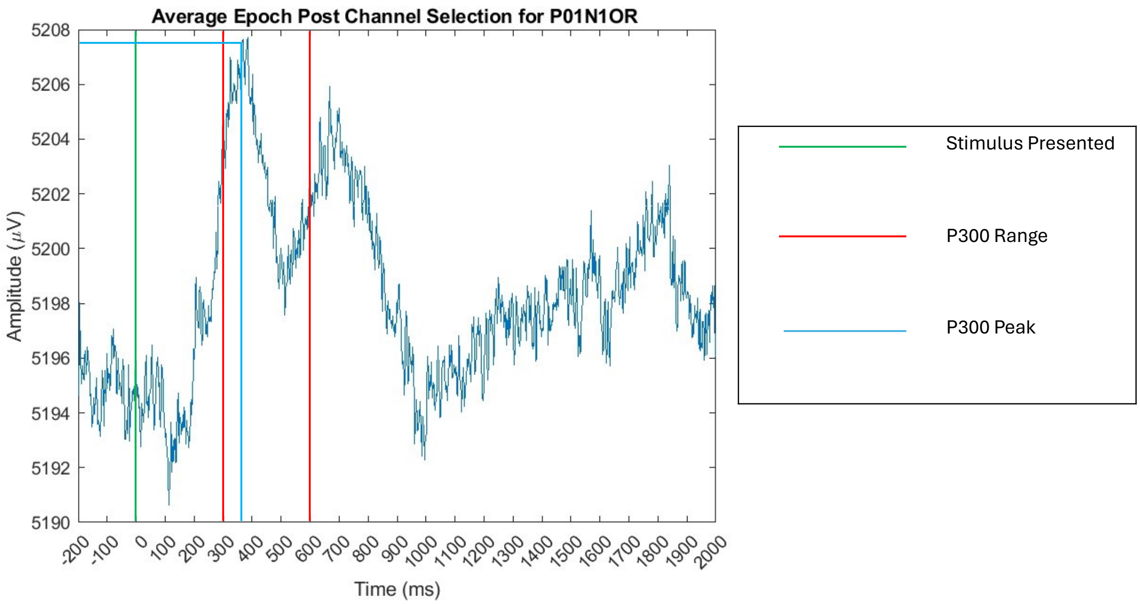

Figure 5.3: An example of a participants average epoch post channel-selection. This diagram represents the data of channel 22 (Pz) i.e. after the "*pop_select*" function removes the other 31 channels of EEG. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation), within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. This waveform was generated by averaging EEG data time-locked to the presentation of a stimulus, this data is yet to be filtered, and baseline corrected. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

5.3.2.2.3 Defining the Time Window of epoch

Each epoch will have a time window that surrounds a stimulus that caused a correct answer. The purpose of this is to measure P300, an ERP component revealed after a target stimulus is shown. To conduct baseline correction on the epoch, the epoch window start time was set to 200 milliseconds before the event (Zhang et al., 2023). The baseline was then used to adjust the post stimulus EEG.

The epoch window end time was set to two seconds after the event to ensure all potentially useful data was retained. This decision was made as the maximum time a stimulus was presented during the N-Back task was 1.5 seconds. In addition to this, retaining data taken up to two seconds after the stimuli presentation would enable P300 to be observed, this component is observed 300 to 600 ms after a participant sees the target stimulus.

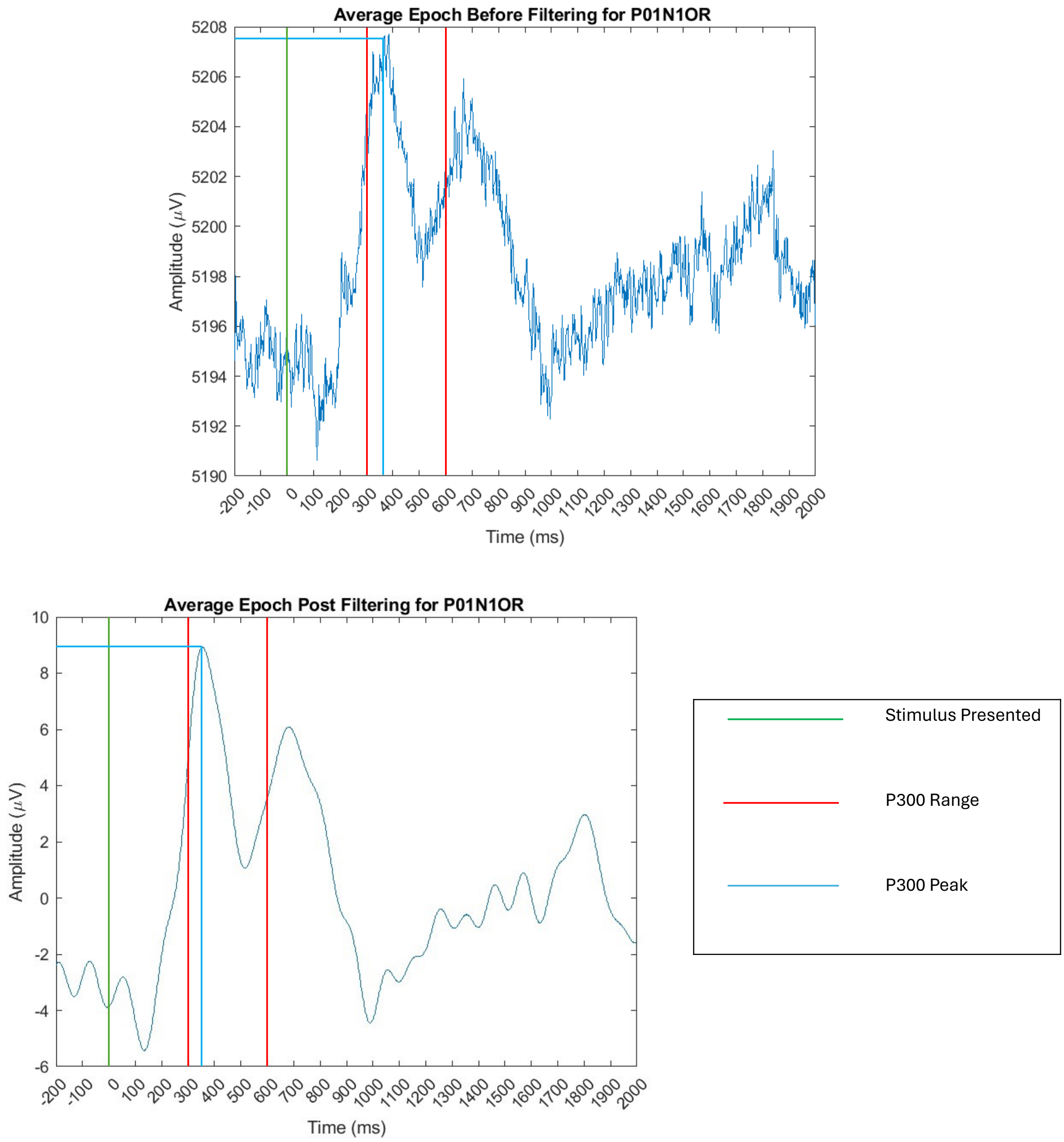

Figure 5.4: An example of a participant's average epoch before and after filtering the EEG data with a 0.2-10hz filter. The 1st plot represents the data before it is filtered, and the 2nd plot represents the data after it is filtered. The data was processed using a bandpass filter that preserves all frequencies between 0.2-10hz. This operation is executed by the "*pop_eegfiltnew*" function provided by EEGLAB in MATLAB. This filter ensures that noise from sources such as electronic devices and eye blinks, is removed from the data, leaving only the neural activity relevant to this study (as seen in the resulting 2nd plot). The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation), within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

5.3.2.2.4 Baseline Correction

After the time window of the epoch is defined, baseline correction is conducted. This entails adjusting the epoch by removing the average signal (baseline) calculated over the pre-stimulus time period. In this instance the pre-stimulus time period being used is 200 ms before the stimulus to the stimulus presentation time itself. To start, the epoch is extracted from the EEG recording, a segment that spans from 200 ms before the stimulus to 2 seconds after the stimulus is created. Then the average value of the EEG signal during the baseline period (-200 ms to 0 ms) is calculated (Zhang et al., 2023). This mean value is then subtracted from the epoch. Finally, the corrected epoch is then stored with other epochs relative to the specific EEG recording they originate from (e.g. Epochs of recording P01N1OR). The operation of baseline correction has been conducted as it helps to adjust the brain data to variations pre-existing before the stimulus. In the case of this study, it allows the researcher to observe the occurrence of P300 and its magnitude by giving the post stimulus segment of the epoch more clarity to the response. After baseline correction has been conducted on all epochs, the average of all epochs is calculated, producing the ERP to be analysed for P300. This is done for each recording session.

5.3.2.3 Find and Group Max Positive Peaks and Matching Latencies

Now that ERPs for each respective EEG recording session have been calculated, the maximum positive peaks within the P300 range (300 to 600ms post-stimulus) are calculated as well as their latencies. The measurement and comparison of the P300 amplitude and latency is necessary as this data correlates with attentional and memory related performance as discussed in previous work (Scharinger et al., 2017; Ren et al., 2023; Luck & Kappenman, 2013).

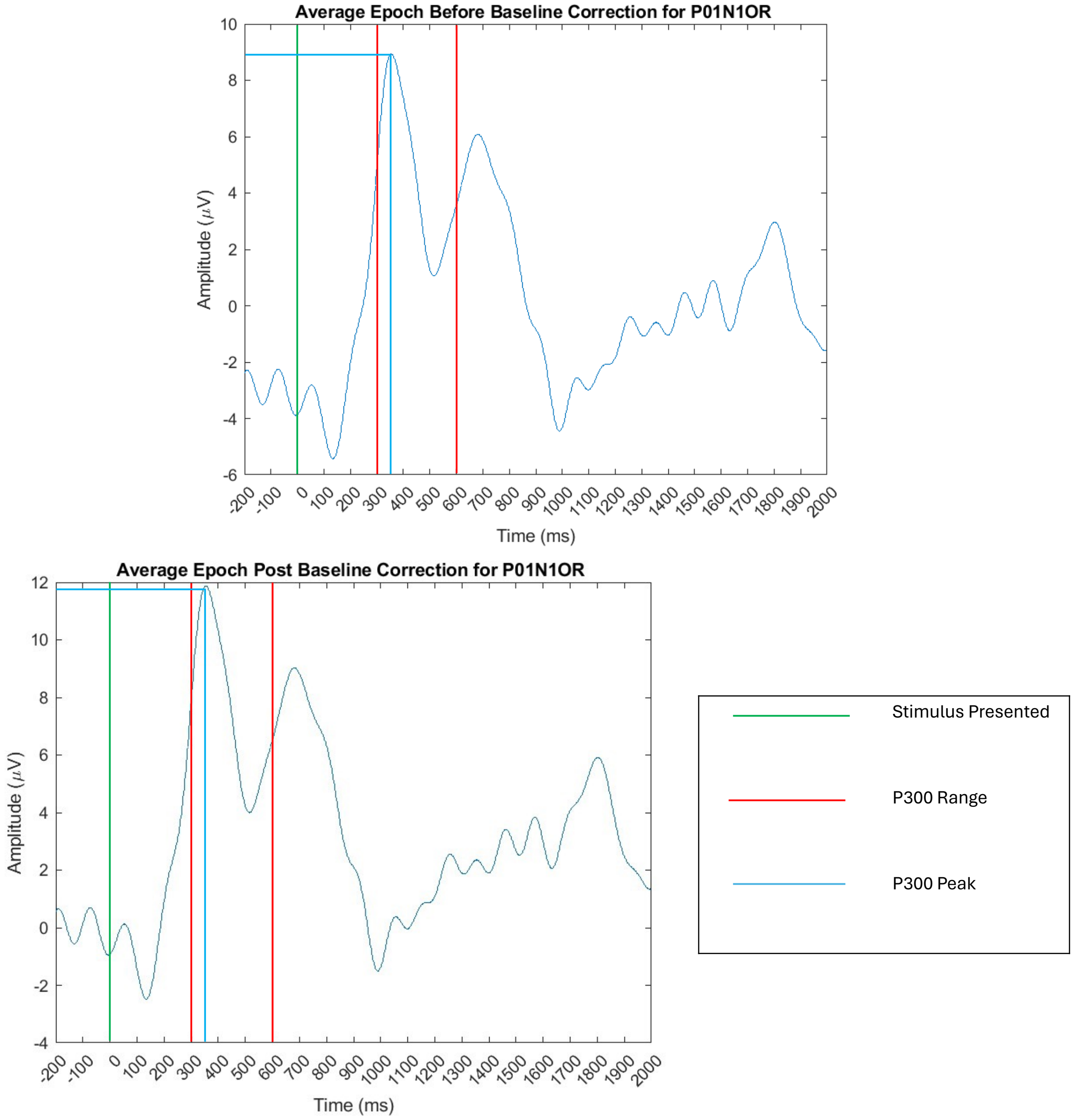

Figure 5.5: An example of a participant's average epoch before and after baseline correction. The 1st plot represents the data before it is corrected, and the 2nd plot represents the data after it is corrected. Baseline correction was conducted by subtracting the average EEG signal in the pre-stimulus period (-200 to 0ms relative to the presentation of the stimulus) from the epoch. This process ensures that the resulting waveform (i.e. the 2nd plot) accurately reflects the participants neural responses to the stimulus, by removing any voltage offsets present before the presentation of the stimulus. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation), within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

The maximum P300 positive peak and matching latency of each ERP is calculated by extracting all peaks within the P300 range of the epoch, then selecting the highest peak (i.e. the highest amplitude) recorded as well as the time it occurred within the P300 range (latency). This operation is completed using the `findpeaks` function on the segment of the epoch within P300 range, this results in a set of 'peaks'. Using this resulting set, the `max` function is used to reveal the highest of the resulting peaks. Once calculated, the maximum positive peak (i.e. amplitude) and its latency are grouped based on their relative EEG recording characteristics.

For example, for each ERP that describes 'N1OR' stimuli that resulted in a correct answer, a max peak amplitude is calculated. These peaks are grouped together resulting in a set of 12, representing 12 max peaks from the recording sessions of 12 participants. The same is done for the latencies of these amplitudes and so on for all other stimuli and N-Back combinations.

5.3.2.4 Statistical Analysis

Matching the analysis made on the event data, the statistical analysis of the EEG data used the Wilcoxon Signed Rank Test. During this, each group of data was compared based on their stimuli type and N-back type. The testing scheme in Table 5.2 was used to compare the max positive peak amplitudes and accompanying latencies of each stimuli type and N-back combination.

5.4 Wilcoxon Signed Rank

This thesis employs the Wilcoxon signed rank test to compare the event data and EEG data recorded under different conditions (N-Back type and Stimuli type). The use of this test entails comparing sets of data that are relative to each other to observe the impact of varying conditions. For example, a comparison of the average reaction times is made between the Outline Resemblance stimuli and the Solid Resemblance stimuli for the 1-Back test. In this case the icons form is the varying condition, whilst the two sets of data are related through the icon semantic and N-Back type used. Because of this, the Wilcoxon signed rank test is ideal for this use case as it is used to compare two related or matched samples. The test is also non-parametric making it appropriate to the data in this study as the analysis uses a small population i.e. each data group compared had a population of 12 (Xia, 2020), hence normality is not tested.

5.4.1 Wilcoxon Signed Rank – Effect Size

To judge the magnitude of any significant difference highlighted by the Wilcoxon Signed Rank test, the effect size (Pallant, 2020, pp. 254–255; Mcleod, 2019) of the test was calculated using the following formula:

$$\frac{|Test\ Statisitic|}{\sqrt{Number\ of\ Observations}} \tag{4}$$

6. Results

6.1 Introduction

Both data types collected from the experimentation were processed and analysed. The significant results obtained from this analysis are presented in this chapter. The first part of this chapter reports the results on the reaction times and accuracy of the 1-Back and 2-Back tasks. The second part of this chapter shows the results of the EEG/ERP analysis. A Wilcoxon Signed Rank Test was conducted on the following groups of data:

- Average Reaction Times to Stimuli that caused a Correct Answer.
- Average Accuracy of Correct Answers
- Max Peak Amplitudes for Correct Answers
- Latencies of Max Peak Amplitudes for Correct Answers

  For each stimuli type, the following comparisons were made for each:
  - Outline Resemblance vs Outline Abstract
  - Solid Resemblance vs Outline Resemblance
  - Solid Resemblance vs Outline Abstract
  - Solid Abstract vs Outline Resemblance
  - Solid Abstract vs Outline Abstract
  - Solid Resemblance vs Solid Abstract

Although all comparisons in Table 5.2 were made, only the results that had significance among all comparisons are presented (**with alpha / significance level = 0.05**).

6.2 N-Back Reaction Times and Accuracy - Wilcoxon Signed Rank

6.2.1 Correct Answer Reaction Times - Wilcoxon Signed Rank

Significant difference was observed when comparing the reaction times of correct answers between different stimuli types. During the comparison between Outline Resemblance and Outline Abstract in the 1-Back test, a significant difference was observed. The Wilcoxon Signed Rank test yielded a z-score of 2.079, making the result statistically significant with a p-value of 0.0188. The effect size (r) was 0.4243. The median reaction time for the Outline Resemblance sample was 0.4887 ms, while for the Outline Abstract sample, it was 0.491 ms.

The Outline Abstract group had significantly faster reaction times compared to the Outline Resemblance group.

When combining the 1-Back and 2-Back test data and comparing the Outline Resemblance and Outline Abstract groups again, a significant difference was observed. The z-score from the Wilcoxon Signed Rank test was 1.922 with a statistically significant p-value of 0.0273. The effect size (r) was 0.3923. The median reaction times were 0.5724 ms for the Outline Resemblance sample and 0.5469 ms for the Outline Abstract sample. The Outline Abstract group had significantly faster reaction times. Lastly, comparing the Solid Abstract and Outline Abstract groups (combining 1-Back and 2-Back test data), a significant difference was observed. The Wilcoxon Signed Rank test produced a z-score of 1.843 alongside a statistically significant p-value of 0.0326. The effect size (r) was 0.3763. The median reaction times were 0.5701 ms for the Solid Abstract sample and 0.5469 ms for the Outline Abstract sample. The Outline Abstract group demonstrated significantly faster reaction times than the Solid Abstract group. (**For Wilcoxon signed rank results for the average correct reaction times, refer to Table A.1**)

6.3 EEG – ERP Analysis - Wilcoxon Signed Rank

6.3.1 Max Peak Amplitudes of ERPs for 'Correct Answers'

Significant difference was observed when focusing on the 1-Back test, the max peak amplitudes of the Outline Resemblance and Outline Abstract stimuli were compared. These amplitudes, specifically observed within a time window of 300 to 600 milliseconds post-stimulus presentation, showed a significant difference. The Wilcoxon Signed Rank test yielded the following results: A z-score of 1.687, a p-value of 0.0458, an effect size (r) of 0.3443. The median max peak amplitude for the Outline Resemblance stimuli was 14.9783 uV. In contrast, the Outline Abstract stimuli had a median max peak amplitude of 13.9205 uV. Notably, the Outline Resemblance stimuli exhibited significantly larger max peak amplitudes than the Outline Abstract stimuli i.e. The Outline Resemblance stimuli exhibited a significantly lower memory load compared to the Outline Abstract stimuli. (**Refer to Table A.3**)

Additionally, a comparison was conducted for the 2-Back test between the Solid Resemblance and Solid Abstract stimuli. The Wilcoxon Signed Rank test revealed a z-score of -2.267, a p-value of 0.0117, an effect size (r) of 0.4628. The Solid Resemblance stimuli had a median max peak amplitude of 11.3734 uV, derived from 12 values, one of which was classified as 'NaN' – MATLAB's `findpeaks` was unable to find positive peaks for P01N2SR thus, the N2SR population of averages was padded with an "NaN" value as MATLAB would only allow a Wilcoxon Signed Rank test to be conducted with two equal populations. On the other hand, the Solid Abstract stimuli showed a median max peak amplitude of 14.01 uV, based on 12 values. The Solid Abstract stimuli had significantly larger max peak amplitudes than the Solid Resemblance stimuli i.e. The Solid Abstract stimuli exhibited a significantly lower memory load compared to the Solid Resemblance stimuli. (**Refer to Table B.3**).

Lastly, a comparison was conducted for the 2-Back test between the Solid Resemblance and Outline Resemblance stimuli. The Wilcoxon Signed Rank test revealed a z-score of -2.356, a p-value of 0.0092, an effect size (r) of 0.4809. The Solid Resemblance stimuli had a median max peak amplitude of 11.3734 uV, derived from 12 values, one of which was classified as 'NaN' – MATLAB's `findpeaks` was unable to find positive peaks for P01N2SR thus, the N2SR population of averages was padded with an "NaN" value as MATLAB would only allow a Wilcoxon Signed Rank test to be conducted with two equal populations. On the other hand, the Outline Resemblance stimuli showed a median max peak amplitude of 11.7179 uV, based on 12 values. It's noteworthy that the Outline Resemblance stimuli had significantly larger max peak amplitudes compared to the Solid Resemblance stimuli i.e. The Outline Resemblance stimuli exhibited a significantly lower memory load compared to the Solid Resemblance stimuli. (**Refer to Table B.3**).

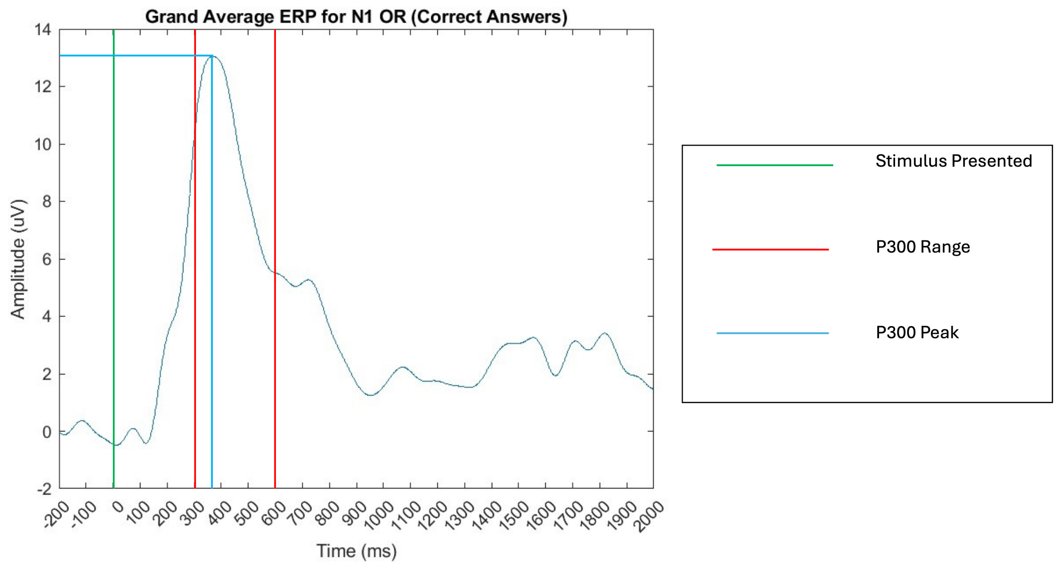

Figure 6.1: The grand average event related potential for outline resemblance stimuli that resulted in a correct answer during the 1-back task. This plot is an average of every instance where a participant submitted a correct answer, after witnessing an outline resemblance stimulus during their respective 1-Back tasks. Each instance was represented by an epoch that was used in the averaging. Each epoch was filtered (0.2-10hz) and baseline corrected using the pre-stimulus period (-200 to 0ms) before the epochs were averaged together. The P300 peak was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation). Within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

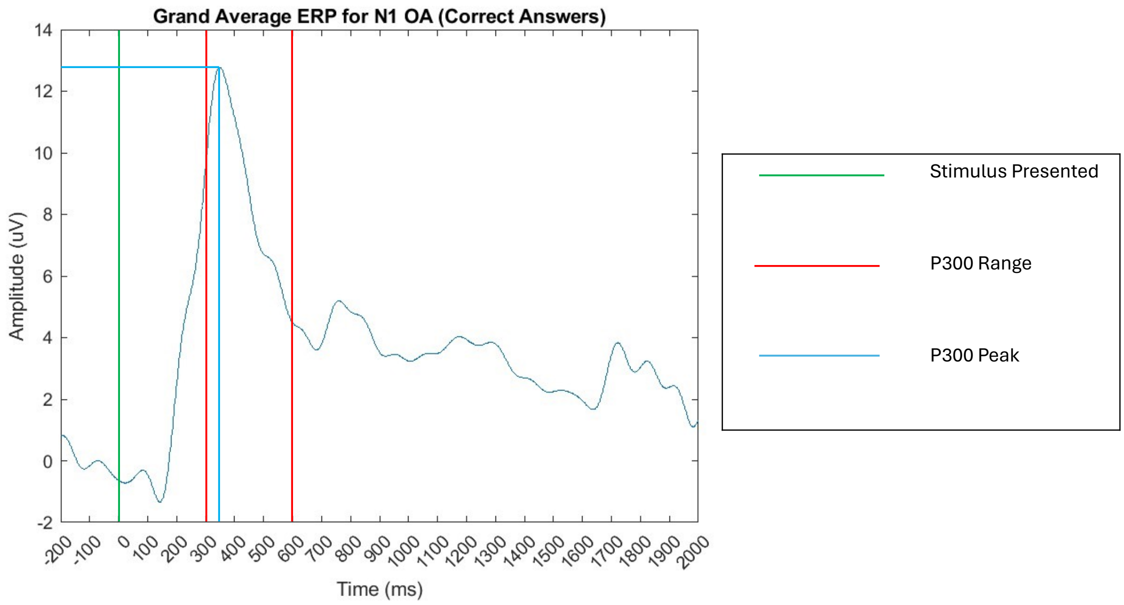

Figure 6.2: The grand average event related potential for outline abstract stimuli that resulted in a correct answer during the 1-back task. This plot is an average of every instance where a participant submitted a correct answer, after witnessing an outline abstract stimulus during their respective 1-Back tasks. Each instance was represented by an epoch that was used in the averaging. Each epoch was filtered (0.2-10hz) and baseline corrected using the pre-stimulus period (-200 to 0ms) before the epochs were averaged together. The P300 peak was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation). Within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

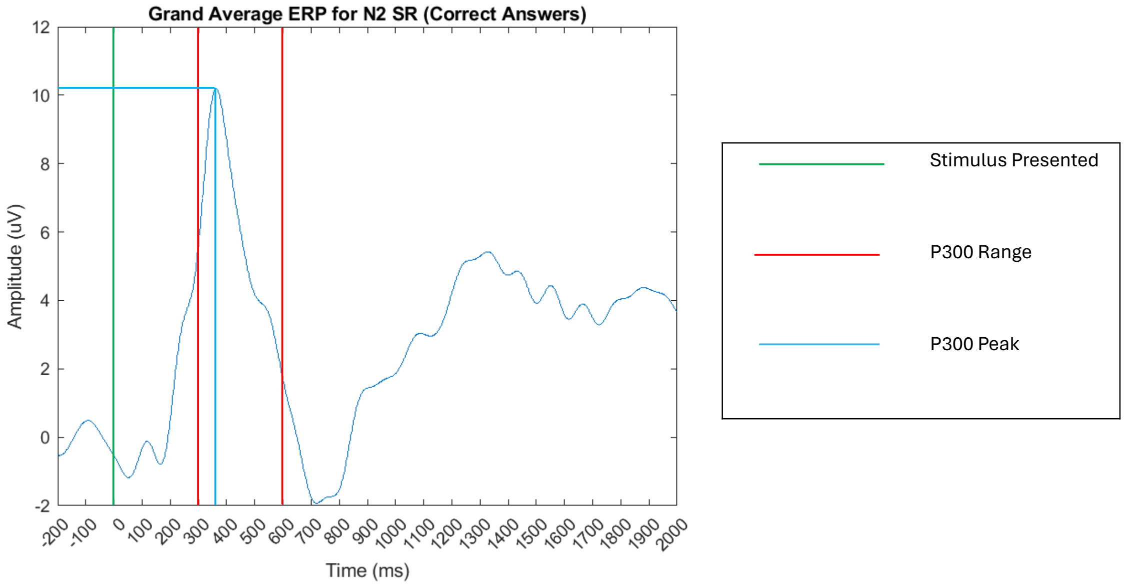

Figure 6.3: The grand average event related potential for solid resemblance stimuli that resulted in a correct answer during the 2-back task. This plot is an average of every instance where a participant submitted a correct answer, after witnessing an solid resemblance stimulus during their respective 2-Back tasks. Each instance was represented by an epoch that was used in the averaging. Each epoch was filtered (0.2-10hz) and baseline corrected using the pre-stimulus period (-200 to 0ms) before the epochs were averaged together. The P300 peak was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation). Within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

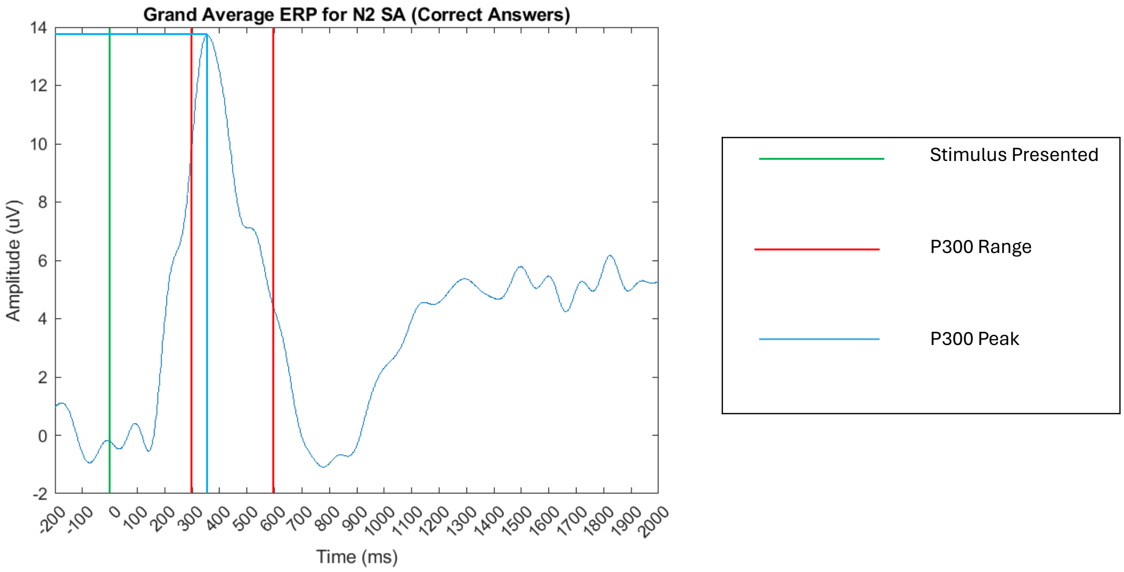

Figure 6.4: The grand average event related potential for solid abstract stimuli that resulted in a correct answer during the 2-back task. This plot is an average of every instance where a participant submitted a correct answer, after witnessing an solid abstract stimulus during their respective 2-Back tasks. Each instance was represented by an epoch that was used in the averaging. Each epoch was filtered (0.2-10hz) and baseline corrected using the pre-stimulus period (-200 to 0ms) before the epochs were averaged together. The P300 peak was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation). Within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

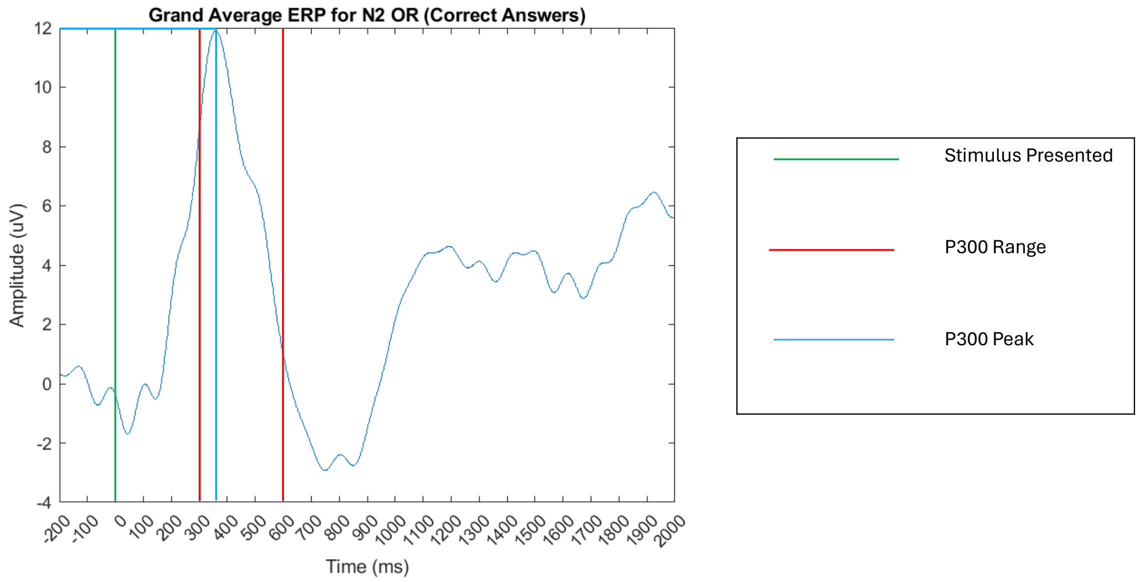

Figure 6.5: The grand average event related potential for outline resemblance stimuli that resulted in a correct answer during the 2-back task. This plot is an average of every instance where a participant submitted a correct answer, after witnessing an outline resemblance stimulus during their respective 2-Back tasks. Each instance was represented by an epoch that was used in the averaging. Each epoch was filtered (0.2-10hz) and baseline corrected using the pre-stimulus period (-200 to 0ms) before the epochs were averaged together. The P300 peak was calculated by identifying the highest positive amplitude within a 300-600ms window post stimulus presentation. These trends correspond to the increase and decrease of the participants memory load. The green line indicates the onset of the stimulus presentation. The red lines indicate the P300 time window (300-600ms post stimulus presentation). Within this window is where the P300 peak usually occurs. The blue line indicates the P300 peak, i.e. the highest positive amplitude within 300-600ms, post stimulus presentation. The x-axis represents time in milliseconds and the y-axis represents amplitude in microvolts.

7. Discussion

7.1 Introduction

This section will act as an overview and interpretation of the key findings. This section will also summarise the implications of these findings in relation to the three hypotheses. Given the results interpreted, the results will be compared to the prior work of others, explaining similarities and differences, helping to define the work of this thesis in relation to the wider history of work. This project's limitations will also be discussed, establishing any inefficiencies this study has had and how those limitations can be improved on in future work. The discussion on future work will also include the identification of key areas where additional research is needed or could be provided for a particular purpose.

7.2 Hypothesis 1

Hypothesis 1 predicted that one of Arledge's classifications (Solid or Outline) will perform significantly better than the other in memory performance i.e. The speed at which the icon is recalled from memory. This hypothesis was designed to follow up on the work of Arledge, specifically in finding a style used by Arledge that performed significantly better than the other in memory performance – Arledge's study concluded in a "winner" between the two not being declared. Nielsen's classifications were used in conjunction with Arledge's Icons to produce a different result.

Given each comparison conducted between Solid and Outline stimuli regardless of the accompanying Nielsen semantic (Resemblance or Abstract), a conclusion on the hypothesis has been reached. All comparisons between Solid and Outline stimuli (regardless of semantic or N-Back used) produced a statistically insignificant result, except for one comparison – The reaction times of correct answers for Solid Abstract and Outline Abstract stimuli (Both N-Back Combined). Although this comparison returned a result that would suggest the difference between the two stimuli groups is significant, this comparison alone is not sufficient evidence to support hypothesis 1. As a result of neither icon style producing a substantial number of significant results that would highlight that specific style as the superior influence on memory performance, the conclusion reached on this hypothesis is that neither style used by Arledge impacted memory performance significantly better than the other.

As solid icons were not recalled significantly faster than outline icons, this reflects Arledge's findings of there being no superior icon style between the two. Despite this result, the current study is exploratory as a result of the limited number of participants available and consequentially focussed demographic; all participants were students from the University of Kent. Due to its exploratory nature and difference in the number of participants used compared to Arledge (2014), the results should be viewed as an initial contribution to a larger future study, rather than a complete replication of Arledge (2014). The results offer insight into the memory performance of Arledge’s icon styles and Nielsen’s (2014) semantic classifications, however based on the participant size of the current study, this insight should be built upon in future studies more comparable in scale to Arledge (2014), in order to improve the reliability of the current study’s results.

In the literature review, it was discussed that articles had been published around Arledge’s study, but did not review or question the paper’s reliability. Due to the current study’s use of practical means to reach its findings, the result of hypothesis 1 allows the audiences of these articles to be more accurately informed.

Moreover, despite the results being refutable due to the significantly low number of participants used compared to Arledge (2014), this study is valuable as it has introduced a new source of primary data to the work of Arledge (2014). This challenges claims made in the articles written in response to Arledge (2014) that do not feature any primary data (liu, 2017; uxmovement, 2019; KrishaWeb, 2020; Wake, 2021).

When focussing on this study’s contribution, this current study expanded Arledge’s original paper by incorporating Nielsen's semantic classifications when designing the study’s stimuli. The inclusion of Nielsen’s semantic classifications served the purpose of examining how the different classifications interacted with Arledge's icon styles in terms of the speed at which each style was recalled from memory. Despite the results not showing any major significant difference between Arledge’s styles, this study still breaks new ground by introducing a new dimension of icon characteristic in the form of Nielsen’s classifications, thus it is advisable to continue this line of investigation in future work to see how a varying secondary design feature can impact the memory performance of Solid and Outline stimuli.

Exploring the result of hypothesis 1 further, previous studies provide perspective on the rejection of hypothesis 1. Jin (2020) challenges the rejection of hypothesis 1 and the conclusion that an icon's style has no effect on the participants that took part in this study.

Jin (2020) found that in an eye tracking experiment using skeuomorphism and flat icons, the style of icon had an impact on the accuracy and efficiency of user cognition, with skeuomorphism icons having better accuracy in user cognition and flat icons having the better efficiency of user cognition.

Skeuomorphism Icons are based on "Skeuomorphism" which is defined as a design approach that imitates real life objects by making a design look 3D by using design components such as texture and colour to make the design / icon look more realistic (Dovetail Editorial Team, 2023). Moreover, flat icons are defined as icons that are in 2D (Mu et al., 2022).

As cognition encompasses memory (Mak, 2015) and the rate of recall from memory as well as recall itself relates to memory (The Human Memory, 2019), this makes Jin's result directly relevant to this study as hypothesis 1 explored the recall speeds of each of the stimuli types.

Jin's result in relation to this current study suggests that integrating different icon styles such as 2D and 3D icons might provide a different result when measuring the speed at which varying icon styles are recalled from memory. The application of varying dimensions to this study's current set of stimuli is valid as Jin's conclusion might suggest that Arledge's Solid and Outline icons, both 2D in presentation, are perceivably not different enough from each other, this would explain why no substantial significant difference was observed when comparing how quickly each style was recalled nor when comparing how many instances each style was successfully recalled. Because of this, future work should explore if applying varying dimensions to Arledge's icons can lead to a significant difference in recall speed not yet observed in this study. Overall, Jin's study represents an opportunity to improve on this current study through its use of design choices novel to the work of Arledge.

Mu et al. (2022) supports this direction of future work as their study explored the impact of icon depth on user attention and search efficiency, comparing flat (i.e. 2D) icons and neumorphic (i.e. 3D) icons. Neumorphic icons are a 3D style of icon inspired by Skeuomorphism, Mu et al. (2022) makes the distinction between the two by referring to the neumorphic style as containing less detailed information compared to skeuomorphic icons. The study found that 3D icons generated more visual attention than 2D icons.

While this study does not directly investigate the Solid and Outline styles used in this study, it does imply that certain design aspects, such as the icon's dimensions, has an impact on visual attention, which is correlated with recall efficiency, i.e. the retrieval of memory (Chun & Turk-Browne, 2007).

Mu et al. (2022) reinforces this study's conclusion that future work should explore varying dimensions of icon design. Mu's observation of 2D and 3D icons eliciting different levels of attention suggests the question - Could the significant difference in visual attention observed between 2D and 3D icons promote a significant difference in the memory performance if varying dimensions were applied to Arledge's Solid and Outline icons? Applying alternative dimensions to Arledge's icons might highlight significances not yet seen in the work of Arledge.

To summarise, Mu et al., (2022) highlights the impact of icon depth on user cognition. This results potential for significant difference in cognition could be applied to the future work of this study in an attempt to observe significant difference in the cognitive performance of Arledge's icons, specifically in their recall speeds.

Moreover, these papers that relate to hypothesis 1 and its rejection (Jin, 2020; Mu et al., 2022) imply that this current study would benefit from using a wider range of icon dimensions in order for significant differences in memory performance to be observed. Although revealing a flaw in this study's methodology, this implication only contributes to the field of icon recognition and Human Computer Interaction as it provides potential insight into the result of this hypothesis, providing an alternate way in which the current study can be expanded on in future work, but to conclude, Hypothesis 1 argues that the usability of Arledge's two styles are equal as the speed at which both were recalled from memory are equal with the exception of one test - The reaction times of correct answers for Solid Abstract and Outline Abstract stimuli (Both N-Back Combined).

7.3 Hypothesis 2

Hypothesis 2 predicted that Nielsen's resemblance icons would perform better than Abstract Icons based on their ability to be associated with real world objects. Again, this prediction was made specific to memory performance i.e. reaction times, predicting that resemblance icons would produce faster reaction times than abstract icons. This hypothesis was designed to test Nielsen's claim on resemblance icons having the best usability. The results produced by this study reject this hypothesis. Overall Resemblance icons and Abstract icons failed to significantly outperform each other; therefore, this study rejects Nielsen's original statement on the "superior" usability of resemblance icons as well as Hypothesis 2. This rejection is justified by the observed result that resemblance icons failed to significantly outperform abstract icons in each and every instance where the two semantic classifications were compared to each other.

Linking this finding to the literature review, this insight is significant especially as it represents the inverse of Nielsen's original statement, claiming that, resemblance icons have the best usability. Secondly, with Nielsen directly naming memorability as one of five quality components of usability (Nielsen Norman Group, 2012), i.e., something resemblance icons are described as usually having better than abstract icons, hypothesis 2's result directly challenges Nielsen's conclusion.

Regarding the reliability of the results, each permutation of icon style and semantic was compared, making the rejection of Hypothesis 2 clear. This is a result of the resemblance icons failing to significantly outperform abstract icons in every instance where the two semantic classifications were compared to each other. The study's comparison of resemblance icons and abstract icons substantiates its rejection of Nielsen's (2014) conclusions. However, this result does not consider icon design components such as colour and size and how they might have impacted this result. In account of this, it is appropriate to label the rejection of hypothesis 2 as tentative given the study's specific focus on icon style and semantic. Despite this, the result still reflects a conclusion valuable to the field of iconography research, as unlike Nielsen (2014), this current study contributes towards the understanding of icon semantics used through the contribution of primary data.

In summary, the results enable this study to reject the conclusion of Nielsen (2014) because it puts Nielsen's ideas into practice. However, the rejection of hypothesis 2, despite being quantitatively supported in the scope of the current study, does lack robustness when applying this result to icon design features not used in this study. This lack of robustness could be rectified through the implementation of more icon design features such as size and colour in future work. Despite this, the current study's results still represent an instance where Nielsen's claim on the superiority of resemblance icons is false, thus Hypothesis 2 is rejected.

This result provides more quantitative data for professionals in the HCI and iconography fields to base their design decisions on. To extend this contribution and ensure this result's reliability, future research should challenge Nielsen's statement further.

Linking this results potential contribution to previous work, Ng, A. W. Y., & Chan, A. H. S. (2008) uses classifications which have a strong similarity to the classifications of Nielsen. This study's "Concrete" icons depict real objects and materials, whilst icons that do not depict real objects are described as "Abstract". Based on the descriptions provided by Ng and Chan, their "features" and Nielsen's classifications could be considered the exact same concept.

Ng and Chan argue that their review on these classifications provides graphic designers with a clearer understanding of these classifications, allowing them to create more accessible and effective icons (p. 1). This is a clear contribution to the field of icon design and Human Computer Interaction as more accessible and effective design is likely to improve the usability of the icons as well as the user experience the icons are featured in. Ng and Chan's paper is only a review and not an independent study that produces its own data in order to bolster this contribution, whilst this current study has used, tested and concluded on the usability of some of the classifications used in the review thus providing a strong argument towards this study also providing the same contribution as Ng and Chan.

A separate argument can be made that this current study continues the work of Ng and Chan by not only reviewing past work as Ng and Chan have done, but practically investigating the classifications used, allowing this current study to adopt the same contribution as Ng and Chan's work but also be considered more reliable than their review, due to this study producing its own data to back its conclusions.

Although relevant literature highlights the contribution of the results of Hypothesis 2, Nielsen's statement on resemblance icons still challenges the rejection of Hypothesis 2, however given that this study is using actual results from experimentation, perhaps this study represents the need to explore the usability of Nielsen's classifications in more nuanced ways.

For example, Nielsen's original statement on the apparent superior usability of Resemblance icons fails to mention how styling or colour could impact this usability of resemblance and abstract icons. In addition to this, the argument can be made that although Nielsen's statement challenges the result of this hypothesis, Nielsen's statement lacks ecological validity as it does not describe the specific conditions in which their statement would be observable in practice. Taking this into account, as this study uses design elements beyond just semantics (i.e. varying styling, a specific colour set etc) as well as puts Nielsen's theory into practice by using an actual experimental task (N-Back), this study concludes that despite the rejection of Hypothesis 2 and Nielsen challenging the rejection of Hypothesis 2, the results of this study contributes to the work of Nielsen by challenging the ecological validity of Nielsen's work by putting their theory into practice and producing a different result. Putting Nielsen's theory into practice also benefits designers and researchers in iconography design and human computer interaction, as the proving or disproving of Nielsen's claims would allow these groups of people to make more informed decisions on which classification of icon semantic to use.

7.4 Hypothesis 3

Hypothesis 3 is rejected by default in the face of no one classification in the Arledge family or the Nielsen family showing any significant difference to each other across the max peak amplitudes or latencies thus not showing any difference between the memory loads of the various icon classifications.

Despite this, the conclusion of Hypothesis 3 is not that P300 was not exhibited under any of the classifications. On the contrary, ERP samples for "Correct Answers" from each of the N-Back tests for each stimuli type used show a clear peak in the P300 range (300 to 600 ms after the target stimuli is observed).

The rejection of Hypothesis 3 is due to their being no observable, significant difference between the memory loads of the icon classifications used, as indicated by their respective P300 performance. The ERP analysis across all EEG recordings shows a narrative of neutrality i.e. there is no major overall difference in Arledge's or Nielsen's families of icons, especially in the ERP data pertaining to the correct answers where P300 is the most likely to be present. The lack of significant difference between the icon classifications suggests that each family elicited similar levels of memory load.

This result reflects the same lack of significance in memory performance observed in Hypothesis 1, suggesting that this hypothesis might share a limitation and direction for future work proposed through reviewing Mu's and Jin's study. The limitation in question was that

Arledge's Icons did not provide a large enough variation in style, causing a lack of significance in the observed icon recall speeds, however, due to an observed difference in visual attention in Mu's study due to the use of icons that varied in dimensions, future work could further the use of varying dimensions in order to observe significance in the P300 ERP analysis as well as the aforementioned reaction times.

The relevance of this limitation and direction of future work to Hypothesis 3 is justified by past research discussed in the literature review, which states that attentional resource allocation is correlated with the P300 amplitude (Scharinger et al., 2017; Ren et al., 2023; Polich & Kok, 1995; Luck & Kappenman, 2013; Amin et al., 2015; Tao et al., 2022; Linden, 2005). In addition to this, if the use of icons with varying dimensions causes an observable difference in attention, which can be measured using P300, this would suggest that a lack of varying dimensions in this current study could be the cause of why a significant observable difference has not been observed in the comparisons of the P300 amplitudes, their latencies and by extension the memory loads of each stimuli classification as a result. Because of this, the result of Hypotheses 1 and 3 are potentially symptoms of the same limitation – a lack of stylistic variance capable of causing a significant observable difference in the reaction times or the P300 performance of the stimuli used.

To summarise, Hypothesis 3 is rejected due to the comparisons of P300, not showing any significant difference between the memory loads of the icon classifications used. This gives evidence that each stimuli classification had similar memory loads during the experiment as indicated by the P300 amplitudes. However, despite this, the result of Hypothesis 3 shows a clear similarity to the result of Hypothesis 1 which also resulted in a lack of significant difference between the stimuli used. Because of this, this study argues that the limitation and future work of Hypothesis 1 is also relevant to Hypothesis 3 as the measurement used in this case, P300, is also correlated with attention as well as memory load - Mu et al. (2022) showed significantly different levels of attentional resource allocation when comparing icons of varying dimensions. Because of this, future work could implement varying dimensions into its existing stimuli families in order to potentially trigger significant differences in attentional resource allocation, consequentially also triggering significant differences in P300 comparisons. Triggering such a variety in P300 performance could lead to new conclusions being reached on the memory loads of the stimuli families due to memory load being indicated by P300 as well as the allocation of attentional resources.

7.5 A finding related to Hypothesis 2 and 3

Despite the rejection of Hypothesis 2, the results of this study produced a finding separate to the hypotheses. Results show that Outlined Abstract icons were recalled significantly faster than Outlined Resemblance icons. This is shown when comparing the performance of the 1-Back test as well as both N-Back tests on average. Using the Wilcoxon signed rank test, the p-values on both comparisons made, are below the significance level of 0.05, showing that the difference in recall speeds between Outline Resemblance and Outline Abstract stimuli is statistically significant with the recall speeds of Outline Abstract stimuli being significantly faster.

This result is also supported by the effect size of both tests. The effect size was observed to be moderate in size, illustrating a moderately significant difference in the speed of recall between Outline Resemblance and Outline Abstract stimuli. This in comparison to a small effect size, makes this significant difference less debatable as a small effect size might make this result subject to error checking or suggested scrutiny. These results compellingly suggest that in the case of outlined icons, the 'abstract' semantic classification, improves recall speeds. This has been observed in all comparisons with this pairing, except in the 2-Back test. Moreover, in Hypothesis 3, significant difference was also observed when comparing the max peak amplitudes (P300) of Outline Abstract and Outline Resemblance stimuli in a 1-Back test. The max peak amplitudes of Outline Abstract stimuli were significantly smaller than that of Outline Resemblance stimuli. These specific differences from Hypothesis 2 and 3 suggests a potential relationship between the reaction times and amplitudes of Outline Abstract and Outline Resemblance stimuli. This relationship entails a faster reaction time with an increased memory load shown by the P300 results. Future research should take note of this finding and further investigate whether this represents a distinction among the stimuli types used or a coincidental statistical significance.

The speculation surrounding why abstract icons were recalled faster than resemblance icons alongside a higher memory load is assumed to be because resemblance icons caused the participant to make associations with memories in the real world, which distracted the participants, interfering with the process of memory recall, therefore slowing it down. Abstract icons on the other hand, do not resemble real world objects, therefore they potentially avoided this interruption in recall. Cherng et al. (2016) agrees with this speculation, whilst the use of N400 in Yang et al. (2021) provides a way of investigating this theory. Cherng et al. (2016) reported a significantly larger P3b amplitude (a subcomponent of P300) with semantically “close” i.e. Resemblance icons compared to semantically “far” i.e. Abstract icons. Cherng justifies this difference with the notion that “far” icons do not trigger the cognitive functions associated with memory as easily as “close” icons (p. 4387). Given Cherng’s stance on the matter, this suggests that abstract icons might not trigger cognitive functions associated with memory as easily, as they do not cause real world associations in memory thus avoiding the delay in recall speeds as speculated.

To investigate if this difference in recall speed is potentially due to a difference in cognitive processing, we turn to Yang (2021). Based on their study, after analysing N400, this study puts forward the argument that different levels of semantic complexity have a significant impact on user cognition. As this current study suspects that a difference in cognitive processing was due to the use of different icon semantics, thus causing significantly different recall times in OR and OA icons, could N400 be analysed to reveal and confirm this suspected difference in semantic related processing? This idea is supported by the fact that Yang (2021) directly stated that N400 is related to semantic processing (p. 171), this suggests that N400 would be appropriate for the comparison of data sets suspected to vary in processing because of their use of two different semantic classifications.

The use of N400 can also support the investigation of Outline Abstract icons having a significantly larger memory load compared to Outline Resemblance icons. The speculation surrounding this result, is that abstract icons required more cognitive processing for the encoding and retrieval of memory, due to their form being arguably harder to process as they do not resemble any real-world objects that the participants might be experienced with. As N400 has been described to be linked to semantic processing by Yang (2021), could the analysis of N400 prove that Outline Abstract icons yielded higher levels of semantic processing potentially justifying the higher memory load indicated by this study’s analysis of P300? Lastly, this current study’s 1-back results should be analysed using N400 to conclude on why an initial significant difference was observed with outline stimuli and not solid stimuli.

7.6 A finding related to the ERP analysis of OA vs OR

Despite the rejection of Hypothesis 3, a separate finding regarding the ERP analysis was found. Returning to the comparison made between the max peak amplitudes of OA and OR, in the 1-Back test, a significant difference was observed - OA had a significantly smaller max peak P300 amplitude compared to OR.

Based on previous work showing an increasing memory load reduce the P300 amplitude, (Scharinger et al., 2017, p. 3; Ren et al., 2023, p. 1; Luck & Kappenman, 2013, p. 165) this indicates that OR elicited a significantly lower memory load compared to OA. As discussed in the introduction and motivation, HCI researcher Ben Scheiderman instructed their readers to lower the short term memory load required to use an interface. This has been advised as according to Scheiderman, humans have a limited capacity for information processing in short term memory and require user interface designers to avoid creating interfaces that require users to remember information from one display and use the information in another display. This rule sets a precedent for how usability can be improved in a user interface, specifically regarding the user's short term memory.

According to Mcleod (2023), short term memory is described as a part of memory that actively holds information for a brief period (i.e. usually for seconds or a minute). In this study, the N-Back task required participants to hold previous stimuli shown in their memory in order to judge if the current stimulus matched. As the stimulus presentation time was 1500 milliseconds and the interstimulus interval was 1500 - 2000 milliseconds, this process took no longer than 7 seconds in the 1-Back task, supporting the case that participants engaged their short term memory during this experiment.

Based on the result of the comparison made, the known impact of memory load on the P300 amplitude, Scheiderman's rule and the definition of short term memory itself, the conclusion can be made that OR icons have significantly better usability than OA icons based on the fact that they elicited a significantly lower memory load when being successfully recalled from memory. This conclusion is supported by the significant difference observed between the P300 amplitudes of the two icon types.

This conclusion shows that semantic has had a significant impact on the memory load of outline icons as results show that that an icon is likely to require less memory load when recalled from memory if that icon resembles a real-life object compared to when the icon has an abstract form.

7.7 Limitations and Future Work

During Hypothesis 1 and Hypothesis 3, using Arledge's icon styles potentially led to this study's first limitation. The stylistic variance between the Solid and Outline icons produced insignificant differences in the comparisons made on the reaction times and ERP data. This limitation presents the question - Can the introduction of a more significant stylistic variety, trigger a more substantial difference in the reaction times and ERP data of the stimuli types used. As discussed in the summary of Hypothesis 1, Jin (2020) supports this direction of future work as their study showed that using 2D and 3D styles of icons had different effects on user cognition. User cognition can affect memory recall (Fitzgerald, 2016, p. 7) as well as involve attention (Cherry, 2024), which is measurable using the P300 amplitude. Mu et al. (2022) also supports this direction of future work as their results showed a significant difference in attention between 2D and 3D icons, again, highlighting that these differences can be observed using the P300 amplitude as attentional resource allocation is correlated with P300. In addition to this, attention directly impacts recall efficiency (the efficiency of the ability to retrieve information from memory (Chun & Turk-Browne, 2007)), making Mu's conclusion relevant to both Hypothesis 1 and 3.

Overall, applying varying dimensions to this study's icons, might lead to significant differences in the reaction times and ERP data being revealed. This proposed result will give researchers and designers working in the field of iconography and human computer interaction, further insight into how their design choices can impact a user's cognition.

Another limitation is the participant population size. This study, with its 12 participants, stands in contrast to Arledge's study which involved 1,260 participants. Despite methodological differences between the two studies, namely this study's use of EEG and N-Back to study Arledge's icons, this inequality in participant numbers could potentially invalidate the comparability of our findings. Future research should consider using a larger and more diverse participant pool than what the population of the University of Kent can offer. This would not only bolster the reliability of this study's findings, but also offer a more appropriately sized study to be compared with Arledge's 2014 paper. A replication of

Arledge's work using a broader participant base, could provide a clearer understanding of the impact of icon styles on usability.

In addition to this, future research could continue to integrate more variables into the methodology. This study introduced Nielsen's semantic classifications to Arledge's work, whilst future work could expand on this by including, variables such as colour, animation, or cultural differences to add more ecological validity to this body of work. Moreover, conducting studies in real-world settings such as testing the different icons in user interfaces and mobile apps, could provide an immediate and practical solution to the lack of ecological validity of testing in controlled environment such as the lab used in this study.

Future research regarding the pre-experimentation phase, could use user centered design research to gain qualitative information on the icons before they are shown to the population of participants. By identifying which icons are preferred over others, this could guide research into developing a group of icons that test the differences in style and semantic more effectively.

Focusing on this study's insignificant result, when judging this study critically, it can be argued that in a computing perspective, this study's contribution despite being novel because of its use of a specific methodology, is lacking. This could be justified by the fact that a majority of the results are insignificant thus not concluding on which icon classification is more usable and why.

Despite this, this study's results show a contribution to computing, as evidence suggests that OA had a significantly larger memory load when compared to OR, providing the conclusion that OR has superior usability based on Shneiderman (2016). This conclusion provides a unique contribution, as this result provides user interface designers with immediate guidance on how the implementation of semantics, might impact any outline styled icons included in their user interface. Moreover, if a user interface designer wishes to use outline icons, but also intends to fulfil Scheiderman's eighth rule and lower the short term memory load required to use an interface (Shneiderman, 2016), this study provides a direct answer to how that might be accomplished.

In connection to this study's purpose of better understanding the circumstances in which icons are most effective in use cases that pertain to memory, this result in accordance to Shneiderman (2016), suggests that the use of outline icons are most effective in adhering to Shneiderman's rule of reducing short term memory load when presented in a resemblance form. This consequently provides a limited answer to the research question: “How do the pre-existing icon classifications discussed by Arledge and Nielsen impact memory performance?” - the semantic classifications of Nielsen, have a significantly varying impact on the memory load of Arledge’s outline style.

The word “limited” is used when addressing the project’s research question, as although this result provides evidence towards semantic having a significantly varying impact on the memory load of outline icons, this result was only observed in one comparison. This means that when judging this result critically, the argument could be made that future work would require this result to be tested with a more direct methodology, centred around the testing of outline icons and how they interact with different icon semantic classifications. This would be done to further validate this study’s result. Despite this, the contribution and evidence provided by this result cannot be ignored based on its independence, as based on the comparisons effect size ($r = 0.3443$), the significant difference is in fact moderate in size.

The argument that this study’s contribution is lacking, can also be discredited by referring to the literature review. In the literature review, N100 as well as P300 was highlighted as an ERP that is relevant to the N-Back task. As this study only analysed one out of 32 channels of EEG (Pz) and N100 was not analysed due to time constraints, this discredits the argument that this study’s contribution is lacking, as the current study’s results do not provide a holistic perspective on the significance of this study’s data, therefore this study's conclusion of insignificance, is limited in scale.

Alternatively, this study may have a larger contribution to psychology, as it puts forward the idea that each classification’s impact on user cognition is, based on these results, equal. In addition to this, although this study’s contribution can be bolstered by the immediate analysis of N100 and N400, a task made feasible by the fact that the data has already been recorded, this study has already analysed P300 results, which is described to be an important ERP component for evaluating cognitive function (e.g. attention and working memory (Zhong et al., 2019, p. 4)).

Furthermore, on a macro level, the data recorded is a contribution to the future research on iconography. This is a valid point as if this data was used in a project larger than the current study, the dataset recorded could be analysed using several more channels for several more ERPs. The results could be compared to other studies that used a different methodology in order to reach further insights into the icons used in this study. As a result, the further use of this study's data will enable the scale of comparison to improve past a single study, and will instead enable the conversation on iconography to be continued not in the qualitative conversation of theory, but in the comparison of data.

To conclude, this section discusses the areas where this study can improve, identifying ways in which future research can expand this study's investigation into icon usability. Addressing this study's limitations and proposing solutions to them, in the form of future work has been done for the purpose of producing an even greater understanding of icon usability and its impacts on human-computer interaction past this current study.

8. Conclusion

To achieve this project's purpose of better understanding the circumstances in which icons are most effective, in use cases that pertain to memory such as reaction times and brain responses, this project aimed to investigate the research question: "How do the pre-existing icon classifications discussed by Arledge and Nielsen impact memory performance?".

This study's results present an interesting narrative. Hypothesis 1 predicted that either Solid or Outline Icons would be recalled significantly faster than the other style of icon. This was not supported by the results recorded. Despite a fraction of the findings showing that outline icons in an abstract form were recalled significantly faster than their solid counterparts, this observation was not large enough across all the results to support Hypothesis 1, suggesting the equality of memory performance between the two styles.

Hypothesis 2's anticipation of Resemblance icons having faster recall speeds compared to Abstract icons was contradicted by the results. Results showed a faster recall speed for Abstract icons in an Outline style compared to Resemblance icons in the same style. Despite being a statistically significant result, this finding represented a minority of the comparisons made between Abstract and Resemblance icons. Combining this result with the fact that neither Resemblance or Abstract icons had significantly faster reaction times, in each instance the two semantic groups were compared, Hypothesis 2 was rejected.

Hypothesis 3 predicted that there would be a clear distinction between the memory loads of the icon classifications. This study chose to use the P300 component in order to measure this. The results indicated that there was no significant differentiation between the memory loads of the icon classifications used, because of this, Hypothesis 3 is rejected. Despite Hypotheses 2 & 3 rejecting their initial predictions, a potential relationship between the ERPs and reaction times of OA and OR stimuli was revealed, showing potential for future research.

In addition to this, a comparison showed that OA had a significantly smaller max peak P300 amplitude compared to OR. This provides evidence towards OR having significantly better usability compared to OA, due to its ability to invoke a significantly smaller memory load compared to OA (Shneiderman, 2016). This conclusion on memory load and the P300 amplitude is supported by previous studies (Scharinger et al., 2017, p. 3; Ren et al., 2023, p. 1; Luck & Kappenman, 2013, p. 165). Combining the significant results on the reaction times and max peak amplitudes of OR vs OA, this shows potential for future investigation, hence its speculation and future work being detailed in the discussion.

Speaking of future research, this project's limitations must be acknowledged to inform future work. In hindsight, this study's scope was limited by the variety of icon dimensions and styles used in experimentation. This points towards the need for a wider implementation of style and form in future work. This implementation will aid our understanding of the memory performance of different iconography beyond this study. For example, varying dimensions of icons have been indicated to potentially influence user cognition. Future studies could utilise this potential, in order to observe significant differences in user cognition between the styles used in this study, be that with a slightly altered design. For example, the implementation of other icon dimensions and styles, could involve the use of various sizes of icons. The comparison of the different memory loads elicited by different sizes of icon, might provide insight into how much of an interface or screen an icon should occupy and if there is a potential trade-off between the usability of an icon and its size. For example, larger icons could be perceived as more visually prominent, compared to smaller icons, causing users to recall such icons faster with an increased memory load due to a larger size conveying a more visually emphasised presence. Dually, smaller icons might be recalled slower, as their complexity and memorable design components might be obscured by their size, making it harder for the user to completely understand the design of the icon. Smaller icons might yield a smaller memory load as their smaller size might limit the icons intelligibility. This could cause the user to prioritise larger, more visually understood icons in memory. If a smaller memory load is elicited, small icons could be argued as having better usability compared to large icons. Despite this, if the cause of this reduced level of memory load is due to the small icons being unidentifiable because of their size, this might reveal a trade-off between the usability of an icon and its size. In terms of the other design components this study could explore, the same dilemma of intelligibility vs memory load could be observed when comparing different icon colours, furthermore, the concept of colour vision deficiency, otherwise known as colour blindness, might complicate this dilemma further.

These potential directions in future research are important as they will allow future studies to better understand the usability of icons featured heavily in modern day user interfaces. As a result, this improvement of understanding might also lead to the improvement of icon design in the field of user experience design and human computer interaction.

Alongside the suggested future work, the study's results have a broader applicability to applications in the real world. An example of how the results could be applied to the real world can be found in the comparison between OR and OA icons. With OR icons requiring a significantly lower memory load (as demonstrated by the ERP analysis), use cases that could benefit from a low memory load might benefit from including OR icons instead of OA icons. An example is the multitasking functionalities featured in different mobile and computer operating systems. It could be argued that an action such as transferring information between two windows on a screen or quickly taking information from one app to another, might be more efficient if the information conveyed by an icon, causes a reduced cognitive strain. In this case, featuring an OR icon might achieve this more efficiently compared to an OA icon. Moving past the example of simply taking information from one screen or application to another, perhaps more mission-critical operations could benefit from this finding. Interfaces featured on tools and vehicles used in remote or extraterrestrial areas might require a design that prevents cognitive overload, due to the potential danger of the environments these tools and vehicles are used in. Using icons that require less memory load (such as OR icons instead of OA icons) could ensure the safety of the end user. In fact, this study's result puts more emphasis on why ensuring the resemblance of outlined icons is important. If the icon design is perceived to be more abstract rather than something that depicts a real-life object, this could lead to the user becoming cognitively overwhelmed due to a heightened memory load caused by the OA icon.

Focussing on the practical implications of this study's results, the overall insignificance between the memory loads of the icon combinations, provides interface designers with an opportunity to focus on other design components such as the size and colour of an icon. This means that user interface designers do not have to be overly concerned about design components that, given the results of this study, have no optimal combination for eliciting the lowest possible memory load. Despite this short-term practical implication, future work should aim to build on this by better understanding if there are other icon design components that might impact memory load more significantly, compared to the style and semantic classifications used in this study. Subsequently, if the field of UI design research can better understand the design components that do significantly impact memory load, a long-term practical implication of both this study and its future work, could be a more defined list of design constraints that ensure the highest possible level of icon usability.

The contribution this study makes to Computing, Psychology, Human Computer Interaction and User Experience Design is significant. This study has challenged the perspectives of previous research and has produced novel and rivalling understandings of icon usability. As a result, this thesis supplements the previous research it references by testing their ideas. The results of this investigation add to an ongoing conversation on how users interpret icons based on a myriad of factors i.e. Style and Semantic. To summarise, this thesis contributes to the ongoing discussion of icon usability, whilst also suggesting future directions on how our understanding of icons can be improved through a closer look at the visual and semantical components they're comprised of.

9. Reflection on Learning

As part of this masters degree by research, to ensure the completion and feasibility of this project, the researcher introduced themselves to MATLAB programming and signal analysis. Under the supervision of Dr Palaniappan Ramaswamy, the researcher was able to refine the skills developed in these areas through trial and error. This project has provided the researcher with an invaluable lesson on time management and independent work.

Coming from a computer science undergraduate background with no prior research experience, the researcher thought the commencing of a masters degree by research would provide them with a practical exposure to the computing industry, especially in the execution of research in the field.

Dually, this research has encountered many obstacles, both internally regarding the execution of the research and externally regarding the researcher's health. Because of this, the support and mentorship of Dr Palaniappan Ramaswamy was critical to this project and is appreciated beyond measure. Overall, this masters by research has been very challenging undoubtedly like all research respectively, its exposed the researcher to a tangible research experience that has garnered an entry level of expertise that will prove useful in the beginning of the researcher's career commencing later in the year.

10. References

Amin, H. U., Malik, A. S., Kamel, N., Chooi, W.-T., & Hussain, M. (2015). P300 correlates with learning & memory abilities and fluid intelligence. *Journal of NeuroEngineering and Rehabilitation*, *12*(1). https://doi.org/10.1186/s12984-015-0077-6

Arledge, C. (2014). *Filled-in vs. Outline Icons: The Impact of Icon Style on Usability*. Cdr.lib.unc.edu. https://doi.org/10.17615/jqx8-3d26

Cambridge Dictionary. (2024). *memorability*. @CambridgeWords. https://dictionary.cambridge.org/dictionary/english/memorability

Cherng, F.-Y., Lin, W.-C., King, J.-T., & Lee, Y.-C. (2016). An EEG-based Approach for Evaluating Graphic Icons from the Perspective of Semantic Distance. https://doi.org/10.1145/2858036.2858133

Cherry, K. (2024). *Cognition in Psychology*. Verywell Mind. https://www.verywellmind.com/what-is-cognition-2794982

Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184. https://doi.org/10.1016/j.conb.2007.03.005

Dovetail Editorial Team. (2023). *What Is Skeuomorphism? Pros, Cons, Evolution, Examples*. Dovetail.com. https://dovetail.com/ux/skeuomorphism/#:~:text=Skeuomorphism%20is%20a%20design%20trend

Ece Aksen, D. (2023). *Impact of Working Memory Load and Emotional Valence on State Impulsivity: An Experimental Study*[Dissertation/Thesis]. https://www.proquest.com/openview/8f10cbb90bb703210a4f7fd60d0ccc87/1?pq-origsite=gscholar&cbl=18750&diss=y

Eide, S. A., Poljac, A.-M., & Sandnes, F. E. (2021). Image Search Versus Text Search Revisited: A Simple Experiment Using a Kahoot Quiz. *Universal Access in Human-Computer Interaction. Design Methods and User Experience*, 186–196. https://doi.org/10.1007/978-3-030-78092-0_12

Fitzgerald, L. (2016). *ScholarWorks Cognitive Processes and Memory Differences in Recall and Recognition in Adults*. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=2599&context=disser

tations&httpsredir=1&referer=#:~:text=Cognitive%20psychology%20studies%20the%20processes

Gittins, D. (1986). Icon-based human-computer interaction. *International Journal of Man-Machine Studies*, *24*(6), 519–543. https://doi.org/10.1016/s0020-7373(86)80007-4

Grissmann, S., Faller, J., Scharinger, C., Spüler, M., & Gerjets, P. (2017). Electroencephalography Based Analysis of Working Memory Load and Affective Valence in an N-back Task with Emotional Stimuli. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00616

Guo, Q. (2016). An icon recognition study on different simplicity levels. *Dr.lib.iastate.edu*. https://doi.org/10.31274/etd-180810-4604

Harley, A. (2014). *Icon Usability*. Nielsen Norman Group. https://www.nngroup.com/articles/icon-usability/

Harvard Medical School. (2022). *Memory. Harvard Health*. Www.health.harvard.edu. https://www.health.harvard.edu/topics/memory

Jin, H. (2020). *Influence of Icon Design Style on User's Cognition*. https://doi.org/10.2991/assehr.k.200428.118

Kleiner, M., Brainard, D., Pelli, D., Broussard, C., Wolf, T., & Niehorster, D. (2018a). *Psychtoolbox-3 - GetSecs*. Psychtoolbox.org. http://psychtoolbox.org/docs/GetSecs

Kleiner, M., Brainard, D., Pelli, D., Broussard, C., Wolf, T., & Niehorster, D. (2018b). *Psychtoolbox-3 - Overview*. Psychtoolbox.org. http://psychtoolbox.org/overview.html#:~:text=The%20routines%20at%20the%20core

Kleiner, M., Brainard, D., Pelli, D., Broussard, C., Wolf, T., & Niehorster, D. (2018c). *Psychtoolbox-3 - WaitSecs*. Psychtoolbox.org. http://psychtoolbox.org/docs/WaitSecs#:~:text=WaitSecs%20uses%20Windows%20QueryPerformanceCounter()

Ko, Y.-H., Shen, I-Hsuan., & Lee, D.-S. (2010). Color Combinations of Visual Display Terminal (VDT) Icon on User Preferences and EEG Response. *Perceptual and Motor Skills*, *110*(2), 411–428. https://doi.org/10.2466/pms.110.2.411-428

KrishaWeb. (2020). *Outline Icons vs. Solid Icons: Which are Faster to Recognize and When?* KrishaWeb. https://www.krishaweb.com/blog/outline-icons-vs-solid-icons-which-are-faster-to-recognize-and-when/#:~:text=In%20general%2C%20the%20solid%20icons

Lahr, J., Minkova, L., Tabrizi, S. J., Stout, J. C., Klöppel, S., & Scheller, E. (2018). Working Memory-Related Effective Connectivity in Huntington's Disease Patients. *Frontiers in Neurology*, *9*. https://doi.org/10.3389/fneur.2018.00370

Legleiter, A., & Caporusso, N. (2020). *Flat-Design Icon Sets: A Case for Universal Meanings?* https://doi.org/10.1007/978-3-030-51828-8_27

Leuchs, L. (2019). *Choosing your reference & why it matters - a BrainVision Analyzer Tip*. Brain Products Press Release. https://pressrelease.brainproducts.com/referencing/#:~:text=The%20reference%20influences%20the%20amplitude

Linden, D. E. J. (2005). The P300: Where in the Brain Is It Produced and What Does It Tell Us? *The Neuroscientist*, *11*(6), 563–576. https://doi.org/10.1177/1073858405280524

liu, T. (2017). *Hollow Icon vs. Solid Icon: Which Is More Friendly for Recognizing?* Medium. https://medium.com/@tristaljing/hollow-icon-vs-solid-icon-which-is-more-friendly-for-recognizing-3cf032849263

Luck, S. J., & Kappenman, E. S. (2013). *The Oxford handbook of event-related potential components*. Oxford University Press.

Lystad, R. P., & Pollard, H. (2009). Functional neuroimaging: a brief overview and feasibility for use in chiropractic research. *The Journal of the Canadian Chiropractic Association*, *53*(1), 59–72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652631/#:~:text=Temporal%20resolution%20refers%20to%20how

Mak, Y. (2015). *What is cognition?* Cambridge Cognition. https://cambridgecognition.com/what-is-cognition/

McGovern Medical School. (n.d.). *Ear Anatomy – Outer Ear | Otorhinolaryngology – Head & Neck Surgery | McGovern Medical School*. Uth.edu. https://med.uth.edu/orl/online-ear-disease-photo-book/chapter-3-ear-anatomy/ear-anatomy-outer-ear/

Mcleod, S. (2019). *Sampling Distribution*. Simplypsychology.org; Simply Psychology. https://www.simplypsychology.org/effect-size.html

Mcleod, S. (2023). *Short Term Memory | Simply Psychology*. Simplypsychology.org. https://www.simplypsychology.org/short-term-memory.html

Mele, G., Cavaliere, C., Alfano, V., Orsini, M., Salvatore, M., & Aiello, M. (2019). Simultaneous EEG-fMRI for Functional Neurological Assessment. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00848

Meule, A. (2017). Reporting and Interpreting Working Memory Performance in n-back Tasks. *Frontiers in Psychology*, *8*(352). https://doi.org/10.3389/fpsyg.2017.00352

Microsoft. (2024). *QueryPerformanceCounter function - Win32 apps*. Learn.microsoft.com. https://learn.microsoft.com/en-us/windows/win32/api/profileapi/nf-profileapi-queryperformancecounter

Mu, D., Huang, Y., Wang, Y., Yang, J., Li, J., & Kang, Z. (2022). Neumorphic or flat? Impact of icon depth on user attention and visual search efficiency. *International Journal of Industrial Ergonomics*, *89*, 103281–103281. https://doi.org/10.1016/j.ergon.2022.103281

Nawaz, R., Wood, G., Nisar, H., & Yap, V. V. (2023). Exploring the Effects of EEG-Based Alpha Neurofeedback on Working Memory Capacity in Healthy Participants. *Bioengineering*, *10*(2), 200. https://doi.org/10.3390/bioengineering10020200

Neuro Electrics. (n.d.-a). *https://www.neuroelectrics.com/*. Neuroelectrics: Reinventing Brain Health. Retrieved February 26, 2024, from https://www.neuroelectrics.com/solution/spareparts-consumables/cap

Neuro Electrics. (n.d.-b). *Neuroelectrics Starstim 20 / Starstim 32 -Part 1- Instructions for Use*.

Neuroelectrics. (2020). *Neuroelectrics Frequently Asked Questions (FAQs) - Neuroelectric's Wiki*. Www.neuroelectrics.com. https://www.neuroelectrics.com/wiki/index.php/Neuroelectrics_Frequently_Asked_Questions_(FAQs)#What_are_the_sampling_rate_of_our_devices.3F

Ng, A. W. Y., & Chan, A. H. S. (2008). *Visual and Cognitive Features on Icon Effectiveness*. 1856-1859. Paper presented at Proceedings of the 2008 IAENG International

Conference on Industrial Engineering (ICINDE 2008), International MultiConference of Engineers and Computer Scientists 2008 (IMECS 2008), China.

Nielsen, J. (2012). *Usability 101: Introduction to Usability*. Nielsen Norman Group. https://www.nngroup.com/articles/usability-101-introduction-to-usability

Nielsen, J. (2014). *Icon Classification: Resemblance, Reference, and Arbitrary Icons*. Nielsen Norman Group. https://www.nngroup.com/articles/classifying-icons

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59. https://doi.org/10.1002/hbm.20131

Pallant, J. (2020). *SPSS Survival Manual* (pp. 254–255). Routledge. https://doi.org/10.4324/9781003117452

Pearson Talent Lens. (n.d.). *Raven's Advanced Progressive Matrices*. Www.talentlens.com. https://www.talentlens.com/recruitment/assessments/raven-s-advanced-progressive-matrices--apm-iii-.html#:~:text=RAVEN

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. M. (2018). *N*-back training and transfer effects revealed by behavioral responses and EEG. *Brain and Behavior*, *8*(11), e01136. https://doi.org/10.1002/brb3.1136

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. M. (2019). Mental workload of young and older adults gauged with ERPs and spectral power during N-Back task performance. *Biological Psychology*, *146*, 107726. https://doi.org/10.1016/j.biopsycho.2019.107726

Perrier, M. J. R., Louw, T. L., & Carsten, O. M. J. (2022). Usability testing of three visual HMIs for assisted driving: How design impacts driver distraction and mental models. *Ergonomics*, 1–22. https://doi.org/10.1080/00140139.2022.2136766

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of P300: an integrative review. *Biological Psychology*, *41*(2), 103–146. https://doi.org/10.1016/0301-0511(95)05130-9

Punchoojit, L., & Hongwarittorrn, N. (2018). *The Influence of Icon Background Colors and Icon Symbols on Menu Item Selection for Smartphone*. IEEE Xplore. https://doi.org/10.1109/NICS.2018.8606854

Ren, M., Xu, J., Li, Y., Wang, M., Georgiev, G., Shen, L., Zhao, J., Cao, Z., Zhang, S., Wang, W., Xu, S., Zhou, Z., Chen, S., Chen, X., Shi, X., Tang, X., & Shan, C. (2023). Neural signatures for the n-back task with different loads: An event-related potential study. *Biological Psychology*, *177*, 108485–108485. https://doi.org/10.1016/j.biopsycho.2023.108485

Rossi, A., & Lenzini, G. (2021). Which Properties Has an Icon? A Critical Discussion on Data Protection Iconography. *Socio-Technical Aspects in Security and Trust*, 211–229. https://doi.org/10.1007/978-3-030-55958-8_12

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2017). Comparison of the Working Memory Load in N-Back and Working Memory Span Tasks by Means of EEG Frequency Band Power and P300 Amplitude. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00006

Science Direct. (2017). *N-Back - an overview | ScienceDirect Topics*. Sciencedirect.com. https://www.sciencedirect.com/topics/neuroscience/n-back

Shneiderman, B. (2016). *The Eight Golden Rules of Interface Design*. Umd.edu. https://www.cs.umd.edu/users/ben/goldenrules.html

Simpson, T. G., & Rafferty, K. (2021). Investigating the P300 Response as a Marker of Working Memory in Virtual Training Environments. *IEEE Transactions on Human-Machine Systems*, *51*(3), 265–277. https://doi.org/10.1109/thms.2021.3064815

Sturzbecher, M. J., & de Araujo, D. B. (2012). Simultaneous EEG-fMRI: Integrating Spatial and Temporal Resolution. *The Relevance of the Time Domain to Neural Network Models*, *3*, 199–217. https://doi.org/10.1007/978-1-4614-0724-9_11

Sur, S., & Sinha, V. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal*, *18*(1), 70-73. https://doi.org/10.4103/0972-6748.57865

Tao, Q., Jiang, L., Li, F., Qiu, Y., Yi, C., Si, Y., Li, C., Zhang, T., Yao, D., & Xu, P. (2022). Dynamic networks of P300-related process. *Cognitive Neurodynamics*, *16*(5), 975–985. https://doi.org/10.1007/s11571-021-09753-3

The Human Memory. (2019). *Memory Recall/Retrieval | Types, Processes, Improvement & Problems*. The Human Memory. https://human-memory.net/memory-recall-retrieval/

uxmovement. (2019). *Solid Vs. Outline Icons: Which Are Faster to Recognize?* Uxmovement. https://uxmovement.com/mobile/solid-vs-outline-icons-which-are-faster-to-recognize/#:~:text=Viewing%20icons%20as%20outlines%20is

Wake, L. (2021). *Should icons be solid or hollow? | Digital Communications team blog*. Digitalcommunications.wp.st-Andrews.ac.uk. https://digitalcommunications.wp.st-andrews.ac.uk/2021/09/23/should-icons-be-solid-or-hollow/

Wang, S., Gwizdka, J., & Chaovalitwongse, W. A. (2016). Using Wireless EEG Signals to Assess Memory Workload in the N-Back Task. *IEEE Transactions on Human-Machine Systems*, *46*(3), 424–435. https://doi.org/10.1109/thms.2015.2476818

Xia, Y. (2020). *Wilcoxon Signed Ranks Test - an overview | ScienceDirect Topics*. Www.sciencedirect.com. https://www.sciencedirect.com/topics/medicine-and-dentistry/wilcoxon-signed-ranks-test#:~:text=Wilcoxon%20rank%2Dsum%20test%20is

Xiao-Fan, Q., Ying, Y., & Shou-Qian, X. U. N. (2014). Exploring the Effects of Icon Concreteness on the Recognition Process: An Event-Related Potential Study. *Journal of Psychological Science*, *37*(1), 27. http://www.psysci.org/EN/abstract/abstract9153.shtml

Yang, C., Peng, Y., & Zeng, J. (2021). Research on cognition and application of icon complexity based on EEG. *CCF Transactions on Pervasive Computing and Interaction*, *3*(2), 170–185. https://doi.org/10.1007/s42486-021-00058-2

Yeh, Y.-Y., Lee, D.-S., & Ko, Y.-H. (2013). Color combination and exposure time on legibility and EEG response of icon presented on visual display terminal. *Displays*, *34*(1), 33–38. https://doi.org/10.1016/j.displa.2012.11.007

Yen, C., Lin, C.-L., & Chiang, M.-C. (2023). Exploring the frontiers of neuroimaging: A review of recent advances in understanding brain functioning and disorders. *Life*, *13*(7), 1–27. https://doi.org/10.3390/life13071472

Zhang, G., Garrett, D. R., & Luck, S. J. (2023). Optimal Filters for ERP Research II: Recommended Settings for Seven Common ERP Components. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.06.13.544794

Zhong, R., Li, M., Chen, Q., Li, J., Li, G., & Lin, W. (2019). The P300 Event-Related Potential Component and Cognitive Impairment in Epilepsy: A Systematic Review and Meta-analysis. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00943

Appendix A - Right Tail Wilcoxon Signed-Rank Test Results

Table A.1: Correct Reaction Times- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	2.078831	0.0188	0.4243	0.4887	0.491
(N2) OR VS OA	0.902134	0.1835	0.1841	0.5899	0.5867
(N1 & N2) OR VS OA	1.921938	0.0273	0.3923	0.5724	0.5469
(N1) SR VS OR	-0.353009	0.638	0.0721	0.4893	0.4887
(N2) SR VS OR	0.196116	0.4223	0.04	0.6103	0.5899
(N1 & N2) SR VS OR	-0.588348	0.7219	0.1201	0.5577	0.5724
(N1) SR VS OA	1.294366	0.0978	0.2642	0.4893	0.491
(N2) SR VS OA	1.137474	0.1277	0.2322	0.6103	0.5867
(N1 & N2) SR VS OA	1.21592	0.112	0.2482	0.5577	0.5469
(N1) SA VS OR	-0.902134	0.8165	0.1841	0.4981	0.4887
(N2) SA VS OR	0.823688	0.2051	0.1681	0.6337	0.5899
(N1 & N2) SA VS OR	0.039223	0.4844	0.008	0.5701	0.5724
(N1) SA VS OA	1.137474	0.1277	0.2322	0.4981	0.491
(N2) SA VS OA	1.529706	0.063	0.3122	0.6337	0.5867
(N1 & N2) SA VS OA	1.843492	0.0326	0.3763	0.5701	0.5469
(N1) SR VS SA	-0.196116	0.5777	0.04	0.4893	0.4981
(N2) SR VS SA	-0.902134	0.8165	0.1841	0.6103	0.6337
(N1 & N2) SR VS SA	-1.372813	0.9151	0.2802	0.5577	0.5701

Table A.2: Correct Answer Accuracy (as a percentage)- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	-0.744092	0.7716	0.1519	99.0	98.0
(N2) OR VS OA	-1.67034	0.9526	0.341	95.0	97.0
(N1 & N2) OR VS OA	-1.48741	0.9315	0.3036	96.5	98.5
(N1) SR VS OR	-0.274721	0.6082	0.0561	100.0	99.0
(N2) SR VS OR	0.256326	0.3988	0.0523	95.0	95.0
(N1 & N2) SR VS OR	0.204658	0.4189	0.0418	97.0	96.5
(N1) SR VS OA	-1.474087	0.9298	0.3009	100.0	98.0
(N2) SR VS OA	-0.83295	0.7976	0.17	95.0	97.0
(N1 & N2) SR VS OA	-1.012335	0.8443	0.2066	97.0	98.5
(N1) SA VS OR	-0.282138	0.6111	0.0576	99.0	99.0
(N2) SA VS OR	0.65504	0.2562	0.1337	95.0	95.0
(N1 & N2) SA VS OR	0.918559	0.1792	0.1875	96.0	96.5
(N1) SA VS OA	-0.636396	0.7377	0.1299	99.0	98.0
(N2) SA VS OA	-0.833688	0.7978	0.1702	95.0	97.0
(N1 & N2) SA VS OA	-0.758747	0.776	0.1549	96.0	98.5
(N1) SR VS SA	-0.281788	0.6109	0.0575	100.0	99.0
(N2) SR VS SA	0.222828	0.4118	0.0455	95.0	95.0
(N1 & N2) SR VS SA	-0.078811	0.5314	0.0161	97.0	96.0

Table A.3: Correct Max Peak Amplitudes- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	1.686599	0.0458	0.3443	14.9783	13.9205
(N2) OR VS OA	0.666795	0.2525	0.1361	11.7179	14.0166
(N1 & N2) OR VS OA	1.529706	0.063	0.3122	13.3853	12.2716
(N1) SR VS OR	-0.588348	0.7219	0.1201	14.4654	14.9783
(N2) SR VS OR	-2.445048	0.9928	0.4991	11.3734	11.7179
(N1 & N2) SR VS OR	-1.137474	0.8723	0.2322	12.7718	13.3853
(N1) SR VS OA	0.588348	0.2781	0.1201	14.4654	13.9205
(N2) SR VS OA	-0.844653	0.8008	0.1724	11.3734	14.0166
(N1 & N2) SR VS OA	-0.039223	0.5156	0.0080	12.7718	12.2716
(N1) SA VS OR	-0.902134	0.8165	0.1841	14.7053	14.9783
(N2) SA VS OR	0.431455	0.3331	0.0881	14.01	11.7179
(N1 & N2) SA VS OR	-0.039223	0.5156	0.008	15.4215	13.3853
(N1) SA VS OA	0.196116	0.4223	0.04	14.7053	13.9205
(N2) SA VS OA	1.529706	0.063	0.3122	14.01	14.0166
(N1 & N2) SA VS OA	0.666795	0.2525	0.1361	15.4215	12.2716
(N1) SR VS SA	0.11767	0.4532	0.024	14.4654	14.7053
(N2) SR VS SA	-2.356137	0.9908	0.4809	11.3734	14.01
(N1 & N2) SR VS SA	-1.451259	0.9266	0.2962	12.7718	15.4215

Table A.4: Correct Max Peak Latencies- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	0.039299	0.4843	0.008	362	365
(N2) OR VS OA	0.311496	0.3777	0.0636	367	366
(N1 & N2) OR VS OA	0.431622	0.333	0.0881	373	366.5
(N1) SR VS OR	-0.824322	0.7951	0.1683	363	362
(N2) SR VS OR	0.489979	0.3121	0.1000	366	367
(N1 & N2) SR VS OR	-0.745528	0.7720	0.1522	359.5	373
(N1) SR VS OA	-1.201484	0.8852	0.2453	363	365
(N2) SR VS OA	0.000000	0.5000	0.0000	366	366
(N1 & N2) SR VS OA	-0.509902	0.6949	0.1041	359.5	366.5
(N1) SA VS OR	-0.400297	0.6555	0.0817	362	362
(N2) SA VS OR	0.274774	0.3917	0.0561	360	367
(N1 & N2) SA VS OR	-0.588575	0.7219	0.1201	362.5	373
(N1) SA VS OA	0.432455	0.3327	0.0883	362	365
(N2) SA VS OA	0.311188	0.3778	0.0635	360	366
(N1 & N2) SA VS OA	0.117715	0.4531	0.024	362.5	366.5
(N1) SR VS SA	-0.816497	0.7929	0.1667	363	362
(N2) SR VS SA	-0.408248	0.6585	0.0833	366	360
(N1 & N2) SR VS SA	-0.078507	0.5313	0.0160	359.5	362.5

Appendix B - Left Tail Wilcoxon Signed-Rank Test Results

Table B.1: Correct Reaction Times- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	2.157277	0.9845	0.4404	0.4887	0.491
(N2) OR VS OA	0.980581	0.8366	0.2002	0.5899	0.5867
(N1 & N2) OR VS OA	2.000385	0.9773	0.4083	0.5724	0.5469
(N1) SR VS OR	-0.274563	0.3918	0.056	0.4893	0.4887
(N2) SR VS OR	0.274563	0.6082	0.056	0.6103	0.5899
(N1 & N2) SR VS OR	-0.509902	0.3051	0.1041	0.5577	0.5724
(N1) SR VS OA	1.372813	0.9151	0.2802	0.4893	0.491
(N2) SR VS OA	1.21592	0.888	0.2482	0.6103	0.5867
(N1 & N2) SR VS OA	1.294366	0.9022	0.2642	0.5577	0.5469
(N1) SA VS OR	-0.823688	0.2051	0.1681	0.4981	0.4887
(N2) SA VS OR	0.902134	0.8165	0.1841	0.6337	0.5899
(N1 & N2) SA VS OR	0.11767	0.5468	0.024	0.5701	0.5724
(N1) SA VS OA	1.21592	0.888	0.2482	0.4981	0.491
(N2) SA VS OA	1.608152	0.9461	0.3283	0.6337	0.5867
(N1 & N2) SA VS OA	1.921938	0.9727	0.3923	0.5701	0.5469
(N1) SR VS SA	-0.11767	0.4532	0.024	0.4893	0.4981
(N2) SR VS SA	-0.823688	0.2051	0.1681	0.6103	0.6337
(N1 & N2) SR VS SA	-1.294366	0.0978	0.2642	0.5577	0.5701

Table B.2: Correct Answer Accuracy (as a percentage) - Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	-0.531494	0.2975	0.1085	99.0	98.0
(N2) OR VS OA	-1.55103	0.0604	0.3166	95.0	97.0
(N1 & N2) OR VS OA	-1.368417	0.0856	0.2793	96.5	98.5
(N1) SR VS OR	0.0	0.5	0.0	100.0	99.0
(N2) SR VS OR	0.358856	0.6401	0.0733	95.0	95.0
(N1 & N2) SR VS OR	0.306987	0.6206	0.0627	97.0	96.5
(N1) SR VS OA	-1.300665	0.0967	0.2655	100.0	98.0
(N2) SR VS OA	-0.713957	0.2376	0.1457	95.0	97.0
(N1 & N2) SR VS OA	-0.893237	0.1859	0.1823	97.0	98.5
(N1) SA VS OR	-0.141069	0.4439	0.0288	99.0	99.0
(N2) SA VS OR	0.774139	0.7806	0.158	95.0	95.0
(N1 & N2) SA VS OR	1.020621	0.8463	0.2083	96.0	96.5
(N1) SA VS OA	-0.494975	0.3103	0.101	99.0	98.0
(N2) SA VS OA	-0.71459	0.2374	0.1459	95.0	97.0
(N1 & N2) SA VS OA	-0.669483	0.2516	0.1367	96.0	98.5
(N1) SR VS SA	-0.140894	0.444	0.0288	100.0	99.0
(N2) SR VS SA	0.31196	0.6225	0.0637	95.0	95.0
(N1 & N2) SR VS SA	0.0	0.5	0.0	97.0	96.0

Table B.3: Correct Max Peak Amplitudes- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	1.765045	0.9612	0.3603	14.9783	13.9205
(N2) OR VS OA	0.745241	0.7719	0.1521	11.7179	14.0166
(N1 & N2) OR VS OA	1.608152	0.9461	0.3283	13.3853	12.2716
(N1) SR VS OR	-0.509902	0.3051	0.1041	14.4654	14.9783
(N2) SR VS OR	-2.356137	0.0092	0.4809	11.3734	11.7179
(N1 & N2) SR VS OR	-1.059027	0.1448	0.2162	12.7718	13.3853
(N1) SR VS OA	0.666795	0.7475	0.1361	14.4654	13.9205
(N2) SR VS OA	-0.755742	0.2249	0.1543	11.3734	14.0166
(N1 & N2) SR VS OA	0.039223	0.5156	0.0080	12.7718	12.2716
(N1) SA VS OR	-0.823688	0.2051	0.1681	14.7053	14.9783
(N2) SA VS OR	0.509902	0.6949	0.1041	14.01	11.7179
(N1 & N2) SA VS OR	0.039223	0.5156	0.0080	15.4215	13.3853
(N1) SA VS OA	0.274563	0.6082	0.056	14.7053	13.9205
(N2) SA VS OA	1.608152	0.9461	0.3283	14.01	14.0166
(N1 & N2) SA VS OA	0.745241	0.7719	0.1521	15.4215	12.2716
(N1) SR VS SA	0.196116	0.5777	0.04	14.4654	14.7053
(N2) SR VS SA	-2.267226	0.0117	0.4628	11.3734	14.0100
(N1 & N2) SR VS SA	-1.372813	0.0849	0.2802	12.7718	15.4215

Table B.4: Correct Max Peak Latencies- Wilcoxon Signed-Rank Test Results

Test	Z-value	p-value	Effect size (r)	Sample 1 Median	Sample 2 Median
(N1) OR VS OA	0.117897	0.5469	0.0241	362.0	365.0
(N2) OR VS OA	0.400495	0.6556	0.0818	367.0	366.0
(N1 & N2) OR VS OA	0.510098	0.695	0.1041	373.0	366.5
(N1) SR VS OR	-0.745815	0.2279	0.1522	363.0	362.0
(N2) SR VS OR	0.579066	0.7187	0.1182	366.0	367.0
(N1 & N2) SR VS OR	-0.667051	0.2524	0.1362	359.5	373.0
(N1) SR VS OA	-1.112485	0.133	0.2271	363.0	365.0
(N2) SR VS OA	0.101929	0.5406	0.0208	366.0	366.0
(N1 & N2) SR VS OA	-0.431455	0.3331	0.0881	359.5	366.5
(N1) SA VS OR	-0.311342	0.3778	0.0636	362.0	362.0
(N2) SA VS OR	0.353281	0.6381	0.0721	360.0	367.0
(N1 & N2) SA VS OR	-0.510098	0.305	0.1041	362.5	373.0
(N1) SA VS OA	0.511083	0.6954	0.1043	362.0	365.0
(N2) SA VS OA	0.400099	0.6555	0.0817	360.0	366.0
(N1 & N2) SA VS OA	0.196192	0.5778	0.04	362.5	366.5
(N1) SR VS SA	-0.714435	0.2375	0.1458	363.0	362.0
(N2) SR VS SA	-0.306186	0.3797	0.0625	366.0	360.0
(N1 & N2) SR VS SA	0.000000	0.5000	0.0000	359.5	362.5

Appendix C - N-Back Task Technical Implementation

The N-Back Task used in this experiment is comprised of the following MATLAB files:

1. `PsychtoolboxExperiment.m`
2. `Preexperiment.m`

This appendix will illustrate the purpose of each file.

C.1 Preexperiment.m

In this file, the code enables data sharing via the Lab Streaming Layer (LSL) framework. LSL facilitates real-time data sharing between software applications and devices. The program first checks the value of `using_lsl` to activate LSL if set to true. The code then loads the LSL library, defines a data stream for event markers and creates an LSL outlet to send event markers (outlets in LSL facilitate the transmission of data with timestamps, e.g. event markers). This file enables the timestamps to be sent in PsychtoolboxExperiment.m thus this file was run before the experiment.

C.2 PsychtoolboxExperiment.m

C.2.1 Functions Used

The following functions are used in the following segments of code.

- The `screen(maketexture)` function is used to load images into the program, this converts the image into a usable texture so it can be drawn on a hidden screen.
- The `screen(drawtexture)` function is used to draw pictures / textures into a hidden screen.
- The `screen(flip)` function is used to display pictures / textures on the hidden screen.
- The `screen(FillRect)` function is used to draw a coloured box in to the program's hidden screen.
- The `outlet.push_sample()` function pushes an array of values into the outlet. Every time this function is used in this program, it acts as an event marker for the EEG recording in place, these markers describes an event based on the users interaction,

for example, `outlet.push_sample('Correct')` has been used to indicate in the EEG recording when the participant gets a correct answer.

C.2.2 Experiment Settings

In this section, specific settings for the N-Back are set. These settings are:

1. A record of the participant identification number.
2. A record of the stimuli type to be used.
3. The N-Back type to be used (1-Back or 2-Back)
4. The stimuli type to be used.

   Options limited to:

   a. Solid Resemblance icons
   b. Solid Abstract icons
   c. Outline Resemblance icons
   d. Outline Abstract icons
5. The stimuli sequence (a generated `.mat` file contains this sequence)
   a. Each unique sequence (`.mat` file) generated by the `genStimuliOrder.m` file and was given a specific name based on which participant and which N-Back it would be used for. For example, `stimuliOrderS12N2.mat` was used for participant 12's 2-Back recording. This unique sequence would then in turn determine which icons are shown and when in the N-Back task.

C.2.3 Load Images of Icons

This section loads the following images into the program:

- Two welcome screens, the welcome screen shown is based on the N-Back type being used.
- Images of the stimuli type specified to be used in the experiment settings.
- The end screen.

These images are loaded into the program using the `screen(maketexture)` function.

C.2.4 Preparation & N-Back Task

C.2.4.1 Preparation

Before the N-Back task commences, the following happens:

1) If LSL has been enabled, the program is allowed to continue.
2) The cursor on screen is hidden to remove the potential distraction from the participants view.
3) The usable keys on the keyboard are restricted to only the space bar (this is done to prevent accidental responses).
4) The welcome screen appropriate to the N-Back type is shown, prompting the participant to press SPACE when they are ready to begin.
5) The program waits for SPACE to be pressed before starting the N-Back task.
6) When SPACE has been pressed, the screen goes blank, giving the participant 10 seconds to get ready.
7) A timestamp is recorded in the EEG recording, this timestamp reports the program (N-Back task) being started.

C.2.4.2 N-Back Task: Stimuli Presentation

The program conducts the N-Back task. The current stimulus is shown for 1500 milliseconds. The stimulus shown is based on the number at the current position in the stimuli sequence specified in the "Experiment Settings" section of the program.

If the participant presses the SPACE (to indicate that they think the current stimulus matches the stimulus n places back), one of three outcomes occur:

1) If the position of the current stimulus is below the "Index Limit", the input is ignored as there have not been enough stimuli shown to give an answer to the N-Back task. i.e. the participant can only answer once shown enough stimuli to successfully or unsuccessfully judge if the current stimulus matches the stimulus shown n places ago. The Index Limit variable was created to prevent this error.
2) If the position of the current stimulus exceeds "Index Limit" and the current stimulus matches the stimulus shown n places ago, a green square is shown in the bottom of the screen as a response to a correct answer. A "Correct" timestamp is also recorded using the `outlet.push_sample()` function.

3) If the position of the current stimulus exceeds “Index Limit” and the current stimulus does not match the stimulus shown n places ago, a red square is shown in the bottom of the screen as a response to an incorrect answer. A “Incorrect” timestamp is also recorded using the `outlet.push_sample()` function.

These outcomes only occur if the participant has not already answered, the stimulus is presented for the stimulus presentation time regardless of the user’s interaction.

C.2.4.3 N-Back: Interstimulus Interval

A blank screen is shown during the ISI for a random duration between 1500 and 2000 milliseconds. This decision was made to prevent periodic responses.

During the interstimulus interval, the participant can still give their answer based on the stimulus they have just been shown, if the participant has not already answered.

If the participant presses the SPACE (to indicate that they think the current stimulus matches the stimulus n places back), one of three outcomes occur:

1) If the position of the latest stimulus is below the “Index Limit”, the input is ignored as there have not been enough stimuli shown to give an answer to the N-Back Task. i.e. The participant can only answer once shown enough stimuli to successfully or unsuccessfully judge if the latest stimulus matches the stimulus shown n stimuli ago.
2) If the position of the latest stimulus exceeds “Index Limit” and the latest stimulus matches the stimulus shown n places ago. A green square is shown in the bottom of the screen as a response to a correct answer. A “Correct” timestamp is also recorded using the “outlet.push_sample()” function.
3) If the position of the latest stimulus exceeds “Index Limit” and the latest stimulus does not match the stimulus shown n places ago. A red square is shown in the bottom of the screen as a response to a incorrect answer. A “Incorrect” timestamp is also recorded using the “outlet.push_sample()” function.

C.2.5 End Screen

In this section, the end screen is drawn to the hidden screen using the screen `DrawTexture` function. The end screen is displayed on screen using the screen `flip` function. Using the `outlet.push_sample()` function, the text “Program Ended” timestamp is created. The cursor is revealed, and the program execution concludes.